Reprezentatywność badań reprezentatywnych

UNIWERSYTET IM. ADAMA MICKIEWICZA W POZNANIU
SERIA SOCJOLOGIA NR 77

Piotr Jabkowski

Reprezentatywność badań reprezentatywnych

Analiza wybranych problemów metodologicznych oraz praktycznych w paradygmacie całkowitego błędu pomiaru

WYDAWNICTWO NAUKOWE

POZNAŃ 2015

ABSTRACT. Jabkowski Piotr, *Reprezentatywność badań reprezentatywnych. Analiza wybranych problemów metodologicznych oraz praktycznych w paradygmacie całkowitego błędu pomiaru* [Representativeness of a Representative Study: Analysis of Selected Methodological and Practical Problems within the Total Survey Error Paradigm]. Wydawnictwo Naukowe Uniwersytetu im. Adama Mickiewicza (Adam Mickiewicz University Press). Poznań 2015. Seria Socjologia nr 77. Pp. 339. ISBN 978-83-232-2887-5. ISSN 0554-8225. Text in Polish with a summary in English.

This monograph deals with the issue of the (non-)representativeness of research surveys and is situated within the context of methodological reflection on the quality of quantitative social research that mainly aims to identify errors related to the representativeness of a sample. The entire discussion is based on the theory of total survey error, which is widely recognized as the paradigm of quantitative research methodology. The author begins by describing the basic assumptions of this paradigm and analyzes errors that occur at different stages of the research process. The main part of this monograph focuses on four topics: (1) survey sampling frames, (2) sampling schemes, (3) errors resulting from an incomplete response rate, and (4) data weighing procedures. The main aim of this monograph was to analyze the methodological consequences of certain practical actions.

Piotr Jabkowski, Uniwersytet im. Adama Mickiewicza w Poznaniu, Wydział Nauk Społecznych, Instytut Socjologii, ul. Szamarzewskiego 89C, 60-568 Poznań, Poland

Recenzent: prof. dr hab. Franciszek Sztabiński

Praca dofinansowana przez Rektora Uniwersytetu im. Adama Mickiewicza w Poznaniu, Wydział Nauk Społecznych oraz Instytut Socjologii UAM

ISBN 978-83-232-2887-5
ISSN 0554-8225

Projekt okładki: Helena Oszmiańska-Napierała

Redaktor: Bożena Kapusta

Redaktor techniczny: Dorota Borowiak

WYDAWNICTWO NAUKOWE UNIWERSYTETU IM. ADAMA MICKIEWICZA W POZNANIU
UL. FREDRY 10, 61-701 POZNAŃ
www.press.amu.edu.pl
Sekretariat: tel. 61 829 46 46, faks 61 829 46 47, e-mail: wydnauk@amu.edu.pl
Dział sprzedaży: tel. 61 829 46 40, e-mail: press@amu.edu.pl

Wydanie I. Ark. wyd. 24,25. Ark. druk. 21,25.

DRUK I OPRAWA: EXPOL, WŁOCŁAWEK, UL. BRZESKA 4

Opisanie błędu świadczy o stopniu rozwoju nauki. Każde dochodzenie naukowe pozostaje podatne na błąd, znacznie lepiej zdawać sobie z tego sprawę, studiując potencjalne źródła błędów, dążąc do ich ograniczenia oraz do szacowania ich wielkości, niż być nieświadomym istnienia błędów. Nieświadomość błędu nie oznacza jego nieistnienia.

Herbert H. Hyman, 1954

Spis treści

Wprowadzenie

Książka ta traktuje o *reprezentatywności* sondaży. Ponieważ tytułowe pojęcie jest niezwykle szerokie i stanowi przedmiot ożywionej dyskusji metodologicznej, to owo dopełnienie badań sondażowych – przykuwające w tej pracy szczególną uwagę autora – wymaga już teraz doprecyzowania.

W ścisłym znaczeniu *reprezentatywny* to mający cechy jakiejś szerszej zbiorowości. W badaniach reprezentatywnych określenie to wiąże się silnie z pojęciem *próby badawczej*. Jeśli mówi się o reprezentatywnym sondażu, to ma się najczęściej na myśli badanie prowadzone na próbie składającej się z jednostek posiadających cechy populacji, co przy założonym współczynniku ufności oraz znanej wartości błędów daje możliwość uogólniania wyników pomiaru z próby na całą populację. Świetnie obrazuje to definicja *próby reprezentatywnej* przedstawiona w podręczniku *Podstawy statystyki dla socjologów* autorstwa Grzegorza Lissowskiego, Jacka Hamana oraz Mikołaja Jasińskiego (2008):

> Próba reprezentatywna dla danej populacji ze względu na badaną zmienną (lub zespół zmiennych) jest to próba, która dla badanej zmiennej (zespołu zmiennych) daje oszacowanie parametrów populacji zawierające się w granicach określonych przez wymagania w kwestii dokładności, wokół odpowiadających wartości w populacji, przy czym procedura, za pomocą której została wylosowana, daje takie reprezentatywne próby z prawdopodobieństwem określonym w przyjętym współczynniku ufności. (Lissowski i in. 2008: 511)

Ciekawą, choć mniej formalną interpretację reprezentatywności znaleźć można w monografii Jolanty Kowal (1998: 24–25), w której autorka zaproponowała trzy znaczenia tego pojęcia. Pierwsze z nich dotyczy zmiennej, a dokładniej mówiąc, obejmuje warunek zakładający, że w pomiarze reprezentatywnym muszą się pojawić wszystkie wartości analizowanej zmiennej, które rzeczywiście występują w populacji. W zasadzie mowa tutaj o reprezentacji typologicznej, a nie statystycznej, gdyż nie bierze się pod uwagę, z jaką częstością owe wartości powinny wystąpić. Kryterium to jest jednak nieścisłe, zwłaszcza że dla zmiennych ciągłych lub dyskretnych o nieskończonej liczbie wartości nigdy nie będzie się w stanie ich wszystkich uwzględnić. Dopiero drugie z zaproponowanych przez J. Kowal znaczeń *reprezentatywności* przybliża do

jakiegoś praktycznego zdefiniowania tego pojęcia. O próbie można powiedzieć jako o reprezentatywnej, jeżeli empiryczne rozkłady zmiennych (lub wartości estymatorów) odpowiadają – z określonym przybliżeniem – rzeczywistym rozkładom (lub wartościom parametrów) w całej populacji. Ten sposób ujmowania reprezentatywności jest zbieżny z propozycją Lissowskiego i in. (2008: 511). Wreszcie w trzecim znaczeniu o próbie można powiedzieć, że jest reprezentatywna, gdy ustalone na jej podstawie zależności pomiędzy zmiennymi odpowiadają rzeczywistym zależnościom występującym w populacji.

Doskonałym przykładem obrazującym wieloznaczność pojęcia *reprezentatywności* jest seria artykułów autorstwa Williama Kruskala oraz Fredericka Mostellera zamieszczonych w latach 1979–1980 w czterech numerach renomowanego czasopisma „International Statistical Review". Poświęcone zostały one analizie kontekstów, w jakich pojęcia *próbkowanie reprezentatywne* oraz *próba reprezentatywna* pojawiają się w literaturze popularnonaukowej (por. Kruskal i in. 1979a: 13–24), naukowej niestatystycznej (por. Kruskal i in. 1979b: 111–127) oraz statystycznej (por. Kruskal i in. 1979c: 245–265), jak i też rysowi historycznemu obu pojęć (por. Kruskal in. 1980: 169–195). Przeprowadzona w tych pracach rekonstrukcja sposobów rozumienia pojęcia „próby reprezentatywnej" umożliwia zidentyfikowanie dziewięciu najbardziej charakterystycznych jego interpretacji. Pięć z tych znaczeń, takich jak: (1) brak selektywności doboru jednostek (oryg. *absence of selective forces*), (2) miniatura populacji (oryg. *miniature of the population*), (3) dobór typowych przypadków (oryg. *typical or ideal cases*), (4) metoda doboru próby, posiadająca określone właściwości (oryg. *representative sampling as a specific sampling method*), i (5) metoda doboru umożliwiająca dobrą/dokładną estymację parametrów populacyjnych (oryg. *representative sampling as permitting good estimation*), zbliżone są do znaczeń, jakie obecnie nadaje się reprezentatywności.

Interesującą próbą (re)definicji pojęcia reprezentatywności (ograniczoną jednak do zjawiska niepełnej realizacji próby sondażowej) są też prace teoretyczne i empiryczne nad tak zwanym wskaźnikiem reprezentatywności (oryg. *Representativity Indicator* lub *R-Indicator*), prowadzone obecnie w ramach programu badawczego *Representativity Indicators for Survey Quality*. Autorzy tej dość oryginalnej koncepcji wiążą pojęcie reprezentatywności z wzorcem losowych braków danych oraz z prawdopodobieństwem uzyskania (udzielenia) odpowiedzi (*response probability*) (por. Schouten i in. 2009: 101–113; Schouten i in. 2011: 1–24; Schouten i in. 2012: 382–399; Luiten i in. 2013: 165–189). W mocnej wersji definicji reprezentatywności przyjmuje się, iż zbiór danych jest ściśle reprezentatywny, jeżeli prawdopodobieństwa udzielenia odpowiedzi pozostają jednakowe w obrębie wszystkich jednostek próby badawczej. W słabszej wersji zakłada się, iż próba jest reprezentatywna względem pewnej zmiennej (kategorialnej/skokowej), jeśli tylko przeciętne jednostkowe prawdopodobieństwa udzia-

łu w badaniu pozostają równe w ramach każdej z klas wartości takiej zmiennej. Innymi słowy, mocna wersja zakłada, że braki danych mają charakter całkowicie losowy w odniesieniu do wszystkich zmiennych, w słabej wersji reprezentatywności braki mogą być losowe dla pewnych zmiennych, a dla innych już nie.

Na potrzeby analiz prowadzonych w tej pracy przyjmę, iż reprezentatywność nie jest dychotomiczną cechą prób badawczych. Za niewłaściwe należy tym samym uznać mówienie, jakoby próba badawcza miała być reprezentatywna albo też niereprezentatywna. Po pierwsze, poziom reprezentatywności jest kwestią stopnia dokładności, z jaką próba badawcza pozwala szacować wartości parametrów. Po drugie, próba może być bardziej reprezentatywna z uwagi na pomiar pewnych zmiennych, a innych już nie. Stąd też Michael E. Davern (2008: 720––722) – autor hasła *próba reprezentatywna* zamieszczonego w *Encyclopedia of Survey Research Methods*, opracowanej pod redakcją Paula J. Lavrakasa (2008) – dowodzi, że jedną z czynności badawczych koniecznych do wykonania w trakcie realizacji badań sondażowych jest oszacowanie stopnia reprezentatywności próby. M.E. Davern daje też „praktyczne" rady, w jaki sposób osiągnąć odpowiedni poziom reprezentatywności. W jego opinii może być on zapewniony dzięki: (a) kompletnej liście danych, zawierającej informacje o wszystkich jednostkach w populacji, czyli możliwości wykorzystania takiego operatu losowania, który zapewnia każdej jednostce niezerowe (choć niekoniecznie równe) prawdopodobieństwo wyboru do próby, (b) zastosowaniu zrandomizowanego schematu doboru jednostek do próby badawczej, jak również (c) zgromadzeniu danych dla każdej bez wyjątku jednostki wylosowanej do próby. Niestety, w zdecydowanej większości sondaży udaje się jedynie spełnić (często znacznym nakładem działań) warunek losowego doboru jednostek do prób badawczych. Możliwość spełnienia dwóch pozostałych wymogów, (a) i (b), pozostaje niezwykle ograniczona z jednej strony przez błąd pokrycia populacji operatem losowania (*coverage error*), a z drugiej – przez błąd braku odpowiedzi (*nonresponse error*). Wynika z tego, że poziom reprezentatywności próby sondażowej obniżają pewne specyficzne typy błędów, a prowadzone w tym zakresie analizy skoncentrowane są głównie wokół takich kwestii jak: schematy doboru próby, pokrycie populacji operatem losowania, niedostępność jednostek (*unit nonresponse*) oraz braki w odpowiedziach na pewne pytania kwestionariuszowe (*item nonresponse*).

* * *

O reprezentatywności sondażowych prób badawczych powiedziano w literaturze już wiele. Nasuwa się zatem zupełnie naturalne, a przy tym kłopotliwe pytanie, czy w metodologii badań reprezentatywnych jest jeszcze miejsce na coś nowego lub, przynajmniej, czy da się doprecyzować coś, co nie zostało jeszcze do końca powiedziane.

Warto od razu zastrzec, że autor nie aspiruje do przedstawienia nowej metodologii badań reprezentatywnych. Główna uwaga skupiać się będzie przede wszystkim na styku teorii oraz empirii, a dokładniej na analizie metodologicznych konsekwencji wynikających z podejmowania pewnych działań praktycznych. Praca ta ma dostarczyć także narzędzi do oceny poziomu reprezentatywności próby. Takie są dwa pierwsze z jej podstawowych celów. Szczególnego znaczenia nabierze krytyczna analiza pewnych procedur badawczych oraz postbadawczych, których wykorzystywanie uważano dotąd za oczywiste. Wykazane zostanie, że wiedza o mechanizmach stojących za reprezentatywnością prób sondażowych nie ma prostego przełożenia na praktykę lub, inaczej, że rzeczywistość badawcza pozostaje dużo bardziej złożona, niż przewiduje to teoria próbkowania reprezentatywnego. Wypełnienie sformułowanego w ten sposób zadania wymagać będzie – rzecz jasna – pokazania, jaki jest stan wiedzy w tym zakresie, a także – z czego wynikają oraz jakie są konsekwencje komplikacji związanych z przeniesieniem tej wiedzy do świata praktyki badawczej. Jest to tym samym trzeci cel tej pracy. Zresztą to właśnie napięcia pomiędzy teorią i empirią pozostają obecnie obszarem intensywnych dociekań wielu badaczy.

Skoro powiedziane zostało już, jakie są cele tej książki, można wskazać, czego będzie ona dotyczyć. Przedmiotem tej pracy jest metoda reprezentatywna, czyli metoda częściowego badania zbiorowości generalnej, oparta na pomiarze jednostek dobranych – z populacji do próby – w sposób losowy. Ponieważ pojęcie metody reprezentatywnej, podobnie jak i samej reprezentatywności, jest niezwykle szerokie, konieczne wydaje się dalsze doprecyzowanie zakresu prowadzonych tutaj analiz.

W pierwszej kolejności trzeba zatem wskazać, że studia nad reprezentatywnością prób badawczych ograniczone zostaną do badań o charakterze surveyowym, te zaś do tych technik sondażowych, w których pozyskiwanie danych odbywa się w drodze interrogacji, czyli poprzez prowadzenie osobistych – tradycyjnych lub wspomaganych komputerowo – standaryzowanych wywiadów kwestionariuszowych. A zatem całkowicie poza polem zainteresowania pozostaną wszystkie techniki ankietowe, a także – w zasadzie – te spośród technik wywiadów (np. *CATI*), w których, co do istoty, nie dochodzi do bezpośredniego kontaktu badacza (lub jego przedstawiciela, tj. ankietera) z osobą potencjalnego respondenta. Za podjęciem takiej decyzji przemawia fakt, iż techniki ankietowe oraz techniki wywiadów telefonicznych (w zestawieniu z technikami wywiadów prowadzonych bezpośrednio) mają na tyle odmienną specyfikę oraz swoiste problemy natury praktycznej i metodologicznej, że stanowią – same w sobie – przedmiot całkowicie odrębnych analiz metodologicznych.

Prowadzone w tej książce analizy ograniczone zostaną także do badań sondażowych o charakterze naukowym. Nie oznacza to bynajmniej, że w jakimś

stopniu praca ta próbuje deprecjonować surveye komercyjne i przeciwstawia je badaniom akademickim. Takie wartościowanie byłoby nie tylko krzywdzące (i to zarówno dla praktyków, jak i naukowców), ale przede wszystkim okazałoby się błędne. Nie jest przecież niczym nadzwyczajnym korzystanie – w badaniach naukowych – z doświadczeń badań marketingowych; zresztą transfer doświadczeń oraz wiedzy przebiega też w odwrotnym kierunku. Chodzi przede wszystkim o to, że badania naukowe mają odmienne cele, różne są też kryteria stosowane do oceny sondażu. O jakości badań naukowych decyduje bowiem nie tylko jego ostateczny rezultat oraz stojąca za nimi wartość użytkowa, ale przede wszystkim proces badawczy oraz prawidłowość jego przebiegu. W pracy tej na drugi plan schodzą również kwestie *kosztochłonności* badań. Ten celowy zabieg autora ma uwypuklić problematykę *jakości* próby reprezentatywnej. Nie zmienia to faktu, że przebieg procesu badawczego oraz podejmowane działania terenowe pozostają zawsze wynikiem konsensusu pomiędzy możliwą do osiągnięcia jakością badania oraz wielkością budżetu.

Trzecie ograniczenie dotyczy typu analizowanej próby badawczej. Przedmiotem rozważań uczynione zostaną wyłącznie tzw. próby realizowane do wyczerpania (por. Sawiński 2005: 84–85). Nie będzie tutaj mowy o próbach rezerwowych, w tym o próbach dopuszczających zamianę jednostek niedostępnych (*non-respondents*) na inne osoby dobierane celowo lub kwotowo. W badaniach (nie tylko) akademickich takie postępowanie nie ma żadnego merytorycznego uzasadnienia (por. Jabkowski 2007). Oczywiście praktyka komercyjnych instytutów badawczych jest często odmienna, nie zmienia to faktu, że działania polegające na doborze respondentów „zastępczych" w żaden sposób nie poprawiają jakości wnioskowania indukcyjnego. Należy to wyraźnie podkreślić. Odbiorcami tej książki będą bowiem nie tylko studenci (w jej literaturowej części) oraz metodologowie badań reprezentatywnych, ale też przedstawiciele instytucji badawczych. Ci ostatni mogliby odnieść wrażenie istnienia poważnych luk w tej książce. Co więcej, praca ta dotyczy wyłącznie takich schematów doboru prób badawczych, w których jednostki populacji (osoby) dobierane są w oparciu o procedury losowe lub quasi-losowe. A zatem poza obszarem zainteresowania pozostają schematy doboru celowego, kwotowego i systematycznego, w tym procedury ustalonej ścieżki (tzw. *random route*), a zwłaszcza związana z tą metodą praktyka polegająca na pozyskiwaniu respondentów charakteryzujących się określonymi cechami demograficznymi (czyli działania zmierzające do wypełniania, w próbie zrealizowanej, założonych *a priori* rozkładów pewnych populacyjnych cech jednostek). Procedury *random route* są co prawda uzasadnione oraz stosowane w doborze budynków mieszkalnych (z wylosowanych punktów adresowych) lub gospodarstw domowych (z wylosowanych budynków) – i w takim kontekście będą w pracy przywoływane – mają jednak wąt-

pliwą wartość (poza ułatwieniem terenowej realizacji próby) w losowym i reprezentatywnym doborze osób do prób badawczych.

Czwarte ograniczenie ma charakter praktyczny i związane jest z prowadzonymi w tej pracy studiami empirycznymi. Bazą źródłową większości analiz uczynione będzie repozytorium Europejskiego Sondażu Społecznego. Dla każdego metodologa badań sondażowych projekt ten jest niezwykle inspirujący i wartościowy. Istotne jest przede wszystkim to, że poza zebraniem kluczowych dla Europy danych społecznych, w programie ESS-u realizuje się również ważne cele metodologiczne. Projekt ten pozostaje, bez wątpienia, przedsięwzięciem o niezwykle wysokiej „kulturze" metodologicznej, co nie oznacza, że osobista fascynacja autora Europejskim Sondażem Społecznym uniemożliwiać będzie krytyczne spojrzenie na jego dokonania.

* * *

Książka składa się z pięciu rozdziałów przyporządkowanych do dwóch umownych części.

Część pierwsza – obejmująca rozdział pierwszy oraz drugi – wprowadza paradygmat całkowitego błędu pomiaru badań sondażowych (*Total Survey Error*) oraz szczegółowo charakteryzuje źródła błędów konstytuujących błąd całkowity. Oba rozdziały mają dla tej pracy znaczenie fundamentalne, wyznaczają bowiem zasięg prowadzonych studiów teoretycznych oraz empirycznych.

Rozdział pierwszy otwiera charakterystyka paradygmatu całkowitego błędu pomiaru badań sondażowych oraz dyskusja nad kontrowersjami związanymi z określaniem wielkości takiego błędu jako różnicy pomiędzy wartością estymatora jakiegoś konkretnego parametru uzyskaną w badaniu oraz jego „prawdziwą" wartością w całej populacji. Przeprowadzone studia literaturowe doprowadzą do wyróżnienia „statystycznego" oraz „psychometrycznego" podejścia do oceny jakości pomiaru, czego konsekwencją będzie poszukiwanie zewnętrznych oraz wewnętrznych standardów oceny sondaży. Kolejna część rozdziału pierwszego poświęcona będzie z kolei identyfikacji głównych źródeł błędów badań reprezentatywnych. Wprowadzone zostanie również rozróżnienie na losowy oraz systematyczny komponent błędu całkowitego. Wprawdzie studia literaturowe będą miały w dużej mierze charakter sprawozdawczy, jednak pozwolą wyróżnić typowy „zestaw" błędów identyfikowanych przez zdecydowaną większość metodologów, którzy odwołują się do paradygmatu całkowitego błędu pomiaru. W czwartej części rozdziału pierwszego przedstawiona zostanie propozycja klasyfikacji źródeł błędów, wprowadzająca podział na błędy obniżające poziom reprezentatywności próby, a także błędy oddziałujące na jakość pomiaru oraz błędy przetwarzania danych wynikowych. Ostatnia

część rozdziału pierwszego poświęcona będzie z kolei miernikowi liczbowemu całkowitego błędu pomiaru badań sondażowych. Zaproponowana zostanie modyfikacja postaci klasycznego estymatora błędu całkowitego, polegająca na odmiennym sposobie definiowania losowego komponentu tego błędu, w którym wielkość wariancji estymatorów pewnych parametrów (w danej próbie) odniesiona będzie do teoretycznej wariancji estymatorów tych parametrów w prostej próbie losowej. Rozdział podsumują uwagi końcowe poświęcone ograniczeniom oraz korzyściom płynącym z analizy jakości danych sondażowych w perspektywie paradygmatu błędu całkowitego.

Rozdział drugi uszczegóławia problematykę błędów popełnianych w badaniach sondażowych. Stanowi tym samym uzupełnienie kwestii poruszanych w rozdziale pierwszym. Struktura rozdziału odpowiada wprowadzonej klasyfikacji źródeł błędów. A zatem, w części pierwszej zdefiniowane zostaną błędy oddziałujące na poziom reprezentatywności próby, w części drugiej – błędy związane z pomiarem, a w części trzeciej – błędy postbadawczego przetwarzania zbiorów danych wynikowych. Wprawdzie druga oraz trzecia część tego rozdziału wykraczać będzie w znacznej mierze poza tematykę tej pracy, to jednak zdefiniuje ona te wszystkie źródła błędów, na które zwrócona zostanie uwaga w rozdziale pierwszym, w ramach charakterystyki paradygmatu całkowitego błędu pomiaru badań sondażowych. Głównym celem tego rozdziału będzie jednak zebranie oraz usystematyzowanie literatury przedmiotu, a także omówienie najważniejszych dylematów metodologicznych oraz praktycznych związanych z oceną jakości pomiaru w perspektywie całkowitego błędu pomiaru badań sondażowych.

W drugiej części książki – obejmującej rozdział trzeci, czwarty oraz piąty – rozpatrywane będą już szczegółowe problemy związane z oceną poziomu reprezentatywności sondażowych prób badawczych. Wprawdzie układ tych rozdziałów odpowiadać będzie kolejnym działaniom badawczym mającym na celu zdefiniowanie badanej populacji, wybór operatu doboru próby, ustalenie schematu losowania jednostek, terenową realizację próby oraz postrealizacyjne ważenie danych, to jednak zakres analizowanych problemów nie będzie przebiegać wzdłuż, ale w poprzek owych etapów postępowania badawczego. Analizy uwzględniać będą bowiem wzajemne oddziaływanie poszczególnych działań badawczych oraz postbadawczych na poziom reprezentatywności próby. Taki układ analiz narzucać będzie – przyjęty w pierwszej części książki – paradygmat całkowitego błędu pomiaru. Jego niezwykłą zaletą jest bowiem ukazanie konieczności spojrzenia na proces badawczy jak na układ naczyń połączonych. Stwierdzenie tego faktu wydaje się być może banalne, wystarczy jednak przeprowadzić pobieżną kwerendę literaturową, by ukazać, w jak wielu przypadkach uwaga metodologów koncentruje się na kwestiach niezwykle szczegółowych, skupionych na jednym problemie, a jak niewielką uwagę poświęca się

szerszemu kontekstowi prowadzonych studiów i analiz. Bodaj najbardziej jaskrawym przykładem takiego „aptekarskiego" podejścia do jakości sondaży są niektóre prace dedykowane fenomenowi niedostępności jednostek. Ponieważ już dawno stwierdzono, że osoby niedostępne różnią się (mogą się różnić) od osób przebadanych, to wobec faktu zmniejszających się – z roku na rok – odsetków realizacji próby, metodologowie badań sondażowych koncentrują się przede wszystkim na tym, jakie szczegółowe działania należy podjąć, aby zmaksymalizować wskaźniki realizacji próby i ograniczyć tym samym maksymalną wielkość błędu związanego z niepełną jej realizacją. Chociaż zagadnienia te są ważne dla osiągnięcia odpowiedniej/zadowalającej jakości badania, to jednak w wielu takich opracowaniach brakuje ogólnej refleksji nad tym, czy wskaźnik realizacji próby ma w ogóle jakieś znaczące przełożenie na wielkość błędów braków danych, lub też, czy procedury na rzecz maksymalizacji wskaźnika realizacji próby nie przekładają się – wbrew oczekiwaniom – na przyrost wielkości błędów badań sondażowych. Ten prosty przykład pokazuje, że podstawową zaletą odwołania się w tej książce do paradygmatu błędu całkowitego będzie wymuszenie refleksji nad tym, czy działania zmierzające do ograniczenia pewnych specyficznych źródeł błędów (na jakie zawsze napotyka się w trakcie doboru, terenowej realizacji oraz postbadawczej „obróbki" wyników pomiaru próby reprezentatywnej) nie mają czasami negatywnego przełożenia na inne komponenty procesu badawczego. W zasadzie, poza tę prostą ideę autor tej książki nie będzie wykraczał.

Rozdział trzeci poświęcony będzie operatom doboru sondażowych prób badawczych. Poprzez odwołanie do koncepcji operatu „idealnego", a także dzięki charakterystyce relacji zachodzących pomiędzy populacją docelową (*target population*) oraz populacją pokrytą operatem losowania (*frame population*), wyprowadzone zostaną cztery główne klasy błędów mających swoje źródło w ułomnościach rejestrów wykorzystywanych do losowania prób reprezentatywnych. Omówione będą procedury służące poprawie jakości operatów losowania próby. Uwaga skupiać się będzie przy tym zarówno na charakterystyce założeń teoretycznych oraz metodologicznych stojących u podstaw owych procedur, jak również na weryfikacji ich efektywności. Studia literaturowe – odwołujące się do analiz empirycznych – uwidocznią przy tym, że korzyści wynikające z redukcji systematycznego błędu niepełnego pokrycia niwelowane są zazwyczaj przez błędy pomiarowe.

O ile zadaniem pierwszej części rozdziału trzeciego będzie zebranie oraz uporządkowanie literatury poświęconej błędom operatów doboru prób badawczych oraz procedurom służącym ograniczaniu ich wielkości, o tyle w części drugiej zaprezentowane zostaną wyniki autorskich analiz dedykowanych empirycznej egzemplifikacji wybranych problemów praktycznych. Bazą analityczną uczynione zostaną repozytoria Europejskiego Sondażu Społecznego (ESS). Ze-

stawienie krajowych populacji docelowych oraz operatów wykorzystywanych do losowania prób badawczych uwidoczni skalę problemów, na jakie narażony jest dobór prób sondażowych w sytuacji niedostępności (lub niedostatecznej jakości) operatów imiennych. Zaprezentowane analizy pozwolą przypisać wszystkie kraje uczestniczące w ESS do pięciu kategorii państw wyróżnionych poprzez skrzyżowanie poziomu agregacji danych w operatach doboru próby (operaty imienne lub zespołowe) oraz typu wylosowanej próby badawczej (imienna, gospodarstw domowych oraz przestrzenna/budynków mieszkalnych). Podjęta zostanie problematyka wewnątrzzespołowej selekcji jednostek indywidualnych z operatów zespołowych, a także zagadnienia niepełnego wewnątrzzespołowego pokrycia populacji. W rozdziale prowadzone będą również rozważania nad możliwością wykorzystania polskich rejestrów administracji publicznej (PESEL oraz TERYT) w losowaniu prób reprezentatywnych.

Rozdział czwarty podejmuje problematykę schematów doboru sondażowych prób badawczych. Punktem odniesienia pozostanie, tym razem, schemat losowania próby w sposób prosty, wraz ze swoimi charakterystykami definicyjnymi, tj. losowaniem nieograniczonym z całej populacji, losowaniem indywidualnym, losowaniem z równymi prawdopodobieństwami selekcji oraz losowaniem jednostopniowym. O ile wprowadzone w rozdziale trzecim rozróżnienie na próby imienne oraz adresowe związane będzie przede wszystkim z uchyleniem warunku jednostopniowości doboru próby, o tyle w rozdziale czwartym uwaga skoncentruje się na trzech pierwszych charakterystykach próby prostej. Przedstawione zostaną zatem schematy losowania stratyfikacyjnego, zespołowego oraz doboru próby ze zróżnicowanymi prawdopodobieństwami selekcji jednostek. W odniesieniu do każdego z tych schematów przeanalizowane oraz opisane zostaną czynniki warunkujące ich mniejszą (lub większą) efektywność (w porównaniu do efektywności schematu losowania prostego), zdefiniowane będą mierniki służące szacowaniu efektywności określonych schematów losowania, a także podjęte zostaną zagadnienia o charakterze praktycznym. Druga część rozdziału czwartego poświęcona będzie z kolei komplikacjom na jakie napotyka próba przeprowadzenia empirycznej oceny efektywności przyjętego schematu doboru próby. Punktem wyjścia będzie charakterystyka kryteriów warunkujących możliwość zastosowania określonych estymatorów efektywności losowania próby. Głównym celem będą natomiast studia nad konsekwencjami wynikającymi z analizy schematu losowania opartej na uproszczonych miernikach oceny jego efektywności.

W ostatnim rozdziale omówione zostaną wybrane problemy niepełnej realizacji sondażowej próby badawczej. Oczywiście zakres zagadnień, jakie przy tej okazji można byłoby poruszyć, jest niezwykle szeroki. Celem rozdziału nie będzie jednak mówienie o wszystkim, ale przedstawienie spójnej metodologii po-

zwalającej na ocenę reprezentatywności sondażowej próby badawczej w świetle jej niepełnej terenowej realizacji.

Rozdział piąty podzielony zostanie na trzy części. W pierwszej rozważone będą kwestie podstawowe, związane z postrealizacyjną klasyfikacją jednostek wylosowanych do sondażowej próby badawczej. Podążając za postulatem konieczności wypracowania jednolitych standardów obliczania wartości wskaźników realizacji próby, przyjęty zostanie podział osób wylosowanych do badania na zbiór: (1) respondentów, (2) jednostek niedostępnych (w tym na osoby niedostępne z powodu braku kontaktu, odmowę lub inny powód niezrealizowania wywiadu), (3) jednostek o nieustalonym statusie przynależności do populacji docelowej, a także (4) jednostek nienależących do populacji. Studia literaturowe oraz autorskie analizy empiryczne oparte na wynikach Europejskiego Sondażu Społecznego pokażą, że podział wprowadzony w warstwie jednostek niedostępnych jest uzasadniony merytoryczne i wynika z tego, że mechanizmy kształtujące gotowość jednostek do udziału w badaniu mają odmienną naturę od mechanizmów warunkujących możliwość dotarcia do wylosowanych osób.

W drugiej części rozdziału piątego scharakteryzowane zostaną schematy terenowej realizacji prób adresowych oraz imiennych. Problematyka ekwiwalentności prób dobieranych z operatów jednostkowych oraz zespołowych podjęta będzie także w rozdziale trzecim, o ile jednak uwaga skupi się wówczas na przełożeniu – charakterystycznego dla prób zespołowych – procesu wewnątrzzespołowej selekcji jednostek na błędy niepełnego (lub nadmiarowego) pokrycia jednostek populacji docelowej, to celem analiz zaprezentowanych w rozdziale piątym będzie znalezienie odpowiedzi na pytanie, czy wykorzystanie odmiennych typów operatów przekłada się w jakiś znaczący sposób na wskaźniki realizacji próby oraz postbadawczą strukturę zbioru respondentów oraz osób niedostępnych.

Ostatnia część rozdziału piątego poświęcona będzie ocenie reprezentatywności sondażowej próby badawczej w świetle jej niepełnej realizacji. W pierwszej kolejności zostaną scharakteryzowane założenia dwóch paradygmatów – deterministycznego oraz probabilistycznego – w ramach których rozważać można błędy wynikające z niedostępności jednostek wylosowanych do badania. W tej części rozdziału opisana zostanie idea tak zwanego wskaźnika reprezentatywności zbioru odpowiedzi oraz możliwość wykorzystania tego miernika w ocenie jakości terenowej realizacji sondażowej próby badawczej. Studia literaturowe oraz analizy empiryczne uwidocznią, że jednym z największych wyzwań metodologicznych i praktycznych związanych z oceną reprezentatywności próby w duchu założeń paradygmatu probabilistycznego staje się identyfikacja mechanizmu kształtującego charakter niedostępności jednostek. Omó-

wione zostaną także procedury ważenia danych oparte na oszacowaniach jednostkowych skłonności do udziału w badaniu. Studia o charakterze teoretycznym zobrazowane zostaną autorskimi analizami empirycznymi.

* * *

Rozważania zawarte w tej książce mają przede wszystkim wymiar użytkowy. Głównym zadaniem metodologii każdej nauki jest dostarczenie instrumentarium do – uprawomocnionego w danej dyscyplinie – badania rzeczywistości empirycznej. W pierwszej kolejności praca ta ma więc zachęcić do krytycznego spojrzenia na wszystkie działania badawcze oraz postbadawcze, które są podejmowane w trakcie realizacji badań reprezentatywnych. Zadaniem tej książki jest także analiza metodologiczna ukierunkowana na identyfikację czynników oddziałujących na reprezentatywność próby. Po trzecie wreszcie – można by przy tej okazji wypowiedzieć niemal sakramentalne *last but not least* – praca ta ma dostarczyć narzędzi empirycznej weryfikacji poziomu reprezentatywności sondażowej próby badawczej. To, w jakim stopniu cele te udało się osiągnąć, a także, na ile praca ta redukuje wspomniane wcześniej napięcie pomiędzy teorią oraz empirią, trzeba – z pokorą – pozostawić ocenie Czytelnika.

ROZDZIAŁ I

Całkowity błąd pomiaru badań sondażowych

Celem rozdziału jest scharakteryzowanie założeń stojących u podstaw paradygmatu całkowitego błędu pomiaru badań sondażowych. Rozdział ma w znacznej mierze charakter sprawozdawczy i porządkujący literaturę przedmiotu, pozwala również na ukierunkowanie prowadzonych dalej studiów na zagadnienia powiązane w sposób jednoznaczny z problematyką reprezentatywności sondażowych prób badawczych. Co prawda, paradygmat całkowitego błędu pomiaru nie jest koncepcją nową, nie znajduje też bezpośredniego przełożenia na praktykę badawczą, nie jest nawet jedyną (i być może nie jest najlepszą) perspektywą, w świetle której rozważać można jakość prowadzonych badań, niemniej jednak uświadamia badaczom, że na proces pomiaru nie można patrzeć fragmentarycznie, przez pryzmat pojedynczych działań służących ograniczaniu lub eliminowaniu pewnych specyficznych źródeł błędów, ale tylko całościowo, z uwzględnieniem wzajemnego oddziaływania wszystkich przedsięwzięć badawczych oraz postbadawczych na jakość otrzymywanych wyników.

Pierwsza część rozdziału skoncentrowana jest na przedstawieniu definicji błędu całkowitego. Poprzez odniesienie do statystycznej teorii próbkowania błąd taki określony zostanie jako różnica pomiędzy wartością estymatora uzyskaną w badaniu oraz „prawdziwą" wartością parametru w całej populacji. Omówione będą również kontrowersje związane z posługiwaniem się pojęciem wartości „prawdziwej" w definiowaniu błędu całkowitego, co doprowadzi do wyróżnienia „statystycznego" oraz „psychometrycznego" podejścia do oceny jakości pomiaru. W drugiej części rozdziału zidentyfikowane zostaną źródła błędów posiadające decydujący wpływ na jakość prowadzonego badania. W trzeciej części zaproponowane będzie usystematyzowanie oraz klasyfikacja źródeł błędów. Z kolei w ostatniej części rozdziału podana zostanie formalna definicja miernika błędu całkowitego (tak zwany *błąd średniokwadratowy*) oraz omówiony będzie sposób empirycznego szacowania wielkości tego współczyn-

nika. Autor zaproponuje modyfikację estymatora błędu średniokwadratowego polegającą na odmiennym – od pojawiającego się w literaturze – sposobie definiowania losowego komponentu błędu. Rozdział podsumują rozważania nad ograniczeniami oraz korzyściami wynikającymi z wykorzystania paradygmatu całkowitego błędu pomiaru w ocenie jakości danych sondażowych.

I.1. Paradygmat całkowitego błędu pomiaru – definicja błędu oraz związane z nią kontrowersje

Zdecydowana większość badaczy zajmujących się paradygmatem *całkowitego błędu pomiaru* (ang. *Total Survey Error*) – w skrócie TSE – definiuje błąd całkowity jako różnicę pomiędzy oszacowaniem pewnego parametru w próbie oraz jego rzeczywistą wartością w całej populacji. Z koncepcją TSE wiązane są takie komponenty procesu zbierania danych jak: populacja, próba, operat losowania, terenowa realizacja próby (w tym niedostępność pewnych jednostek), wybór estymatorów, operacjonalizacja pojęć, uzyskiwanie odpowiedzi (pomiar), wyznaczanie wartości statystyk punktowych i przedziałowych, czy też wreszcie opracowywanie wyników badań. Innymi słowy, błąd może się pojawić na każdym etapie realizacji procesu badawczego (por. Groves i in. 2010: 850). Należy przy tym podkreślić, że choć wielu metodologów zajmujących się problematyką błędów badań reprezentatywnych różni się w szczegółach, jeśli chodzi o identyfikację źródeł błędów, a nawet sposoby wyznaczania ich wielkości, to jednocześnie panuje względna zgoda co do definiowania samego pojęcia TSE. W większości prac eksponowane są takie składniki konstytuujące całkowity błąd jak: (1) końcowy wynik pomiaru, (2) różnica, (3) wartość prawdziwa. Taki sposób rozumienia całkowitego błędu pomiaru okazuje się na tyle powszechny w literaturze przedmiotu, że wielu badaczy podejmujących problematykę jakości sondaży nie przedstawia w ogóle formalnej definicji pojęcia TSE, uznając je najczęściej za oczywiste. Ujmując to inaczej, metodologowie i praktycy skupiają się bardziej na pewnych źródłach błędów niż na skumulowanym błędzie jako całkowitym efekcie poszczególnych uchybień badawczych.

Dla porządku należy przywołać jednak kilka typowych definicji pojęcia TSE pojawiających się w literaturze badań reprezentatywnych. W pierwszej kolejności warto przytoczyć definicję Herberta F. Weisberga (2005) przedstawioną w monografii *The Total Survey Error Approach. A Guide to the New Science of Survey Research*. Autor ten, analizując różne typy błędów, z jakimi badacz ma do czynienia w trakcie realizacji pomiaru reprezentatywnego, wyprowadza jedną z charakterystycznych definicji TSE. W drugim rozdziale przywołanej pracy odnaleźć można następujące stwierdzenie:

> […] ‘błąd’ jest zazwyczaj uważany za synonim ‘pomyłki’, jednakże w kontekście badań surveyowych odnosi się on do różnicy pomiędzy uzyskaną wartością oraz wartością prawdziwą […] w populacji. (Weisberg 2005: 18)

Analogiczny sposób definiowania TSE znajduje się też w drugim rozdziale podręcznika *Introduction to Survey Quality*. Paul P. Biemer oraz Lars E. Lyberg (2003) wskazują w nim, że:

> całkowity błąd pomiaru jest różnicą pomiędzy [wartością – P.J.] estymatora i prawdziwą wartością parametru w populacji. (Biemer i in. 2003: 35)

W podobnej formie definicja ta pojawia się także w tekście P. Biemera (2010a), otwierającym specjalny numer czasopisma „Public Opinion Quarterly" poświęcony tematyce TSE. Całkowity błąd pomiaru przedstawiony został w tym artykule jako skumulowany efekt wielu różnych źródeł błędów:

> Całkowity błąd pomiaru (TSE) jest pojęciem kumulującym wszystkie błędy, które mogą wynikać z projektowania, zbierania, opracowywania i analizy danych sondażowych. W takim ujęciu, błąd pomiaru w badaniach surveyowych definiowany jest jako odchylenie odpowiedzi uzyskanych podczas pomiaru od ich prawdziwych wartości. […] Błędy pomiaru wyrastają z niedoskonałości operatu losowania, procesu próbkowania, prowadzenia wywiadów, zachowań ankieterów oraz respondentów, braków danych, kodowania, kategoryzowania oraz opracowywania wyników. (Biemer 2010a: 817–818)

Przegląd literatury ukazuje zatem, że *całkowity błąd pomiaru* odnosi się do różnicy pomiędzy wartością estymatora pewnego parametru ustaloną na podstawie badania oraz jego rzeczywistą wartością w całej populacji. Jeżeli zatem przez θ oznaczy się prawdziwą (najczęściej niestety nieznaną) wartość pewnego parametru, natomiast przez $\hat{\theta}$ określi się jego estymator, to TSE można zdefiniować za pomocą formuły:

$$TSE(\hat{\theta}) \stackrel{\text{def}}{=} \hat{\theta} - \theta. \tag{I.1.}$$

Zauważyć należy jednak, że taki sposób ujmowania całkowitego błędu pomiaru okazuje się kłopotliwy, bowiem ustalenie wartości błędu wymaga wiedzy o wielkości parametru w całej populacji. W praktyce udaje się jedynie oszacować wartość błędu całkowitego poprzez jego dekompozycję na składniki odpowiadające pewnym klasom lub poszczególnym źródłom błędów. Doskonałym tego przykładem jest definiowanie całkowitego błędu pomiaru jako sumy błędów losowych (*random*) oraz systematycznych (*bias*), czyli takich komponentów TSE, które oddziałują z jednej strony na precyzję estymatorów, z drugiej zaś na ich dokładność. Obie klasy błędów wykorzystuje się zresztą do definiowania mierników liczbowych TSE (por. Biemer 2010b: 45; Alwin 2007: 53; Biemer i in. 2003: 37; Groves 1989: 38).

Definiowanie całkowitego błędu pomiaru poprzez różnicę pomiędzy wartością estymatora jakiegoś konkretnego parametru oraz jego wartością „prawdziwą" prowadzi do kontrowersji związanych nie tylko z tym, w jakich przypadkach daje się taką prawdziwą wartość wyznaczyć, ale także – czym ona w ogóle miałaby być. Problematyka ta podejmowana była już wielokrotnie, zazwyczaj w ramach opisu różnych sposobów definiowania błędów pomiarowych w badaniach psychologicznych, socjologicznych oraz ekonomicznych. Najbardziej interesujące wydaje się przy tym rozróżnienie wprowadzone przez R. Grovesa (1989: 9–10; 18–22), wskazujące na odmienność statystycznego oraz psychometrycznego podejścia do problematyki pomiaru. Abstrahując od trafności użycia obu pojęć, należy wskazać, że funkcjonują one także w polskiej tradycji metodologicznej, o czym wspomniał ostatnio Franciszek Sztabiński (2011: 45–47) w książce poświęconej ocenie jakości danych w badaniach surveyowych. W pierwszym rozdziale tej publikacji, w ramach wykładu poświęconego wewnętrznej walidacji badań reprezentatywnych poprzez ocenę jakości przeprowadzonego pomiaru, odnaleźć można następującą konstatację:

> W ujęciu statystycznym przyjmuje się założenie o istnieniu wartości prawdziwej. Tym samym przez błąd pomiaru rozumie się dowolne odchylenie (różnicę) wyników badania od wartości 'prawdziwej', która jest przedmiotem badania. […] Z kolei w ujęciu psychologicznym przyjmuje się, że wartość prawdziwa nie istnieje (istnieje co najwyżej wartość 'ukryta'), a błąd jest 'zanieczyszczeniem' wyniku pomiaru, spowodowanym wpływem jakiegoś czynnika. (Sztabiński, F. 2011: 45)

Na założeniu o istnieniu wartości „ukrytej" opierają się zresztą dość powszechne w psychologii analizy zmierzające do oszacowania jej wartości poprzez ustalenie wartości innych zmiennych obserwowalnych.

Zauważyć można, iż niezwykle symptomatyczne w statystycznym definiowaniu całkowitego błędu pomiaru jest branie w cudzysłów przymiotnika *prawdziwa* w odniesieniu do słowa *wartość*. Ponieważ w praktyce wyznaczenie wartości prawdziwej jest bardzo trudne, to wyrażenie „prawdziwa" okazuje się w gruncie rzeczy (poza nielicznymi wyjątkami) umowne. Nawet jeżeli w sensie formalnym, w odniesieniu do wielu parametrów, taka prawdziwa wartość istnieje, to jej ustalenie wydaje się niemożliwe. Wysiłek badawczy nie koncentruje się zatem na tym, by ustalić, jaka jest wartość całkowitego błędu pomiaru (bez znajomości wartości parametru nie da się jej wyznaczyć), lecz na tym, aby kontrolować (lub przynajmniej próbować ograniczać) wpływ tych wszystkich źródeł błędów, które mogą w znacznym stopniu wpłynąć na jakość (reprezentatywność) sondaży oraz koszty prowadzonych badań[1].

---

[1] Należy zresztą wspomnieć, że pojęcie TSE pojawia się w literaturze dość często w połączeniu z *jakością* oraz *kosztami* badań. Dla przykładu, w przywoływanym już wcześniej 74. numerze

Powracając do rozróżnienia na statystyczne oraz psychometryczne rozumienie wartości prawdziwej, warto odwołać się ponownie do koncepcji wyłożonej przez R. Grovesa w książce *Survey Errors and Survey Costs*. W pierwszym rozdziale tej publikacji autor wprowadza rozróżnienie terminologiczne ułatwiające klasyfikację omawianych przez siebie źródeł błędów. Prowadzi również rozważania nad istotą pojęcia „wartość prawdziwa", ukazując, iż:

> W niektórych sytuacjach zakłada się, że istnieje obserwowalna wartość prawdziwa pewnych statystyk [parametrów – P.J.] w populacji oraz zgodna z prawdą odpowiedź respondenta wybranego do próby badawczej. […] To domniemane istnienie obserwowalnej wartości prawdziwej odróżnia statystyczny punkt widzenia od perspektywy psychometrycznej […]. Istnieją uzasadnione powody dla takiego rozróżnienia. Pomiar psychologiczny dotyczy zazwyczaj postaw, które nie mogą być zaobserwowane przez nikogo poza samym respondentem. (Groves 1989: 9)

Takie rozróżnienie, obecne w identycznej formie także u F. Sztabińskiego (2011: 45), pozwala przyjąć, że w odniesieniu do statystyk opisujących stan faktyczny, np. cech społeczno-demograficznych respondentów, można mówić o istnieniu wartości prawdziwej, a także zgodnej z prawdą odpowiedzi respondentów, natomiast w odniesieniu do pytań o opinie, postawy, przekonania itd. uznać można, że wartość prawdziwa nie istnieje, istnieją jedynie odpowiedzi respondentów zgodne z ich wyobrażeniami o przedmiocie badania. Terminologia wykorzystana przez Grovesa do charakterystyki podejścia psychometrycznego obejmuje dwa pojęcia bardzo dobrze znane w literaturze poświęconej zagadnieniom konceptualizacji i operacjonalizacji problematyki badawczej,

---

czasopisma „Public Opinion Quarterly", poświęconym tematyce TSE, R. Groves i L. Lyberg definiują miarę liczbową całkowitego błędu pomiaru, podkreślając, iż „[…] może być ona rozpatrywana jako wskaźnik jakości danych" (Groves i in. 2010: 850). Ciekawą analizę koncepcji TSE w świetle jakości pomiaru oraz kosztów badań odnaleźć można także w artykułach P. Biemera (2010a: 818–819, 2010b: 28–30). Odwołując się bezpośrednio do pracy Richarda Plateka oraz Carla-Erika Särndala (2001: 1–20), jak również do znanej książki Grovesa (1989), P. Biemer ukazuje w obu pracach, że identyfikacja głównych źródeł błędów badań sondażowych jest niezbędnym elementem służącym wyznaczeniu takich strategii badawczych, które mają na celu ograniczenie negatywnych skutków błędów, przy jednoczesnym kontrolowaniu związanych z tym kosztów. Interesujący jest także sposób, w jaki P. Biemer patrzy na całkowity błąd pomiaru. Przyjmuje on dwie różne perspektywy, to znaczy badacza (wytwórcy danych) oraz użytkownika (odbiorcy badań), ukazując, że mogą one prowadzić do odmiennego postrzegania jakości pomiaru. O ile bowiem badacze kładą duży nacisk na precyzję oraz dokładność estymacji (przyglądając się wielkości próby badawczej, wskaźnikom realizacji próby, pokryciu populacji operatem losowania, spójności odpowiedzi respondentów, to znaczy rzetelności oraz homogeniczności pomiaru itd.), przeznaczając na to znaczną część wysiłków oraz środków finansowych, o tyle użytkownicy badań przyjmują najczęściej precyzję oraz dokładność za coś oczywistego, z kolei w stopniu największym interesuje ich aktualność danych, ich dostępność oraz użyteczność.

w której mówi się o *nieobserwowalnych konstruktach* (postawach, opiniach, przekonaniach), które badacze próbują zmierzyć za pomocą *indicatorów* (por. Groves 1989: 18). W polskiej literaturze ten „nieobserwowalny konstrukt" zwykło nazywać się *indicatum*, dla którego poszukuje się *wskaźników* (por. Nowak 2007: 165)[2], czyli takich obserwowalnych zjawisk, za pomocą których wnosi się o występowaniu, bądź też stopniu intensywności występowania – wspomnianego przez Grovesa – nieobserwowalnego konstruktu. W konsekwencji główna uwaga badaczy skupia się nie na tym, czy wynik pomiaru empirycznego „odstaje" od wartości prawdziwej, ale w jakim stopniu zakresy wskaźników oraz indicatum pozostają zbieżne (por. Nowak 2007: 177–181)[3]. Dwa pojęcia funkcjonujące w powszechnym użyciu: *trafność* i *rzetelność*, traktowane jako kryteria oceny jakości pomiaru psychometrycznego, odnoszą się przy tym nie tyle do procesu zbierania danych, co do jego rezultatu (por. Sztabiński F. 2011: 61–76; Groves 1989: 19–27). Pojęcia te będą jeszcze przedmiotem analizy w tej pracy w ramach klasyfikacji głównych źródeł błędów badań sondażowych.

Niezależnie jednak od przyjętego punktu widzenia – statystycznego czy też psychometrycznego – rozsądne wydaje się przyjęcie obu stanowisk, lub, inaczej mówiąc, odpowiednie rozłożenie akcentów w zależności od charakteru poszukiwanej informacji: czy chodzi o fakty, czy też o opinie. Świetnie ujął to F. Sztabiński w przywoływanej już publikacji poświęconej ocenie jakości badań sondażowych:

> Otóż w każdym badaniu zadaje się respondentom pytania o pewne ich cechy, tak zwane metryczkowe […], oraz pytania dotyczące ich opinii, poglądów, przekonań […]. Ponieważ w przypadku cech społeczno-demograficznych istnieje 'wartość prawdziwa', należy w odniesieniu do nich mówić o statystycznej koncepcji błędu. Z kolei w przypadku pytań o […] opinie, przekonania […] – o psy-

---

[2] Stefan Nowak, opisując w *Metodologii badań społecznych* zagadnienia pomiaru, rozróżnia przypadki pomiaru takich zmiennych, dla których określony jest jednoznaczny sens empiryczny, od pomiaru charakterystyk „[…] zasadniczo niedostępn[ych] obserwacjom […] lub też trudno obserwowaln[ych]" (Nowak 2007: 165). W tym drugim przypadku, jak wskazuje S. Nowak, „[…] badacz musi postawić sobie pytanie, w jaki sposób może on mimo wszystko określić pośrednio jej wartość w badaniach i w konsekwencji dokonać odpowiednich pomiarów tej zmiennej" (Nowak 2007: 165). Cały wywód dotyczący pomiaru zapośredniczonego kończy się stwierdzeniem wskazującym na konieczność doboru wskaźników umożliwiających identyfikację owego konstruktu, nazywanego za Tadeuszem Pawłowskim (1969) *indicatum*.

[3] S. Nowak definiuje przy tym trzy różne miary trafności wskaźników (por. Nowak 2007: 177–179). O tym, w jakim stopniu wskaźnik pozwala „wychwycić" te wszystkie elementy, które mają cechę *indicatum*, informuje *miara mocy zawierania*. Z kolei *miara mocy odrzucania* pozwala stwierdzić, w jakim stopniu wskaźnik jest w stanie wyróżnić te wszystkie jednostki, które cechy *indicatum* nie posiadają. Ostatnia ze zdefiniowanych miar – *miara mocy rozdzielczej* – określona została jako współczynnik korelacji pomiędzy wskaźnikiem i *indicatum*.

chologicznym ujęciu błędu. Nie można więc powiedzieć, iż określając błąd pomiaru w danym badaniu, przyjmujemy psychologiczne lub statystyczne jego rozumienie. Należy zatem mówić o 'psychologicznym' lub 'statystycznym' podejściu do błędów pomiaru, w zależności od rodzaju pytań: ich przedmiotu czy też poszukiwanej informacji. (Sztabiński, F. 2011: 48)

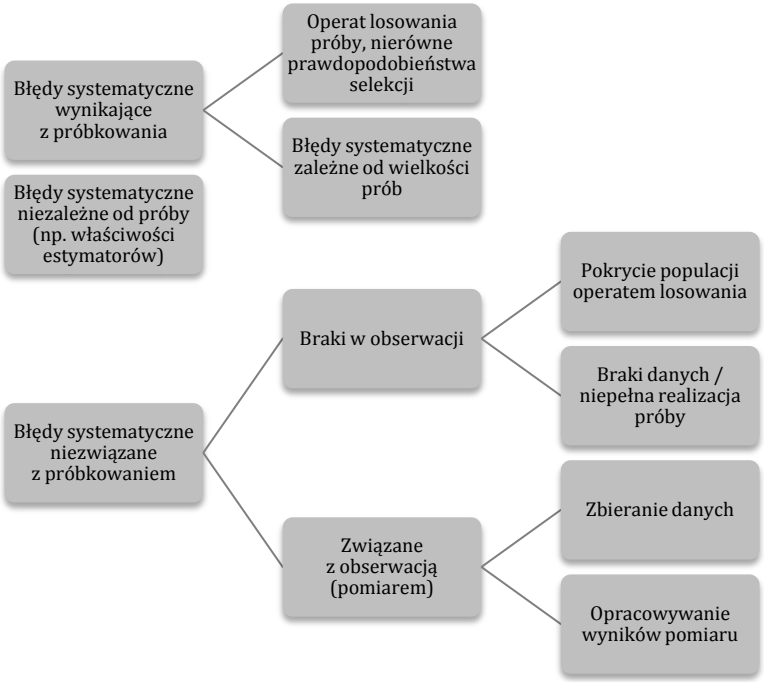
Uzasadnione wydaje się zatem takie podejście, które w ramach analizy potencjalnych źródeł błędów ujmuje psychometryczne komponenty *trafności* oraz *rzetelności* pomiaru. Znajduje to wyraz w wielu współczesnych koncepcjach całkowitego błędu pomiaru badań sondażowych (por. np. Biemer 2010a: 822; Groves i in. 2010: 856; Fuchs 2008: 898; Groves i in. 2004: 48), w tym w przyjętej w tej monografii klasyfikacji źródeł błędów.

I.2. Identyfikacja źródeł błędów w badaniach sondażowych – przegląd literatury

Zainteresowanie metodologów badań reprezentatywnych tematyką błędów sięga początku lat 40. XX wieku. W okresie tym ukazały się pierwsze artykuły dedykowane jakości surveyów, w których brano pod uwagę nie tylko teorię próbkowania reprezentatywnego, ale także specyfikę badań społecznych. Omawiając metodologiczne źródła fascynacji paradygmatem całkowitego błędu pomiaru, R. Groves i L. Lyberg (2010: 851–852) ulokowali genezę owej idei w tekście Edwardsa Deminga (1944: 359–369). Autor ten nie używał wprawdzie *explicite* wyrażenia TSE[4], wskazywał jednak na pewne czynniki, a dokładniej błędy, obniżające jakość przeprowadzonego badania. Wśród nich wymieniał takie komponenty pomiaru jak: próbkowanie, wariancja estymatorów, wpływ ankietera i techniki badawczej na uzyskany wynik pomiaru, braki danych, a także na wiele klas błędów pomiarowych oraz błędów opracowywania wyników. E. Deming nie wspominał o problemach wynikających z niepełnego pokrycia populacji operatem losowania i pomijał w swoich rozważaniach ten (tak powszechny w dzisiejszych koncepcjach TSE) typ błędu, zwracał za to uwagę na pewne specyficzne ograniczenia surveyów związane z etycznymi aspektami badań. Chociaż prace E. Deminga miały duży wpływ na badaczy

[4] W raporcie z projektu badawczego *Errors in Surveys,* opublikowanym przez Tore Daleniusa w 1974 roku, pojawia się pojęcie *Total Survey Design* (por. Groves i in. 2010: 854). Z kolei określenie *Total Survey Error* użyte zostało po raz pierwszy dopiero w 1979 roku. Wykorzystali je R. Andersen, J. Kasper oraz M. Frankel w pracy poświęconej analizie jakości reprezentatywnych badań nad zdrowiem ludności. Autorzy tego opracowania zaproponowali dekompozycję TSE uwzględniającą trzy kryteria. Pierwsze wprowadzało podział na wariancję oraz błędy systematyczne, drugie na błędy próbkowania oraz inne błędy niezwiązane z doborem próby, wreszcie trzecie na błędy niezwiązane oraz związane z obserwacją.

amerykańskich, to jednak przez wiele lat studia poświęcone jakości surveyów koncentrowały się raczej na błędach losowych niż na błędach systematycznych[5]. W zasadzie takie podejście do oceny jakości surveyów przełamane zostało dopiero w pracy Leslie Kisha z 1965 roku, w monografii *Survey Sampling* pojawił się bowiem osobny rozdział dedykowany w całości błędom nielosowym wraz z próbą ich klasyfikacji (por. Kish 1965: 509–573).



Ryc. I.1. Klasyfikacja błędów nielosowych zaproponowana przez L. Kisha

Źródło: opracowanie własne na podstawie Kish 1965: 519

Przedstawiona przez L. Kisha klasyfikacja błędów nielosowych wydaje się interesująca przynajmniej z kilku powodów. Przede wszystkim, po raz pierwszy wprowadzony został w literaturze podział (obecny później w nieco innej posta-

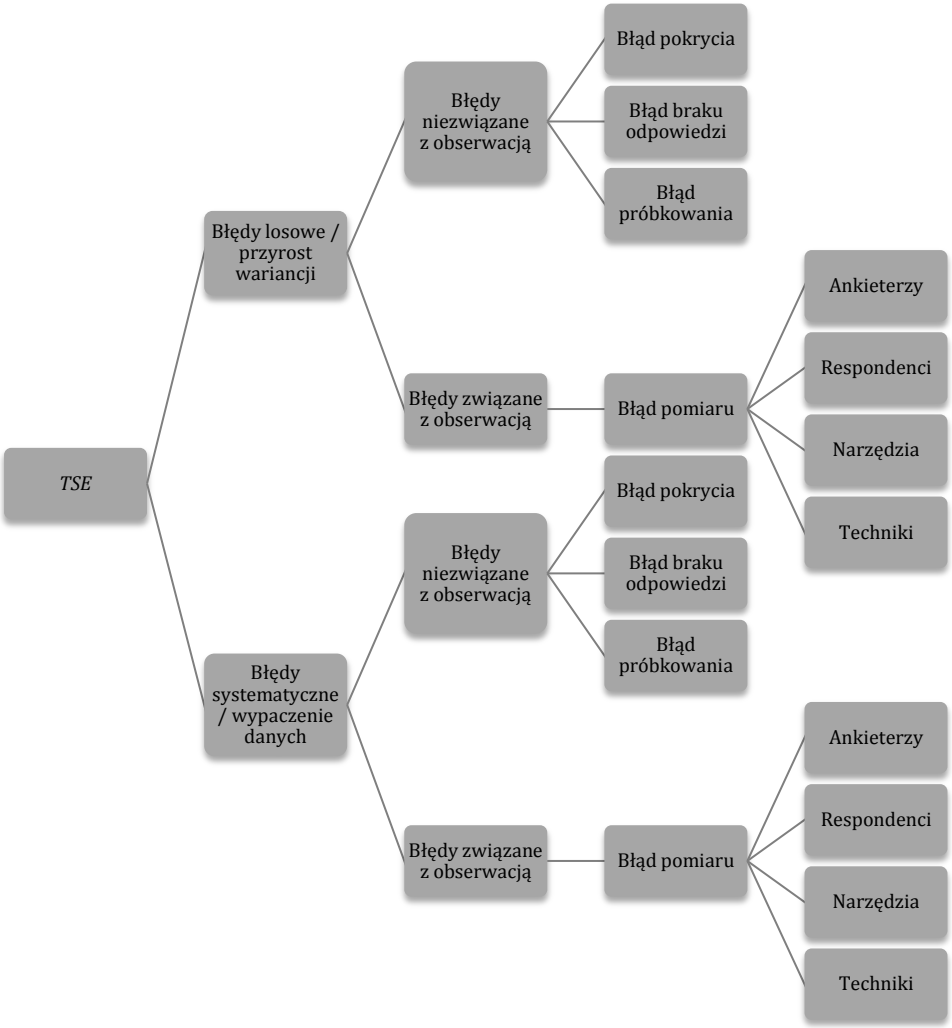
[5] Skupienie uwagi wielu badaczy na błędach losowych H. Weisberg (2005: 13) tłumaczy tym, że ich wielkości daje się w bardzo łatwy sposób wyznaczać za pomocą dobrze znanych formuł matematycznych, podczas gdy wartości błędów nielosowych wydają się najczęściej zbyt trudne do ustalenia. Co więcej, konsekwencje błędów wynikających z próbkowania mogą być redukowane poprzez zwiększanie liczebności próby, nie ma przy tym sprawdzonych reguł ograniczania skutków błędów nielosowych.

ci u R. Grovesa 1989), na błędy związane z obserwacją (*observation error*) oraz błędy niezwiązane z obserwacją (*nonobservation error*). Te pierwsze są domeną pomiaru i wynikają z niedoskonałości procesu zbierania, przetwarzania i opracowywania wyników badań. Drugie z kolei pozostają efektem zarówno niepełnego pokrycia badanej populacji przez operat wykorzystywany do losowania jednostek, jak też skutkiem niepełnej realizacji próby. Istotne jest w sumie to, że dzięki takiemu „nowemu" sposobowi patrzenia na jakość pomiaru metodologowie badań reprezentatywnych wyszli w swoich analizach poza błędy próbkowania, przypisując coraz większe znaczenie innym źródłom uchybień pojawiających się na wszystkich etapach procesu badawczego.

Paradygmat całkowitego błędu pomiaru zawdzięcza swoją niezwykłą popularność przede wszystkim jednak pracom R. Grovesa, w tym analizom zaprezentowanym przez tego znanego metodologa pod koniec lat 80. XX wieku w pracy *Survey Errors and Survey Costs*. Od jej pierwszego wydania w 1989 roku paradygmat TSE stał się dominującą perspektywą wykorzystywaną do opisu jakości surveyów. Nie sposób zatem mówić o całkowitym błędzie pomiaru, nie odwołując się do ustaleń R. Grovesa (1989). Należy zacząć od tego, że we wprowadzeniu do przywołanej książki autor składa deklarację ukierunkowującą analizy na cztery kategorie błędów: (a) pokrycia, (b) braku odpowiedzi, (c) próbkowania (*sampling error*) oraz (d) zbierania danych:

> Zacznę od przyjęcia założenia, że wszystkie próby surveyowe obarczone są różnego rodzaju błędami, takimi jak:
>
> 1. Błąd pokrycia, wyrastający z braku szans wylosowania pewnych jednostek populacji.
> 2. Błąd braku odpowiedzi, wynikający z nieustalenia danych dla pewnych osób wylosowanych do próby.
> 3. Błąd próbkowania, wynikający ze zróżnicowania [jednostek – P.J.] […] w populacji.
> 4. Błąd pomiaru, wynikający z niedokładności procedur ustalania wartości zmiennych. Wyrasta on na bazie:
>   a. efektu oddziaływania ankieterów na odpowiedzi respondentów;
>   b. błędów powiązanych z respondentami […];
>   c. błędów […] narzędzi badawczych […];
>   d. błędów będących efektem wyboru określonej techniki zbierania danych […]. (Groves 1989: vi)

W rozdziale pierwszym tej książki przedstawiona została również dekompozycja całkowitego błędu pomiaru na komponent losowy oraz systematyczny. Groves wprowadza też rozróżnienie na błędy związane z obserwacją oraz te, które z obserwacją nie są powiązane. Podział ten obecny był wprawdzie już u Kisha (1965), jednak pojawił się tam w nieco innej postaci. Groves traktuje



Ryc. I.2. Klasyfikacja błędów zaproponowana przez R. Grovesa

Źródło: opracowanie własne na podstawie Groves 1989: 8–30

bowiem błędy próbkowania jako jeden z komponentów błędów niezwiązanych z obserwacją, inaczej niż Kish, który uznawał je za odrębną klasę zanieczyszczeń wyników pomiaru. A zatem dla R. Grovesa błędy niezwiązane z obserwacją to takie, które „pojawiają się na skutek przeprowadzenia pomiaru nie całej, a jedynie części populacji" (Groves 1989: 11), z kolei te związane z obserwacją (pomiarem) „odnoszą się do różnic pomiędzy odpowiedziami udzielanymi przez respondentów na zadawane im pytania a prawdziwymi wartościami

[tych zmiennych – P.J.]" (Groves 1989: 11). Innymi słowy, błąd pokrycia, błąd próbkowania oraz błąd braku odpowiedzi przyporządkowane są przez tego autora do klasy błędów niezwiązanych z obserwacją, z kolei źródeł błędów związanych z obserwacją należy, jego zdaniem, upatrywać u ankieterów, respondentów, w narzędziach badawczych oraz technikach gromadzenia danych.

Zauważyć można, że w przedstawionym przez Grovesa wyliczeniu źródeł błędów chodzi tak naprawdę o wyróżnienie błędów losowych i nielosowych (źródła tych dwóch klas błędów są takie same). Co więcej, w klasyfikacji tej nie pojawiają się błędy opracowywania wyników pomiaru. Pominięcie ich jest jednak zabiegiem celowym, gdyż przedmiotem zainteresowania Grovesa pozostaje *de facto* wyłącznie proces próbkowania oraz zbierania danych, a nie te etapy badania, które nazwane zostały przez Herberta Weisberga postsurveyowymi (por. Weisberg 2005: 19). Doskonałym tego potwierdzeniem jest następujący fragment z pierwszego rozdziału monografii *Survey Errors and Survey Costs*:

> Czytelnik może zauważyć, że rysunek [na którym Groves zaprezentował dekompozycję błędów – P.J.] nie jest wyczerpującym wyliczeniem wszystkich źródeł błędów badań surveyowych. Do najbardziej znaczących [źródeł błędów – P.J.] pominiętych w tym wyliczeniu należą te, które wyrastają z […] kodowania, edycji, wprowadzania danych i innych operacji na zbiorach wykonywanych po fazie gromadzenia wyników. Te elementy zostały celowo pominięte […] nie dlatego, że są trywialnymi źródłami błędów, ale że nie wiążą się z [fazą doboru respondentów i gromadzenia danych – P.J.]. Wynikają natomiast z działania analityków oraz osób zarządzających bazami danych. (Groves 1989: 12)

Pominięcie błędów procedowania danych jest zresztą wyrazem poglądu Grovesa, wyłożonym w artykule opublikowanym w 2010 roku wspólnie z L. Lybergiem, w którym obaj metodolodzy krytykują dążenie wielu badaczy do wyliczenia wszystkich możliwych źródeł błędów (por. Groves i in. 2010: 854).

Praca Grovesa (1989) jest interesująca również z innych względów. Autor porównał w niej terminologię wykorzystywaną przez przedstawicieli kilku różnych (sub)dyscyplin naukowych do definiowania błędów pojawiających się w badaniach reprezentatywnych. Zestawił przy tym podejście charakterystyczne dla badań społecznych z tym, które pojawia się w literaturze statystycznej, psychometrycznej oraz ekonometrycznej. Analizy Grovesa pokazały, że badacze z kręgu psychometrii koncentrują swoją uwagę na *trafności* i *rzetelności* pomiaru, zupełnie inaczej niż ma to miejsce w statystycznym podejściu do błędów, gdzie zainteresowanie skupia się na *wariancji* estymatorów oraz systematycznym *wypaczeniu* pomiaru. Groves pokazuje, że psychometryczna *rzetelność* nie jest tym samym, czym statystyczna *wariancja*, podobnie zresztą jak *trafność* i *wypaczenie*, które też nie mogą być traktowane jako pojęcia tożsame (por. Groves 1989: 18).

Komplikacje te doskonale opisali również współautorzy monografii *Survey Methodology*, którzy stwierdzają, iż:

> trafna miara jakiegoś konstruktu to taka, która jest z nim idealnie skorelowana. […] Dwie miary mogą być idealnie skorelowane, ale prowadzić do różnych wartości. […] To jest podstawowa rozbieżność teorii pomiaru psychometrycznego oraz statystycznej terminologii błędów. (Groves i in. 2004: 51)

W dalszej części publikacji odnaleźć można jeszcze bardziej precyzyjne uszczegółowienie tej kwestii:

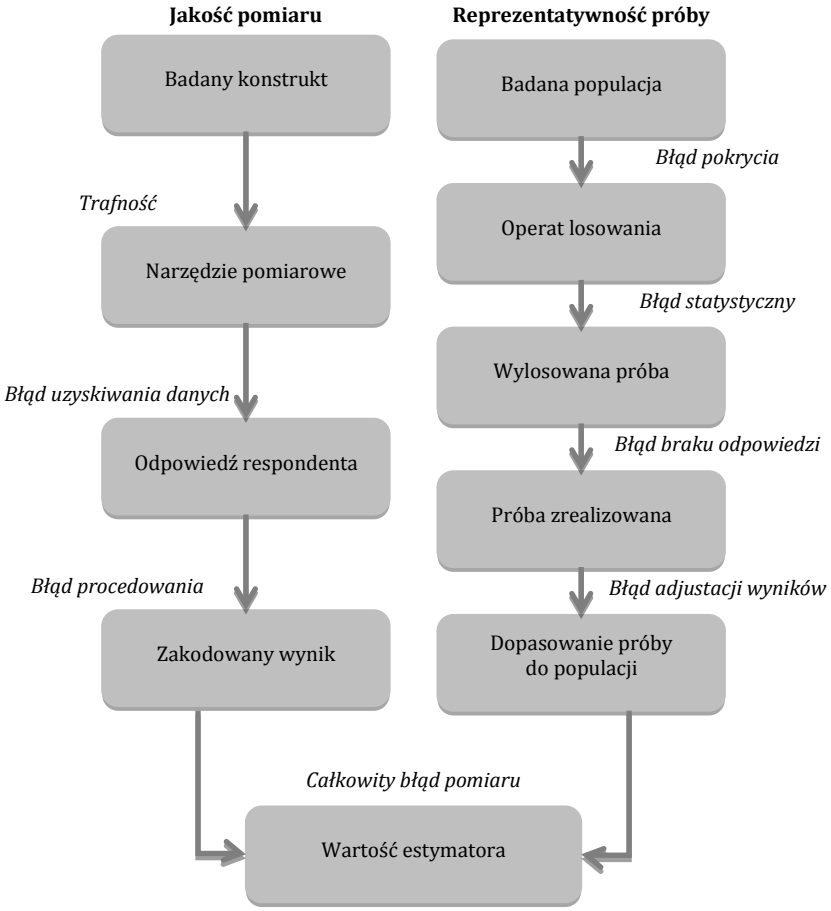
> podstawowym nieporozumieniem terminologicznym w odniesieniu do błędów powiązanych z pytaniem, jest to, w jaki sposób 'trafność' łączy się z 'wypaczeniem'. […] Przeanalizujmy, co stałoby się, gdyby odpowiedzi respondentów różniły się systematycznie od wartości prawdziwej? […] W pewnych warunkach systematyczne niedoszacowanie [prawdziwych wartości – P.J.] nie obniżyłoby współczynnika korelacji pomiędzy odpowiedzią a tą prawdziwą wartością. Dla przykładu, gdyby wszyscy respondenci niedoszacowali swojej wagi [o tę samą wielkość – P.J.] to współczynnik korelacji [z wartością prawdziwą – P.J.] wyniósłby jeden, pomimo tego, że przeciętna wartość w całej próbie byłaby [dokładnie o tę wielkość niedoszacowana – P.J.]. (Groves i in. 2004: 258–259)

Powracając do przeprowadzonego przez Grovesa (1989) porównania różnych sposobów definiowania błędów pomiarowych, można wskazać, że najbardziej zbliżoną terminologię wykorzystywaną do opisu błędów odnaleźć można w literaturze badań surveyowych oraz w statystycznej teorii próbkowania. W obu przypadkach pojawiają się takie określenia jak: *dokładność* (*accuracy*) na oznaczenie błędu całkowitego, *precyzja* na oznaczenie wariancji estymatorów, *błąd statystyczny* na oznaczenie wariancji wynikającej z próbkowania losowego oraz *wypaczenie/błąd systematyczny* na oznaczenie nielosowych odchyleń wyników pomiaru od (prawdziwych) wartości parametrów w populacji. Groves pokazuje jednocześnie, że w statystycznym spojrzeniu na błędy uwaga skupia się bardziej na procesie próbkowania[6], natomiast metodologowie badań surveyowych koncentrują się na błędach operatów, błędach niepełnej realizacji próby oraz błędach gromadzenia danych.

---

[6] W perspektywie statystycznej zwraca się uwagę na wielkości prób badawczych, dążąc w ten sposób do ograniczania błędów statystycznych. Chociaż zdaniem Grovesa takie statystyczne spojrzenie na błędy pomiarowe jest niezwykle wartościowe, to jednak pozostaje niekompletne oraz niesatysfakcjonujące, pomija bowiem praktyczne problemy wynikające z organizacji procesu gromadzenia danych, np. wpływ i efekt ankieterski, różnice w stopniu zaangażowania się respondentów w udzielanie odpowiedzi i kwestie związane z jakością narzędzi badawczych (por. Groves 1989: 13).

W wydanej w 2004 roku książce *Survey Metodology* Groves wraz ze współautorami zaproponował jeszcze inny sposób klasyfikowania źródeł błędów w ramach paradygmatu całkowitego błędu pomiaru. Odnaleźć można tam przyporządkowanie błędów do kolejnych etapów procesu badawczego. Ukazuje ono niezwykle ciekawą konsekwencję dość oczywistej prawidłowości, a mianowicie, że w trakcie badań surveyowych dąży się do jednoczesnego zapewnienia wysokiej jakości pomiaru oraz uzyskania możliwie najlepszego poziomu reprezentatywności próby. W ramach pierwszego przedsięwzięcia chodzi o te wszystkie zadania, które ulokowane są na kontinuum: *odpowiedź respondenta – mierzony konstrukt/indicatum*. Z kolei w drugim działaniu mieszczą się te



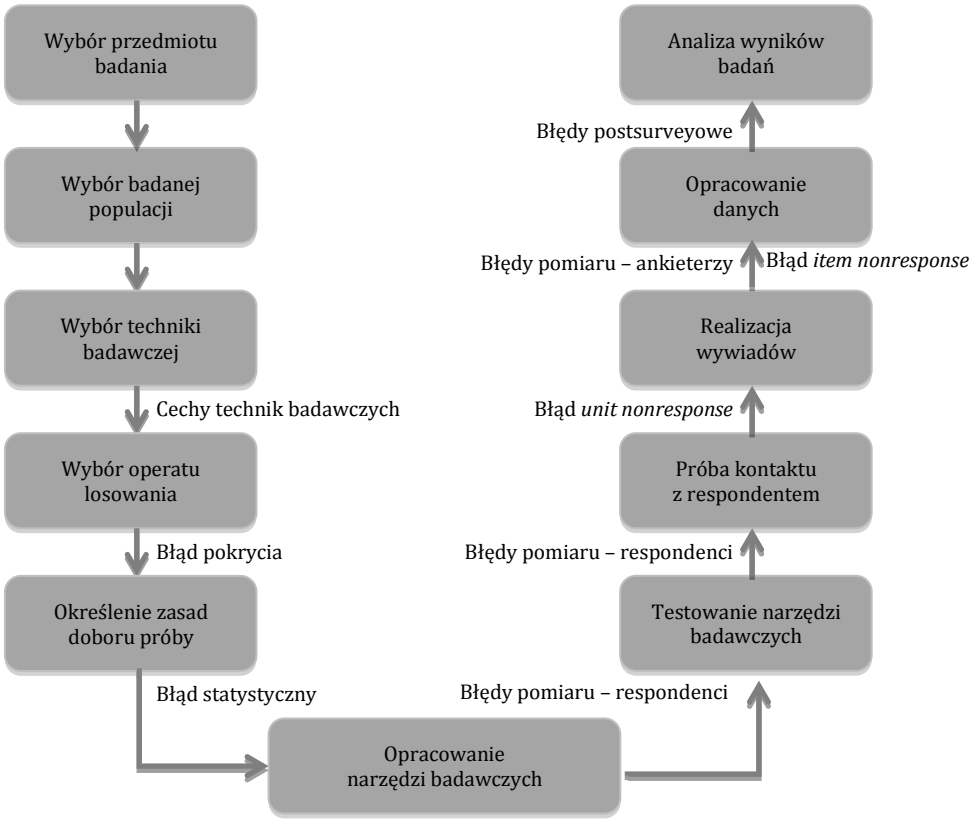
Ryc. I.3. Błędy badań sondażowych a jakość pomiaru oraz reprezentatywność próby

Źródło: opracowanie własne na podstawie Groves i in. 2004: 48

wszystkie zadania, które rozciągają się na linii: *osoby udzielające odpowiedzi – badana populacja*. Poszczególne źródła błędów przypisane zostały właśnie do tych dwóch zasadniczych faz procesu badawczego. O ile zatem – w opinii Grovesa i in. (2004) – jakość pomiaru obciążona jest przez trafność wskaźników, błędy uzyskiwania danych oraz błędy ich opracowywania, o tyle na reprezentatywność próby badawczej wpływ ma błąd pokrycia, błąd próbkowania, niedostępność jednostek oraz ważenie danych (por. Groves i in. 2004: 39–48). Co istotne, w książce *Survey Methodology* omówiono te same źródła błędów, które charakteryzował wcześniej Groves (1989), może z wyjątkiem błędów opracowywania wyników – celowo wykluczonych z analiz prezentowanych we wcześniejszej chronologicznie publikacji *Survey Errors and Survey Costs*. Autorzy opierają się również na przyjętym przez Grovesa (1989) podziale całkowitego błędu pomiaru na komponent błędu losowego/wariancji oraz błędu nielosowego/wypaczenia wyników, a także na błędy związane i niezwiązane z procesem pomiaru (por. Groves i in. 2004: 60–61).

Przypisanie błędów losowych i systematycznych do kolejnych etapów badania sondażowego odnaleźć można również w książce Weisberga (2005). Autor ten, podobnie jak Groves i in. (2004), wyraża przekonanie, iż „błąd może wystąpić na każdym etapie realizacji surveyu" (Weisberg 2005: 16). Prezentowane przez niego ujęcie procesu badawczego jest jednak zasadniczo odmienne od propozycji przedstawionej przez Grovesa i in. (2004). Na proces badawczy H. Weisberg nie patrzy bowiem przez pryzmat działań podejmowanych w celu zapewnienia odpowiedniej jakości pomiaru oraz zadowalającego poziomu reprezentatywności próby, ale jako na łańcuch kolejnych – następujących po sobie w sposób chronologiczny – czynności badawczych[7]. Obejmują one wszystkie działania podejmowane w trakcie realizacji sondaży, to znaczy wstępne określenie przedmiotu i celu badania, zdefiniowanie badanej populacji, wybór operatu losowania, określenie zasad doboru próby, wybór techniki badawczej, opracowanie narzędzi badawczych, pilotaż badania, terenową realizację próby, opracowanie i przygotowanie baz danych oraz analizę wyników. Weisberg przypisuje kolejnym, choć – wbrew wcześniejszym deklaracjom – nie wszyst-

[7] Zaprezentowany przez Weisberga sposób wyróżniania etapów procesu badawczego jest w zasadzie zbieżny z doskonale znanym w polskiej metodologii ujęciem Grzegorza Babińskiego (1980: 19–34), który w artykule *Etapy procesu badawczego* zidentyfikował dwanaście „uniwersalnych" kroków badawczych podejmowanych w trakcie realizacji badań terenowych, w tym: (1) wstępne sformułowanie problemu badawczego, (2) eksplikację problematyki badawczej, (3) operacjonalizację problematyki badawczej, (4) przygotowanie narzędzi badawczych, (5) pilotaż badania, (6) dobór próby, (7) realizację badań empirycznych, (8) weryfikację zebranego materiału empirycznego, (9) wstępne grupowanie materiału surowego, (10) analizę zebranego materiału empirycznego, (11) testowanie hipotez i uogólnianie wniosków, (12) pisanie końcowego raportu z badań.



Ryc. I.4. Błędy pomiarowe a etapy procesu badawczego

Źródło: opracowanie własne na podstawie Weisberg 2005: 17–18

kim etapom procesu badawczego, odpowiednie kategorie błędów. Wyróżnia przy tym błędy związane z (1) doborem respondentów, (2) uzyskiwaniem odpowiedzi oraz (3) zarządzaniem badaniami (por. Weisberg 2005: 17–20). Zdaniem tego autora, dobór respondentów obarczony jest głównie: (a) statystycznym błędem próbkowania, (b) błędem pokrycia populacji operatem losowania oraz (c) błędem wynikającym z niedostępności pewnych jednostek (*unit nonresponse*). Klasa błędów związanych z dokładnością wyników obejmuje natomiast: (d) błąd braku odpowiedzi na niektóre pytania (*item nonresponse*), (e) błąd pomiaru, którego źródłem jest respondent oraz (f) błąd pomiaru, którego źródłem jest ankieter. Z kolei trzy ostatnie źródła błędów: (g) błąd procedowania i analizowania wyników, (h) błąd wynikający z wyboru techniki gromadzenia danych, a także (i) błąd ekwiwalentności przedmiotowej oraz

czasowej przyporządkowane zostały do grupy problemów wynikających z uchybień w administrowaniu i zarządzaniu badaniami[8].

W odróżnieniu od Grovesa i in. (2004), H. Weisberg (2005) nie bierze pod uwagę psychometrycznych miar trafności i rzetelności pomiaru, poświęca za to o wiele więcej uwagi procedowaniu wyników. Nie są one wprawdzie związane z procesem zbierania danych, jednak wpływają bezpośrednio na dokładność pomiaru (por. Weisberg 2005: 261). Komplikacje wynikające z uchybień popełnianych w trakcie opracowywania wyników badań będą jeszcze przedmiotem rozważań w II rozdziale tej pracy.

I.3. Typologia błędów – reprezentatywność próby *vs.* dokładność wyników pomiaru

Przegląd literatury metodologicznej wskazał na znaczne zróżnicowanie w sposobach identyfikacji źródeł błędów oraz na liczne kryteria stosowane do ich klasyfikacji. Warto uporządkować te ustalenia, ukierunkowując prowadzone rozważania na te spośród wszystkich źródeł błędów, które mają przełożenie na *reprezentatywność próby* sondażowej. Zanim to jednak nastąpi, należy najpierw wskazać, że poza błędami próbkowania (*sampling error*), wynikającymi z samej natury badań prowadzonych na próbach losowych, badacze identyfikują także cztery główne klasy błędów nielosowych/systematycznych, tj. błąd operatu losowania (*coverage/frame error*), błąd braku odpowiedzi (*nonresponse error*), błąd pomiaru (*measurement error*), a także błąd opracowywania wyników badań (*processing error*). Wyliczenie to nie jest oczywiście wyczerpujące[9], choć obejmuje pewien typowy wykaz źródeł błędów, charakterystyczny dla większości analiz osadzonych w paradygmacie TSE. Znaczące jest to, że każda z wyróżnionych klas błędów zawiera własne (specyficzne) źródła zanieczyszczeń pomiaru, może też w zróżnicowanym stopniu wpływać na wariancję oraz systematyczne wypaczenie estymatorów (por. Biemer 2010b: 30)[10]. Zupełnie

---

[8] Perspektywa porównawcza nie będzie w tej monografii przedmiotem jakiegoś specjalnego zainteresowania. Specyficzne dla surveyów międzykrajowych źródła błędów, takie jak: ekwiwalentność przedmiotowa, pomiarowa i temporalno-procesualna, były jednak przedmiotem analiz wielu badaczy (por. np. Stoop i in. 2010: 5–8; Cichocki i in. 2009: 195; Słomczyński 2004: 85–116; Sztabiński P.B. 2004: 28–30).

[9] Dla przykładu P. Biemer oraz L. Lyberg (2003: 38–43) włączają w obręb swoich rozważań również błąd pojawiający się na etapie konceptualizacji oraz operacjonalizacji problematyki badawczej, nazywany przez nich błędem specyfikacji (*specification error*), z kolei R. Groves (1989: 41–46) rozpatruje dodatkowo zagadnienia trafności oraz rzetelności pomiaru.

[10] Ciekawą analizę w tym zakresie odnaleźć można w opracowaniu P. Biemera (2010b: 30). Jako potencjalne źródło błędu w ramach konceptualizacji oraz operacjonalizacji autor ten wskazuje na niewłaściwe opracowanie pytań kwestionariuszowych, na przykład nietrafne wskaźniki,

oczywiste wydaje się stwierdzenie, że ryzyko oddziaływania każdej klasy błędów – bądź to na losowy, bądź też na systematyczny komponent TSE – zależy od wielu aspektów. Takim czynnikiem różnicującym może być na przykład sposób doboru próby. W badaniach realizowanych na próbach losowych ryzyko wypaczenia wyników na skutek próbkowania jest niewielkie, inaczej niż w próbach dobieranych celowo czy też systematycznie (por. Park i in. 2004: 183–193; Sirken 2002: 183–190). Innym przykładem mogą być procedury ważenia danych, które będą w większym stopniu oddziaływać na wariancję niż na błąd nielosowy, chyba że zmienne wykorzystane do ustalenia wartości wag będą same w sobie obarczone błędem, co skutkować będzie nie tylko zwiększeniem wariancji estymatorów, ale również ich systematycznym zniekształceniem (por. Billiet i in. 2009: 12–13). Doskonale znane są rozterki badaczy podejmujących problematykę braku odpowiedzi. Ten typ błędu wpływa oczywiście na wariancję estymatorów poprzez zmniejszenie liczebności próby, ale także, a może przede wszystkim, na ich wypaczenie, będące konsekwencją nielosowego odpadu z próby pewnych specyficznych kategorii badanych osób (por. Jabkowski 2011: 41–44; Sztabiński i in. 2007: 31–37; Groves 2006: 657–662; Domański 1999: 72–78). Podobnie rzecz ma się z różnymi typami błędów uzyskiwania i opracowywania wyników, które mogą w niejednorodnym zakresie oddziaływać bądź to na precyzję pomiaru, bądź też systematycznie zniekształcać wyniki pomiaru, bowiem jednym z głównych źródeł błędów pozyskiwania danych są ankieterzy oraz ich zachowania w trakcie realizacji wywiadów (por. Olson i in. 2011: 99–114; West i in. 2010: 1004–1026; Biemer i in. 2003: 156–169; Groves 1989: 395–406). Problematyka ta jest doskonale znana w polskiej literaturze pod pojęciem *wpływu* oraz *efektu ankieterskiego* i była przedmiotem analizy autorów wywodzących się m.in. z łódzkiej szkoły metodologicznej (por. Lutyńska 1998: 17–36; Lutyńska 1997: 52–57; Sztabiński P.B. 1997: 115–124; Sztabiński P.B. 1995: 159–168). Upraszczając w tym momencie ustalenia metodologów, można zauważyć, że jeśli ankieterzy intencjonalnie zmieniają sens wywia-

---

zawieranie pytań, które nie mają związku z odpowiedziami na pytania badawcze, tj. są nieistotne z punktu widzenia problematyki badawczej. W odniesieniu do błędu pokrycia Biemer wskazuje, iż jego źródłem są braki jednostek populacji w operatach losowania, błędne przypisywanie jednostek do populacji, duplikowanie się osób w bazach danych czy też niepełne lub błędne informacje niepozwalające na dotarcie do wylosowanych osób. Do źródeł błędów *nonresponse* zaliczone zostały braki danych powstałe na skutek odmowy udziału w badaniu, niedostępności wylosowanych osób oraz te wszystkie przypadki, w których przeprowadzono wywiad w niepełnym zakresie (np. respondent odmówił dalszej współpracy lub pominął niektóre pytania). Z błędem pozyskiwania danych wiązane są z kolei problemy wynikające z zastosowanych technik badawczych, zachowań ankieterów i respondentów oraz niedoskonałości narzędzi badawczych. Procedowanie danych może być natomiast obarczone błędem powstałym na skutek niewłaściwego wprowadzania wyników do baz danych oraz edycji baz danych, złego kodowania pytań (np. otwartych, wielokrotnego wyboru czy też warunkowych), problemów wynikających z konieczności ważenia wyników oraz niewłaściwego ich przedstawiania.

du (np. źle zadają pytania, sugerują odpowiedzi zgodne z własnymi przekonaniami itd.) lub nieintencjonalnie wpływają na uzyskiwane odpowiedzi (np. z uwagi na swoje cechy osobiste), to wiąże się to z wysokim ryzykiem błędu systematycznego. Jeśli dodatkowo odpowiedzi uzyskiwane przez poszczególnych ankieterów różnią się z uwagi na wariancję (przy założeniu, że respondenci przydzielani byli im losowo), to ankieterzy stają się też źródłem błędu losowego.

Należy zatem postawić ważne pytanie, a mianowicie: czy nie jest przypadkiem tak, że każda klasa błędów oddziałuje zarówno na losowy, jak i nielosowy komponent całkowitego błędu pomiaru? Odpowiedź jest twierdząca, chociaż oddziaływanie to jest nierównomierne dla poszczególnych klas błędów.

Tabela I.1. Ryzyko błędu losowego oraz systematycznego dla głównych komponentów TSE

Komponent TSE	Ryzyko błędu losowego	Ryzyko błędu systematycznego
Statystyczny błąd próbkowania	wysokie	niskie
Błąd specyfikacji	niskie	wysokie
Błąd pokrycia	niskie	wysokie
Błąd braku odpowiedzi	niskie	wysokie
Błąd uzyskiwania danych	wysokie	wysokie
Błąd procedowania danych	wysokie	wysokie

Źródło: opracowanie własne na podstawie P. Biemer (2010b: 45)

Niezwykle pomocne w rozstrzygnięciu tego problemu okazują się analizy metodologiczne przeprowadzone przez Biemera (2010b: 44–45). W drugim rozdziale podręcznika *The Handbook of Survey Research* autor ten przedstawia tabelę przyporządkowującą do kolejnych komponentów całkowitego błędu pomiaru określone ryzyko wystąpienia błędu losowego oraz systematycznego. Ukazuje ona, że chociaż błąd konceptualizacji i operacjonalizacji, błąd pokrycia i błąd braku odpowiedzi niosą ze sobą wysokie ryzyko błędu systematycznego, to jednocześnie ich wpływ na błąd losowy jest niewielki[11]. Niejako w opozycji do tych błędów znajduje się błąd próbkowania, który raczej nie prowadzi do

[11] Takie myślenie o przełożeniu błędów badań sondażowych na losowy lub systematyczny komponent całkowitego błędu pomiaru jest naturalnie dość dużym uproszczeniem. Wystarczy zauważyć, iż niepełna realizacja próby badawczej wiąże się z ryzykiem zniekształcenia wyników pomiaru oraz skutkuje utratą precyzji estymacji. Z drugiej jednak strony to właśnie wypaczenie danych jest głównym komponentem błędu niepełnej realizacji próby badawczej, której wielkość, inaczej niż poziom wariancji, nie ulega redukcji wraz ze wzrostem liczebności sondażowej próby badawczej (por. Brick 2013: 330).

systematycznego zniekształcenia wyników, ale w niezwykle znaczący sposób oddziałuje na precyzję pomiaru. Z kolei błąd uzyskiwania danych oraz błąd ich procedowania oddziałują zarówno na losowy, jak i nielosowy komponent całkowitego błędu pomiaru.

Wszystkie zidentyfikowane w literaturze klasy błędów badań sondażowych zestawiono w tabeli I.2. Przyporządkowano je do odpowiednich kategorii powstałych po skrzyżowaniu dwóch kryteriów podziału. Pierwsze z nich opiera się na odróżnieniu błędów losowych, czyli tych, które powodują wzrost (lub spadek) wariancji estymatorów, od błędów systematycznych, to znaczy takich, które są efektem kierunkowego wpływu jakiegoś czynnika na badaną cechę.

Tabela I.2. Klasyfikacja błędów pomiaru – terminologia przyjęta w monografii

Kryteria podziału	**Błędy losowe** ↓ ***wariancja***	**Błędy systematyczne** ↓ ***wypaczenie***
Błędy niezwiązane z obserwacją ↓ ***reprezentatywność próby***	(1) błąd próbkowania/błąd statystyczny (2) błąd schematu próbkowania (*design effect*, prawdopodobieństwa losowania) (3) błąd adjustacji/ważenie danych	(4) błąd pokrycia/błąd operatu losowania • pominięcie jednostek • duplikowanie jednostek • błędna klasyfikacja jednostek (5) błąd braków odpowiedzi • całkowity (*unit*) • częściowy (*item*)
Uchybienia w pomiarze ↓ ***dokładność pomiaru***	(6) etap konceptualizacji i operacjonalizacji • trafność oraz rzetelność[12] • błąd specyfikacji (7) błąd pomiaru • technika badawcza • narzędzie badawcze • ankieterzy • respondenci	
Przetwarzanie danych ↓ ***dokładność wyników***	(8) błąd procedowania danych • wprowadzanie danych • edytowanie danych • kodowanie pytań • analiza danych • przedstawianie wyników	

Źródło: opracowanie własne

[12] Ponieważ w podejściu psychometrycznym „trafność" oraz „rzetelność" traktowane są jako kryteria oceny jakości wskaźników lub skal pomiarowych, w odniesieniu do tych dwóch kategorii w literaturze metodologicznej nie używa się terminu *błąd*.

Jest to podział stosowany przez badaczy prawie powszechnie (por. Biemer i in. 2010a: 48; Stoop i in. 2010: 3–4; Alwin 2007: 5–6; Weisberg 2005: 13; Groves 2004: 8), głównie z uwagi na fakt, że miara całkowitego błędu pomiaru, czyli błąd średniokwadratowy, daje się rozpisać jako suma kwadratu wszystkich błędów systematycznych oraz wariancji (por. np. Biemer 2010a: 825–829). Drugie kryterium zaczerpnięte zostało od R. Grovesa (1989: 11), choć obecne jest ono również w rozważaniach innych autorów (por. Biemer 2011: 14; Sztabiński F. 2011: 48–60; Stoop i in. 2010: 4). Podział opiera się na odróżnieniu błędów związanych z obserwacją (*error of observation*), czyli tych, które wynikają z niedoskonałości i uchybień w przeprowadzanym pomiarze jakiejś cechy, od błędów niezwiązanych z obserwacją (*error of nonobservation*), czyli tych, które poprzez brak możliwości objęcia badaniem całej populacji (lub próby) obniżają reprezentatywność próby sondażowej. Dodatkową klasę stanowią błędy wynikające z niewłaściwego opracowywania danych – oddziałujące na dokładność wyników (por. Biemer i in. 2003: 219–222). Wpisując kolejne źródła błędów do tabeli I.2., wykorzystano również przywoływane wcześniej ustalenia P. Biemera (2010b: 45) ukazujące ryzyko oddziaływania każdego z tych błędów na przyrost wariancji oraz systematyczne zniekształcenie wyników pomiaru.

Błędy będące konsekwencją niedoskonałości zbierania danych (tj. klasa (7) błędów pomiaru), jak również błędy wynikające z uchybień w przetwarzaniu danych (8), przyporządkowane zostały zarówno do rodziny błędów losowych, jak i też systematycznych. Z kolei wprowadzony przez Biemera oraz Lyberga (2003) błąd specyfikacji lokuje się w klasie błędów systematycznych. W świetle zagadnień poruszanych w tej pracy najbardziej interesujące są jednak te źródła błędów, które obniżają reprezentatywność prób badawczych. Ich natura tkwi w procesie próbkowania i jest efektem pominięcia w badaniu pewnej części jednostek populacji lub próby. Oczywiście idea badań reprezentatywnych zapewnia możliwość wnioskowania o populacji na podstawie jej (niepełnego) losowego podzbioru, jednakże, jak już pokazywano, stopień reprezentatywności przejawia się w dopasowaniu cech posiadanych przez wylosowane jednostki do atrybutów jednostek w całej populacji. Pierwsze dwa źródła błędów, obniżające to dopasowanie, są ściśle związane z probabilistyczną teorią doboru próby. Doskonale znany jest bowiem (1) błąd próbkowania – będący immanentną cechą badań reprezentatywnych. W tym momencie wystarczy przypomnieć, że jego wielkość zależy od liczebności prób badawczych, wariancji wyników oraz założonego poziomu ufności. Drugie źródło problemów związane jest z kolei ze (2) schematem próbkowania i wynika z zastosowania losowania odmiennego od doboru próby prostej. Dla przykładu, w celu uniknięcia znacznego terytorialnego rozproszenia respondentów stosuje się często schematy doboru warstwowego oraz zespołowego/wiązkowanego. Każdy taki odmienny

od losowania prostego schemat doboru próby wpływa jednak na jego efektywność i może powodować przyrost wariancji estymatorów. Dwa kolejne źródła błędów – błąd operatu losowania oraz błąd braku odpowiedzi – zaklasyfikowano z kolei do grupy błędów systematycznych. Są one szczególnym atrybutem badań sondażowych, skupiającym dużą uwagę metodologów. Pierwszy z nich jest przede wszystkim efektem (4) niepełnego pokrycia populacji operatem losowania, tzn. wiąże się z wykluczeniem z populacji pewnej liczby jednostek wchodzących w rzeczywistości w jej skład. Drugi natomiast wynika z faktu (5) niepełnej realizacji próby. Ściśle wiąże się z tym pojęcie *próby zrealizowanej*, różniącej się od *próby wylosowanej* liczbą jednostek, z którymi nie udało się przeprowadzić badania. Warto wreszcie zauważyć, że ostatnim z wymienionych w tabeli I.2. źródłem błędów oddziałujących na poziom reprezentatywności prób badawczych jest ważenie danych. Chociaż w rzeczywistości obejmuje ono działania podejmowane w trakcie opracowywania wyników badań, to jednak konieczność ważenia danych pojawia się na skutek wystąpienia pewnych klas błędów obniżających reprezentatywność próby (por. Groves i in. 2004: 48).

I.4. Błąd średniokwadratowy jako miara liczbowa całkowitego błędu pomiaru

Chociaż istnieje wiele sposobów definiowania miernika całkowitego błędu pomiaru, to jednak najczęściej wykorzystywaną statystyką jest miara nazywana *błędem średniokwadratowym* (w skrócie MSE). Ponieważ całkowity błąd pomiaru jest pojęciem odnoszącym się do wartości pojedynczych estymatorów, a nie do wszystkich estymatorów w danym surveyu, czy też nawet jakiegoś określonego typu statystyk wykorzystanych w badaniu (np. średnich lub wskaźników struktury), to MSE jest miernikiem oceny jakości estymacji każdego parametru z osobna (por. Groves 1989: 8). Ważne jest jednak to, że błąd średniokwadratowy obejmuje wszystkie źródła błędów badań sondażowych, jest więc miarą całkowitego błędu pomiaru. Innymi słowy, niskie wartości MSE oznaczałyby wysoką dokładność pomiaru, z kolei duże wartości MSE wskazywałyby na niską dokładność przeprowadzonego badania (por. Biemer i in. 2003: 45).

W terminologii probabilistycznej błąd średniokwadratowy definiowany jest jako wartość oczekiwana z kwadratu różnicy pomiędzy wielkością estymatora $\hat{\theta}$ ustaloną na podstawie badania oraz prawdziwą wartością szacowanego parametru $\hat{\theta}$. Mówiąc nieco inaczej, MSE jest przeciętną wartością kwadratów różnic pomiędzy wszystkimi tymi estymatorami, które potencjalnie można byłoby wyznaczyć ze wszystkich *n*-elementowych replikacji prób badawczych dobranych z populacji o skończonej liczbie *N* elementów, a rzeczywistą warto-

ścią parametru w całej populacji. W sensie formalnym MSE można zapisać za pomocą równania:

(I.2.) $$MSE(\hat{\theta}) \stackrel{\text{def}}{=} \mathrm{E}(\hat{\theta} - \theta)^2$$ (por. Biemer i in. 2003: 53).

Wystarczy przywołać wzór I.1. (tzn. definicję miary TSE), by zauważyć, że MSE jest wartością oczekiwaną z kwadratu TSE. Stąd wyrażenie I.2. jest równoważne formule:

(I.2'.) $$MSE(\hat{\theta}) \stackrel{\text{def}}{=} \mathrm{E}(TSE(\hat{\theta}))^2.$$

Podkreślić trzeba jednak, że tak długo, jak prawdziwa wartość parametru pozostaje nieznana, niewiadoma jest też wartość miary MSE. Tę niedogodność niezwykle trafnie wyraził Groves, stwierdzając, że „błąd średniokwadratowy jest rzadko kiedy w pełni mierzalny dla wszystkich statystyk surveyowych" (Groves 1989: 8). W wielu przypadkach konieczne jest przeprowadzenie dodatkowych studiów umożliwiających oszacowanie (zatem już nie dokładne wyznaczenie) wartości miary MSE.

Należy zwrócić również uwagę na to, iż w literaturze badań surveyowych spotkać można inną formułę wyrażania błędu średniokwadratowego. Uwzględnia ona dekompozycję tej miary na czynnik losowy oraz systematyczny. Ujmując to bardziej precyzyjnie, MSE przedstawia się często jako sumę (1) wariancji oraz (2) kwadratu skumulowanych błędów nielosowych (por. Biemer 2011: 14; Weisberg 2005: 23; Groves 1989: 8). W sensie formalnym jest to suma: (1) kwadratu odległości (a) przeciętnej wartości estymatorów, wyznaczonych ze wszystkich replikacji *n*-elementowych prób badawczych losowanych z populacji o liczebności *N* jednostek, od (b) prawdziwej wartości parametru, a także (2) przeciętnego kwadratu różnic pomiędzy (a) wartościami poszczególnych estymatorów oraz (b) ich średnią wartością (por. Biemer i in. 2003: 56). Stosując zapis statystyczny, MSE można przedstawić za pomocą doskonale znanej formuły[13]:

(I.3.) $$MSE(\hat{\theta}) \stackrel{\text{def}}{=} \mathrm{B}^2(\hat{\theta}) + \mathrm{Var}(\hat{\theta})$$ (por. Biemer 2010a: 826),

---

[13] Proste przekształcenie wzoru I.2. prowadzi do formuły I.3.: $MSE(\hat{\theta}) = \mathrm{E}(\hat{\theta} - \theta)^2 = \mathrm{E}(\hat{\theta} - \mathrm{E}(\hat{\theta}) + \mathrm{E}(\hat{\theta}) - \theta)^2 = \mathrm{E}\left((\hat{\theta} - \mathrm{E}(\hat{\theta})) + (\mathrm{E}(\hat{\theta}) - \theta)\right)^2 = *$, stosując wzór na wariancję sumy dwóch zmiennych losowych (por. Lissowski i in. 2008: 221) otrzymuje się $* = \mathrm{E}\left(\hat{\theta} - \mathrm{E}(\hat{\theta})\right)^2 + \mathrm{E}(\mathrm{E}(\hat{\theta}) - \theta)^2 + 2\mathrm{Cov}\left((\hat{\theta} - \mathrm{E}(\hat{\theta})); (\mathrm{E}(\hat{\theta}) - \theta)\right) = **$, pierwszy składnik tego wyrażenia jest z definicji – wariancją estymatora, drugi natomiast to kwadrat sumy błędów wypaczenia, wreszcie trzeci to kowariancja obu zmiennych losowych, a ponieważ są one niezależne, to współczynnik kowariancji przyjmuje wartość zero, stąd $** = \mathrm{Var}(\hat{\theta}) + \mathrm{B}^2(\hat{\theta})$. Zatem I.2. jest równoważne I.3.

która sprowadza interpretację wartości błędu średniokwadratowego do kwadratu długości przeciwprostokątnej w trójkącie prostokątnym o przyprostokątnych równych pierwiastkowi wariancji oraz sumie błędów systematycznych.

Należy podkreślić, że chociaż przedstawione sposoby definiowania MSE, to znaczy I.2., I.2'. oraz I.3., dają te same rezultaty, to jednak badacze preferują formułę I.3. pozwalającą na dekompozycję MSE, tj. oddzielenie komponentu losowego od systematycznego. Co ważne jednak, w praktyce wielkość błędu średniokwadratowego da się wyłącznie przybliżyć przez identyfikację niektórych źródeł błędów. Jedną z takich propozycji empirycznego oszacowania MSE odnaleźć można w publikacjach Biemer i in. (2003: 59) oraz Biemer (2011: 14):

$$(\text{I.4.}) \qquad MSE(\hat{\theta}) \cong \left(\sum \mathrm{B}(\hat{\theta})\right)^2 + \mathrm{Var}_{SAMPL}(\hat{\theta}) + \mathrm{Var}_{MSR}(\hat{\theta}) + \mathrm{Var}_{DP}(\hat{\theta}),$$

gdzie:

- $\sum \mathrm{B}(\hat{\theta}) = \mathrm{B}_{NC}(\hat{\theta}) + \mathrm{B}_{NR}(\hat{\theta}) + \mathrm{B}_{SP}(\hat{\theta}) + \mathrm{B}_{MSR}(\hat{\theta}) + \mathrm{B}_{DP}(\hat{\theta})$,

przy czym: $\mathrm{B}_{NC}(\hat{\theta})$ jest przybliżeniem błędu wynikającego z niepełnego pokrycia populacji operatem losowania, $\mathrm{B}_{NR}(\hat{\theta})$ jest estymatorem błędu systematycznego, będącego efektem niepełnej realizacji próby, $\mathrm{B}_{SP}(\hat{\theta})$ jest oszacowaniem błędu systematycznego wynikającego z uchybień w procesie konceptualizacji oraz operacjonalizacji, $\mathrm{B}_{MSR}(\hat{\theta})$ jest przybliżeniem błędu systematycznego powstałego na skutek procesu zbierania danych, $\mathrm{B}_{DP}(\hat{\theta})$ oznacza oszacowanie błędu systematycznego na skutek niewłaściwego postsurveyowego przetwarzania baz danych, $\mathrm{Var}_{SAMPL}(\hat{\theta})$ jest estymatorem wariancji związanej z określonym schematem próbkowania, $\mathrm{Var}_{MSR}(\hat{\theta})$ jest oszacowaniem komponentu wariancji wynikającej z procesu zbierania danych, natomiast $\mathrm{Var}_{DP}(\hat{\theta})$ oznacza estymator wariancji na skutek postsurveyowego przetwarzania wyników. Szacowanie miary MSE wiąże się więc z koniecznością estymacji wielkości pięciu klas błędów nielosowych oraz trzech komponentów wariancji. Wyrażenie I.4. nie jest już jednak tożsame formule I.3., bowiem MSE w postaci I.3. jest konstruktem teoretycznym obejmującym wszystkie bez wyjątku, tzn. znane oraz nieznane źródła błędów badań sondażowych, a I.4. uwzględnia już tylko niektóre z nich.

Chociaż zaproponowane przez Biemera oraz Lyberga oszacowanie MSE jest niezwykle wartościowe, to należy jednak przeprowadzić jego modyfikację w taki sposób, by odpowiadała ona przyjętej wcześniej typologii błędów. Estymator wielkości MSE można bowiem zapisać również jako:

$$(\text{I.5.}) \quad MSE(\hat{\theta}) \cong \left(\sum \mathrm{B}(\hat{\theta})\right)^2 + \mathrm{Var}_{SRS}(\hat{\theta}) \cdot DEFF_{TOTAL} \cdot VIF \cdot DEFF_{ANK} \cdot DEFF_{KOD}$$

gdzie:

- $\sum \mathrm{B}(\hat{\theta}) = \mathrm{B}_{NC}(\hat{\theta}) + \mathrm{B}_{NR}(\hat{\theta}) + \mathrm{B}_{MSR}(\hat{\theta}) + \mathrm{B}_{DP}(\hat{\theta})$.

Symbole $B_{NC}(\hat{\theta})$, $B_{NR}(\hat{\theta})$, $B_{MSR}(\hat{\theta})$ oraz $B_{DP}(\hat{\theta})$ oznaczają jak we wzorze I.4., natomiast $Var_{SRS}(\hat{\theta})$ jest oszacowaniem teoretycznej wariancji estymatorów $\hat{\theta}$ dla prób dobranych zgodnie z zasadą losowania prostego, $DEFF_{TOTAL}$ jest miarą całkowitego efektu schematu losowania próby, innego niż losowanie proste, VIF jest miernikiem przyrostu wariancji na skutek postsurveyowej adjustacji danych, $DEFF_{ANK}$ oznacza miarę przyrostu wariancji wynikającą z efektu ankieterskiego, a $DEFF_{KOD}$ jest oznaczeniem przyrostu wariancji na skutek kodowania wyników badania.

Zasadnicza różnica w definiowaniu MSE za pomocą wzoru I.5. (w porównaniu z I.4.), polega na sposobie wyrażenia losowego komponentu błędu średniokwadratowego. Ujmuje ona wariancję estymatorów relatywnie do teoretycznej wariancji prostej próby losowej, zawiera więc informację o przyroście wariancji (lub analogicznie o utracie precyzji wnioskowania) na skutek przyjęcia określonych schematów losowania prób badawczych, a także zastosowania pewnych procedur uzyskiwania oraz przetwarzania wyników pomiaru. Można również zauważyć, że w I.5. pominięty został – wprowadzony przez Biemera i Lyberga (2003: 38–39) – błąd specyfikacji związany z etapem konceptualizacji oraz operacjonalizacji problematyki badawczej. Przemawia za tym przynajmniej kilka argumentów. Po pierwsze, pojęcie błędu specyfikacji „zarezerwowane" jest do określenia faktu niewłaściwego doboru zmiennych w modelach analitycznych (por. Weisberg 2005: 175; Olsson i in 2004: 453–500). Po drugie, choć Biemer i Lyberg definiują w sposób opisowy błąd specyfikacji, to jednak nie określają go w sposób formalny. Po trzecie, wydaje się, że w ocenie jakości procesu konceptualizacji i operacjonalizacji problematyki badawczej (zwłaszcza w odniesieniu do koncepcji pomiaru wskaźnikowego) bardziej przydatne wydają się, mimo wszystko, psychometryczne kryteria trafności oraz rzetelności pomiaru, których co prawda nie da się implementować w sposób bezpośredni do koncepcji TSE, lecz które są wykorzystywane przez wielu autorów do weryfikacji poprawności doboru wskaźników. Zagadnienia te podjęte będą w rozdziale II, w ramach analizy błędów związanych z konceptualizacją i operacjonalizacją problematyki badawczej. Co oczywiste, formuła I.5. nie wyczerpuje wszystkich źródeł błędów popełnianych w trakcie realizacji badań sondażowych, pozostając jedynie oszacowaniem błędu zdefiniowanego wzorami I.2. oraz I.3.

I.5. Uwagi końcowe

Paradygmat całkowitego błędu pomiaru badań sondażowych nie jest – rzecz jasna – jedyną perspektywą teoretyczną, w świetle której analizować można jakość surveyów. Co więcej, ma on pewne istotne mankamenty opisane między innymi w przywoływanym już artykule *Total Survey Error. Past, Present,*

and Future (por. Groves i in. 2010: 861–867) i w opracowaniu L. Lyberga (2012: 107–130) *Survey Quality*. Jedną z takich najbardziej charakterystycznych ułomności koncepcji TSE jest skupienie uwagi jedynie na wybranych komponentach jakości danych, takich jak precyzja pomiaru czy też wypaczenie estymatorów, a także, choć już w mniejszym stopniu, na rzetelności oraz trafności pomiaru, a pomijanie innych ważnych aspektów jakości danych, takich jak np. ich wiarygodność oraz przydatność (istotność)[14]. Groves i Lyberg (2010) wykazali np., iż jednym z kryteriów wiarygodności danych stosowanym przez odbiorców jest zaufanie do ośrodka badawczego, społeczna opinia o jego profesjonalizmie oraz przekonanie o jego niezależności od wpływów innych instytucji. Wymiary te nie znajdują przełożenia na koncepcję całkowitego błędu pomiaru. Podobnie rzecz ma się z przydatnością danych, która odzwierciedla stopień, w jakim prezentowane wyniki spełniają oczekiwania odbiorców, np. w zakresie ich kompletności, dostępności, aktualności, regularności publikowania itd. Innymi słowy, w koncepcji TSE nieobecna jest w perspektywa odbiorcy surveyów. W konsekwencji estymatory, charakteryzujące się wysoką jakością, mogą być, mimo wszystko, nisko oceniane przez odbiorców badań, lub inaczej, wysoko oceniane mogą być te dane, które są uznawane za istotne i wiarygodne, pomimo iż całkowite błędy pomiaru wykraczać będą poza poziomy możliwe do akceptacji przez badaczy.

Kolejnym mankamentem związanym z paradygmatem całkowitego błędu pomiaru pozostaje to, iż studia metodologiczne nie przekładają się jakoś szczególnie na praktykę surveyową. Nawet w tak znakomitym badaniu, jakim jest projekt Europejskiego Sondażu Społecznego, odnaleźć można szczegółowe informacje o schematach doboru próby, nierównych prawdopodobieństwach selekcji respondentów, efektywnych wielkościach prób badawczych, wielkościach próby zrealizowanej, współczynniku realizowalności próby, a w odniesieniu do niektórych krajów – również informacje o stopniu pokrycia populacji operatem losowania, brakuje jednak wyliczeń wielkości popełnionych błędów

[14] Świetnym studium analizy jakości danych z perspektywy ich odbiorcy są rozważania zaprezentowane w monografii *Ogród metodologii socjologicznej* autorstwa Antoniego Sułka (2002). W rozdziale szóstym pt. *Wiarygodność źródeł i rzetelność danych urzędowych* autor przedstawia pięć praktycznych wskazówek (pytań) ułatwiających użytkownikom danych ocenę ich wiarygodności (por. Sułek 2002: 132–147). Chociaż wskazówki te odnoszą się głównie do zastanych danych statystycznych, to jednak mogą służyć również do oceny wiarygodności danych sondażowych. Ponieważ wiarygodność danych rozpatrywana jest głównie w świetle wiarygodności instytucji, stąd pytania o to: w jaki sposób instytucja zbierała dane?, czy informatorzy instytucji mieli jakieś powody do udzielania informacji stronniczych?, czy dane publikowane przez instytucję są jednocześnie podstawą oceny tej instytucji? A. Sułek proponuje także kryteria wykraczające poza ocenę wiarygodności instytucji, a dokładniej, skupiające się na wiarygodności procesu zbierania danych, np.: czy rejestrowane fakty były jednoznacznie określone, tak aby uniemożliwić ich różną interpretację?, czy zbieranie danych było kontrolowane?, czy osoby udzielające informacji oraz zbierające informacje liczyły się z możliwością kontroli?

systematycznych. Wynika to z oczywistego faktu, że estymacja nielosowych komponentów TSE wymaga dodatkowych, często kosztownych i wyrafinowanych metodologicznie, studiów empirycznych[15]. Ujmując to nieco inaczej, paradygmat całkowitego błędu pomiaru badań sondażowych ukazuje wyłącznie, jakie są główne ograniczenia oraz zagrożenia procesu pomiaru, co nie oznacza, że daje wskazówki umożliwiające łatwe estymowanie wielkości poszczególnych źródeł błędów. Potwierdza to znaną i nieco humorystyczną prawidłowość, iż świadomość występowania problemu nie jest jednoznaczna z uwolnieniem się od niego.

Warto jednak zauważyć, że paradygmat TSE odgrywa niezwykle ważną rolę w praktyce badawczej. Uświadamia bowiem, że na jakość pomiaru reprezentatywnego należy patrzeć całościowo, uwzględniając wzajemne oddziaływanie wszystkich działań badawczych. Istotne jest to, by pamiętać, że chociaż koncentracja wysiłków (i kosztów) na jakimś szczególnym typie błędu (np. pomiaru, niepełnej realizacji próby itd.) może doprowadzić do zredukowania jego wielkości, to jednocześnie działania takie mogą spowodować przyrost wielkości innych błędów. Ta nauka, płynąca z paradygmatu całkowitego błędu pomiaru, stanie się jeszcze bardziej widoczna w kolejnych rozdziałach pracy. Zanim jednak omówione zostaną szczegółowe problemy związane z reprezentatywnością sondażowych prób badawczych, w kolejnym rozdziale zdefiniowane będą wielkości błędów mających najbardziej znaczące przełożenie na jakość badania sondażowego.

---

[15] R. Groves oraz L. Lyberg (2010: 865) ukazują, że nawet w przypadku pomiaru komponentu losowego miary TSE ustalenie wielkości wariancji wymaga i tak znacznego nakładu środków. Stąd badacze stają przed problemem, jaką część budżetu powinni przeznaczyć na pomiar oraz poprawę jakości surveyu, kosztem rezygnacji z podejmowania innych przedsięwzięć.

ROZDZIAŁ II

Źródła błędów badań sondażowych – dylematy metodologiczne i praktyczne

Zakres problemów omówionych w tym rozdziale jest dość szeroki i w znacznej części wykracza poza zagadnienia reprezentatywności prób badawczych, jednak podstawowym celem jest zdefiniowanie tych źródeł błędów, na które zwrócono uwagę w rozdziale pierwszym, w ramach charakterystyki paradygmatu całkowitego błędu pomiaru badań sondażowych. O ile bowiem celem studiów literaturowych była identyfikacja źródeł błędów, to zadaniem podjętym w rozdziale II będzie – poza formalnym określeniem wielkości błędów losowych oraz systematycznych – wskazanie na kluczowe dylematy metodologiczne oraz praktyczne związane z oceną jakości badań sondażowych prowadzoną w paradygmacie błędu całkowitego.

Rozdział składa się z trzech części. Pierwsza z nich skoncentrowana jest na analizie błędów oddziałujących na reprezentatywność prób badawczych. W podrozdziale II.1.1. wprowadzono pojęcia prostej próby losowej oraz statystycznego błędu próbkowania, by następnie, w podrozdziale II.1.2., opisać miernik efektywności zastosowanego schematu doboru próby oraz pojęcie efektywnej wielkości próby badawczej. W dwóch następnych podrozdziałach, II.1.3. oraz II.1.4., scharakteryzowane zostały dwie klasy błędów systematycznych, mających swoje źródło w ułomnościach operatów doboru prób badawczych oraz w niedostępności pewnych jednostek wylosowanych do próby. Pierwszą część rozdziału zamykają studia – zaprezentowane w podrozdziale II.1.5. – poświęcone procedurom ważenia danych oraz relacjom zachodzącym pomiędzy precyzją estymacji a procesem ważenia wyników. Druga część rozdziału zawiera z kolei opis procesu pomiaru. W pierwszej kolejności (podrozdział II.2.1.) scharakteryzowane zostały błędy pojawiające się na etapie konceptualizacji i operacjonalizacji problematyki badawczej, a następnie – miary trafności oraz rzetelności pomiaru. W podrozdziale II.2.2. sformalizowano pro-

ces pomiaru, wyróżniono jego komponenty, a także związane z nimi błędy. Ostatnia część rozdziału, II.3., obejmuje analizę błędów pojawiających się na etapie postterenowego przetwarzania danych wynikowych.

II.1. Błędy związane z reprezentatywnością prób badawczych

II.1.1. Statystyczny błąd próbkowania (*sampling error*)

Z dużym prawdopodobieństwem można powiedzieć, że błąd próbkowania – będący immanentną cechą wszystkich badań reprezentatywnych – jest najlepiej rozpoznanym błędem badań sondażowych. Jego źródłem jest proces losowego doboru części populacji do próby badawczej, a zatem wynika on z faktu przeprowadzenia pomiaru nie na wszystkich, lecz na części jednostek należących do badanej populacji (por. Weisberg 2005: 225). Naturę błędu próbkowania doskonale ujął R. Groves (1989), stwierdzając, iż:

> błąd ten jest błędem wynikającym z braku obserwacji. Estymacja w surveyach ulega błędom próbkowania dlatego, że nie wszyscy członkowie […] populacji pozostali przebadani. Gdyby tak było, to błąd próbkowania zostałby wyeliminowany. Idea [tego – P.J.] błędu [...] opiera się na tym, że powtarzając proces [doboru próby – P.J.] […] wylosowuje się zupełnie inne [zbiory – P.J.] osób z tej samej populacji, co skutkuje różnymi wartościami [oszacowania parametrów – P.J.]. (Groves 1989: 240)

Chociaż w sposób nieuchronny jednostki wylosowane do próby badawczej różnią się od jednostek całej populacji, to jednak na podstawie losowych prób badawczych można dążyć do uogólniania wniosków na populację. Innymi słowy, błąd statystyczny jest miarą tego, jak dalece wartości estymatorów ustalone na podstawie pomiaru próby badawczej różnią się, lub raczej – powinny różnić się od rzeczywistych wartości szacowanych parametrów w całej populacji, bowiem teoria próbkowania statystycznego zakłada pomiar przeprowadzony na wszystkich jednostkach wylosowanych do próby. Gdyby w istocie tak było, to wyniki surveyów obciążone byłyby wyłącznie błędami losowymi. Choć założenie takie przyjmuje się w klasycznych modelach wnioskowania statystycznego, to jednak w rzeczywistości, jak wiadomo, pomiar obciążony jest przez różne kategorie błędów losowych oraz systematycznych. Nie zmienia to faktu, że błąd statystyczny pojawia się zawsze, niezależnie od obecności innych rodzajów błędów.

Wielkość błędu statystycznego zależy oczywiście od liczebności próby badawczej, a także od stopnia zróżnicowania wartości zmiennych, nie mówiąc już

o zależności tego błędu od założonego przez badacza współczynnika ufności. Powszechnie wiadomo, że im większa będzie liczebność próby badawczej, tym większy będzie też poziom precyzji estymatorów. Podobnie zresztą, pomiar okaże się bardziej precyzyjny wtedy, gdy mniejsza będzie wariancja zmiennych poddawanych pomiarowi. W tym kontekście warto wskazać, że w przywoływanej już wcześniej monografii *Survey Errors and Survey Costs* R. Groves (1989) w następujący sposób odpowiada na fundamentalne pytanie o źródła błędów próbkowania:

> heterogeniczność populacji pozostaje sama w sobie powodem występowania wariancji estymatorów […]. Jeżeli populacja byłaby homogeniczna, tzn. jeżeli […] członkowie populacji posiadaliby tę samą cechę, to wszystkie próby (o dowolnych liczebnościach) dałyby te same wartości statystyk. Nie byłyby one zainfekowane błędem próbkowania. (Groves 1989: 240)

Innymi słowy, źródłem błędu statystycznego jest – poza częściową naturą badań reprezentatywnych – także zróżnicowanie jednostek w obrębie populacji.

Dla prostych prób losowych istnieją oczywiście „eleganckie" formuły pozwalające na wyznaczenie wielkości statystycznych błędów próbkowania. Zanim jednak wielkości te – dla estymatorów wskaźników struktury oraz średnich arytmetycznych – zostaną zdefiniowane w sposób formalny, krótkim komentarzem należy opatrzyć pojęcie *prostej próby losowej.* Ponieważ ten schemat losowania będzie jeszcze przedmiotem szczegółowej analizy w rozdziale trzecim, to na tym etapie dyskusji wystarczy podać podstawowe informacje o próbie prostej. W pierwszej kolejności warto odwołać się do pracy Sergeya Dorofeeva oraz Petera Granta (2006: 16), którzy zwracają uwagę na dwie definicyjne właściwości takich „prostych" schematów doboru, to znaczy *losowość* oraz *prostotę* (lub inaczej *brak złożoności* doboru). Losowość odnosi się do jednakowych szans selekcji jednostek z populacji do próby badawczej, z kolei prostota oznacza, że każda osoba dobierana jest do próby wprost z całego operatu jednostkowego, niezależnie od wyboru innej osoby lub zespołu osób. Pierwszy wymóg pociąga za sobą konieczność dostępu do operatów zawierających kompletne i adekwatne rejestry populacji, z kolei drugi oznacza, że wybór pewnych jednostek nie będzie zwiększać lub zmniejszać szans selekcji innych. Co oczywiste, prosta próba losowa jest nieobciążona, to znaczy: nie jest obciążona błędem systematycznym, choć istnieją oczywiście także inne – alternatywne względem losowania prostego – zasady selekcji, które również umożliwiają dobór nieobciążonych prób losowych, niespełniających jednak definicyjnych kryteriów schematu losowania prostego. G. Lissowski, J. Haman oraz M. Jasiński (2008) wskazują z kolei na nieco inne właściwości prostego doboru losowego, podkreślając, że musi on spełniać cztery podstawowe warunki, to jest być: (1) indywidualny, czyli umożliwiać losowanie pojedynczych elementów popu-

lacji, a nie zespołów jednostek, (2) jednostopniowy, czyli umożliwiać wylosowanie próby w jednym etapie, (3) nieograniczony, czyli dawać możliwość losowania elementów do próby z całej populacji, a nie oddzielnie z jej poszczególnych części, oraz (4) zapewniać jednakowe prawdopodobieństwa wyboru. W konsekwencji wszystkie możliwe do wylosowania w ten sposób próby o ustalonej liczbie elementów będą miały jednakowe prawdopodobieństwa selekcji (por. Lissowski i in. 2008: 513).

Jeżeli zatem przyjmie się, że z *N*-elementowej populacji wylosowano w sposób prosty próbę o liczebności *n* elementów (frakcja elementów wybranych do próby wynosi $f = n/N$), to dla estymatora wskaźnika struktury wielkość statystycznego błędu próbkowania przy danym poziomie istotności wynoszącym α można zapisać w postaci wyrażenia:

(II.1.) $$d = \pm z_{\frac{\alpha}{2}} \sqrt{(1-f)\frac{p(1-p)}{n-1}},$$

gdzie:

- $(1-f)$ jest korektą wprowadzoną z uwagi na skończoną liczbę elementów w populacji;
- p oznacza wartość parametru wskaźnika struktury w całej populacji;
- dla przyjętego poziomu istotności wnioskowania α, $z_{\frac{\alpha}{2}}$ jest wartością krytyczną rozkładu normalnego standaryzowanego, taką, dla której $P\left(-z_{\frac{\alpha}{2}} \leq z \leq +z_{\frac{\alpha}{2}}\right) = 1-\alpha$, przy czym $z \sim N(0,1)$;
- wyrażenie $\sqrt{(1-f)\frac{p(1-p)}{n-1}}$ jest niczym innym, jak dobrze znanym błędem standardowym estymatorów parametru *p*, ustalonym dla wszystkich różnych *n*-elementowych prostych prób losowych wybranych z populacji o liczebności *N* elementów[16] (por. Weisberg 2005: 230; Groves i in. 2004: 100).

Podobnie można zapisać wielkość błędu statystycznego dla estymatora średniej arytmetycznej, który przy danym poziomie istotności równym α przyjmuje postać:

(II.2.) $$d = \pm z_{\frac{\alpha}{2}} \sqrt{1-f}\,\frac{\sigma}{\sqrt{n}},$$

gdzie:

- σ jest wielkością odchylenia standardowego wszystkich wartości badanej cechy w populacji (z dzielnikiem $N-1$) od ich średniej arytmetycznej μ,

[16] Liczbę wszystkich *n*-elementowych prób badawczych dobranych w sposób losowy z *N*-elementowej populacji, wyznacza się ze wzoru $\binom{N}{n} = \frac{N!}{n!(N-n)!}$.

- a wyrażenie $\sqrt{1-f}\frac{\sigma}{\sqrt{n}}$ jest błędem standardowym wszystkich wartości estymatorów średniej arytmetycznej, możliwych do wyznaczenia z różnych *n*-elementowych prostych prób losowych (por. Groves i in. 2006: 99; Weisberg 2005: 229).

W praktyce badań sondażowych statystyczny błąd próbkowania wykorzystuje się do wyznaczenia minimalnych wielkości prób badawczych, które przy zachowaniu zasad losowania prostego (oraz przy założeniu, że $f \sim 0$) należałoby dobrać z całej populacji, aby nie przekroczyć wielkości zakładanego *a priori* błędu statystycznego. Jeżeli zatem badacz ustali maksymalną wielkość błędu statystycznego d oraz przyjmie określony poziom istotności wnioskowania α, to dla estymatora wskaźnika struktury minimalną wielkość próby (oznaczoną jako n_{eff}) wyznacza się z nierówności:

(II.3.) $$n_{eff} \geq z^2_{\frac{\alpha}{2}} \frac{p(1-p)}{d^2},$$

natomiast dla estymatora średniej arytmetycznej (przy założeniu normalności rozkładu zmiennej) z nierówności:

(II.4.) $$n_{eff} \geq z^2_{\frac{\alpha}{2}} \frac{\sigma^2}{d^2}.$$

Ponieważ dla estymacji wskaźników struktury minimalne wielkości prób badawczych są zazwyczaj większe od minimalnych liczebności prób wymaganych dla estymacji średnich arytmetycznych, to nierówność (II.3.) można wykorzystać do ustalenia minimalnej liczebności próby badawczej dla wszystkich szacowanych parametrów w ogóle. Wystarczy zauważyć, że przy maksymalnym poziomie błędu statystycznego d oraz przyjętym poziomie istotności α wyrażenie po prawej stronie nierówności (II.3.) przyjmuje największą wartość dla $p = 0{,}5$ (por. Weisberg 2005: 230). Innymi słowy, dla tej najmniej korzystnej sytuacji minimalną wielkość próby można wyznaczyć z nierówności:

(II.5.) $$n_{eff} \geq z^2_{\frac{\alpha}{2}} \frac{1}{4d^2},$$

przyjmując założenie, że dobór próby przeprowadzony zostanie zgodnie z zasadami schematu losowania prostego z populacji o nieskończonej liczbie elementów (lub populacji tak licznej, że $f \sim 0$).

II.1.2. Zmiana precyzji wnioskowania wynikająca z przyjętego schematu doboru próby (*design effect*)

Możliwość wyznaczenia wielkości błędów statystycznych podlega jednak pewnym istotnym ograniczeniom. Najważniejsza pozostaje przy tym konieczność zachowania reguł losowania prostego. Co za tym idzie, wielkości błędów

statystycznych zależą nie tylko od liczebności próby, wariancji wyników lub też ustalonych przez badacza współczynników ufności, ale także, a może przede wszystkim, od przyjętego schematu dobru próby. Innymi słowy, jeżeli badacz nie jest w stanie (lub też nie chce) dobrać respondentów zgodnie z losowaniem prostym, to każdy inny schemat doboru próby pociąga za sobą utratę (lub w niewielu przypadkach wzrost) precyzji estymacji (por. Dorofeev i in. 2006: 4). Do oceny przyrostu wariancji służy tak zwany miernik DEFF (od angielskiego wyrażenia *design effect*), zdefiniowany przez L. Kisha w 1965 roku jako „stosunek wariancji wyników w dobranej próbie badawczej oraz wariancji w prostej próbie losowej" (Kish 1965: 258). Wprawdzie autor monografii *Survey Sampling* rozpatrywał wskaźnik DEFF głównie w odniesieniu do estymatorów parametrów średniej arytmetycznej, jednakże można dokonać prostego uogólnienia tego miernika na dowolny estymator $\hat{\theta}$ parametru θ. Wówczas w sensie formalnym DEFF jest ilorazem wariancji wszystkich estymatorów wyznaczonych na podstawie pomiaru różnych *n*-elementowych prób badawczych dobranych w oparciu o dany schemat losowania oraz wariancji estymatorów, które można by wyznaczyć, gdyby tylko zachowano reguły losowania prostego. Miara DEFF przyjmuje postać wyrażenia[17]:

(II.6.) $$\mathrm{DEFF} \stackrel{\text{def}}{=} \frac{\mathrm{Var}_{SAMPL}(\hat{\theta})}{\mathrm{Var}_{srs}(\hat{\theta})}$$ (por. Gabler i in. 2006: 4),

gdzie:

- $\mathrm{Var}_{SAMPL}(\hat{\theta})$ jest wariancją estymatorów ustalonych na podstawie wszystkich możliwych *n*-elementowych prób badawczych dobranych według przyjętego schematu losowania;
- $\mathrm{Var}_{srs}(\hat{\theta})$ jest wariancją estymatorów wyznaczonych na podstawie wszystkich możliwych *n*-elementowych prób badawczych dobranych według losowania prostego.

---

[17] Uogólnioną wersję wzoru (II.6.) odnaleźć można w artykule Inho Parka oraz Hyunshika Lee opublikowanym w 2004 roku w czwartym numerze czasopisma „Survey Methodology" (por. Park i in. 2004: 5). Wspomniani autorzy, powołując się na argumenty przytoczone przez Särndala i in. (1992: 54), dochodzą do wniosku, że w procesie estymacji wartości parametrów można stosować różne procedury wyznaczania wielkości statystyk opisowych. Dla przykładu, w próbie dobieranej według określonych przez badacza zasad selekcji estymator może być ważony, a w prostej próbie losowej – nieważony. We wszystkich takich przypadkach „DEFF [...] nie zależy jedynie od schematu doboru próby, ale odnosi się także do szczególnego procesu estymacji określonego parametru w populacji" (Park i in. 2004: 5). Innymi słowy, przyrost wariancji może być spowodowany nie tylko schematem doboru próby, ale także przyjętą formułą wyznaczania wartości estymatorów. Choć oczywiście uwaga ta jest słuszna, to jednak utrata precyzji estymacji na skutek adjustacji danych rozpatrzona zostanie (poza ważeniem rekompensującym nierówne szanse selekcji do próby badawczej) dopiero w sekcji II.1.5., gdzie zdefiniowana będzie miara (oparta na tej samej idei co miernik DEFF), służąca oszacowaniu przyrostu wariancji na skutek ważenia wartości wyników pomiaru.

W literaturze badań sondażowych można jednak odnaleźć inne sposoby definiowania miernika DEFF, które, poza konsekwencjami wynikającymi z wyboru pewnego schematu doboru próby (warstwowego oraz wiązkowanego), uwzględniają także wzrost wariancji, będący efektem ważenia wyników rekompensującego na przykład nierówne prawdopodobieństwa selekcji jednostek do próby badawczej. L. Kish (1987: 202), a za nim inni autorzy: Lee (2012: 19; Biemer (2011: 214), Lynn i in. (2007: 107–124), Gabler i in. (2006: 115–116), Gabler i in. (1999: 105–107), oraz Park i in. (2004: 4–14), wprowadzają miarę „całkowitego" efektu doboru próby będącą oszacowaniem DEFF, która przyjmuje postać iloczynu:

$$\text{DEFF}_{TOTAL} = \text{DEFF}_p \times \text{DEFF}_s \times \text{DEFF}_c, \tag{II.7.}$$

gdzie:

- DEFF_p (inne określenie to *VIF*) jest miernikiem przyrostu wariancji, będącego skutkiem losowania z nierównymi prawdopodobieństwami selekcji. Wskaźnik ten zdefiniowany zostanie w sekcji II.1.5. tego rozdziału,
- DEFF_s jest miarą efektu schematu losowania z populacji rozwarstwionych,
- DEFF_c jest miarą efektu losowania wiązek respondentów.

Właściwości tych mierników będą przeanalizowane dokładnie w rozdziale IV, poświęconym w całości schematom doboru prób badawczych.

Powracając do wzoru (II.6), należy zauważyć, że mianownik wyrażenia DEFF (czyli $\text{Var}_{srs}(\hat{\theta})$), jest niczym innym jak kwadratem wyprowadzonego już wcześniej błędu standardowego estymatorów $\hat{\theta}$. Zatem dla parametru wskaźnika struktury wartość wariancji wszystkich estymatorów tego parametru wyznaczonych z *n*-elementowych prób badawczych, dobranych zgodnie z losowaniem prostym z *N*-elementowej populacji, można zapisać jako:

$$\text{Var}_{srs}(\hat{p}) = (1 - f)\frac{p(1-p)}{n-1}, \tag{II.8.}$$

z kolei dla estymatorów średniej arytmetycznej – w postaci wyrażenia:

$$\text{Var}_{srs}(\hat{\mu}) = (1 - f)\frac{\sigma^2}{n}, \tag{II.9.}$$

gdzie: $(1 - f)$, n, N, p oraz σ oznaczają te same wielkości, jak we wzorze (II.1.) oraz (II.2.). W praktyce, przy bardzo licznych populacjach, korekta $(1 - f)$ może być pominięta, gdyż frakcja elementów wylosowanych do próby stanowi niewielką część promila całej populacji, a $(1 - f)$ jest bliskie wartości 1. Jeżeli natomiast $n = N$, co oznacza badania przeprowadzane na całej populacji bez wyjątku, to wówczas $f = 1$, a $(1 - f) = 0$, czyli $\text{Var}_{srs}(\hat{p}) = 0$ oraz $\text{Var}_{srs}(\hat{\mu}) = 0$. W tym drugim przypadku wartość estymatora (tylko jednego możliwego do wyznaczenia) jest dokładną wartością parametru, stąd wariancja estymatorów wynosi zero.

Estymacja wskaźnika DEFF (ogólnie DEFF_{TOTAL}) ma swoje przełożenie na konkretne działania podejmowane w ramach ustalania minimalnych wielkości prób badawczych. Doskonale znane jest pojęcie *efektywnej wielkości próby badawczej* (por. na przykład Biemer 2011: 232; Kohler 2007: 56; Lynn i in. 2007: 112; Gabler i in. 2006: 4; Groves i in. 2004: 108), oznaczające przeskalowaną liczebność próby dobranej zgodnie z określonym schematem losowania, do odpowiadającej jej liczebności prostej próby losowej. Innymi słowy, efektywna wielkość próby odpowiada takiej teoretycznej liczebności prostej próby losowej, która umożliwia estymację wartości parametrów z takim samym poziomem błędu statystycznego, jak liczebność prób dobranych zgodnie z ustalonym przez badacza schematem (por. Dorofeev i in. 2006: 90). Oznacza to, że działania badawcze, zmierzające do wyznaczenia minimalnych wielkości prób badawczych, wymagają *de facto* ich urealnienia do liczebności efektywnych, gdyż w przeciwnym razie otrzymany poziom błędu statystycznego może przekraczać przyjęte wartości maksymalne. Najczęściej wyznacza się zatem pożądaną efektywną wielkość prostej próby losowej (n_{eff}), oblicza (przewidywaną) skalę przyrostu wariancji dla wybranego schematu losowania próby (DEFF_{TOTAL}) oraz ustala wielkość próby n_{SAMPL}, równoważną liczebności n_{eff}. Wielkość n_{SAMPL} można wyznaczyć przy pomocy prostej formuły:

(II.10.) $$n_{SAMPL} = n_{eff} \times \mathrm{DEFF}_{TOTAL}.$$

Zupełnie analogicznie określić można efektywną wielkość zrealizowanej już próby badawczej o liczebności n_{SAMPL}, dobranej zgodnie z określonym schematem losowania. Wystarczy tylko przekształcić wzór (II.10.) do postaci:

(II.11.) $$n_{eff} = \frac{n_{SAMPL}}{\mathrm{DEFF}_{TOTAL}}.$$

Świetną egzemplifikacją tych zależności są dane metodologiczne Europejskiego Sondażu Społecznego, prowadzonego na populacjach mieszkańców kilkudziesięciu krajów europejskich oraz Izraela. W poniższej tabeli zestawiono wielkości krajowych prób badawczych z wartościami współczynników DEFF_p oraz efektywnymi liczebnościami prób dla danych z 5. rundy ESS-u[18].

Wartości mierników DEFF w poszczególnych krajach nie można oczywiście rozpatrywać w oderwaniu od zastosowanych schematów losowania respondentów. Niezwykle ciekawych informacji w tym zakresie dostarcza szczegółowa analiza raportu dokumentującego metodologiczną stronę badań ESS5-2010. Przeglądając opis procedur próbkowania, można zauważyć, że we wszystkich krajach, w których zastosowano operaty umożliwiające losowanie prób imien-

[18] Podane w raporcie metodologicznym dane dotyczące mierników DEFF uwzględniają jedynie efekt nierównych prawdopodobieństw selekcji jednostek do próby badawczej, stąd oznaczenie DEFF_p.

Tabela II.1. Wartości mierników $DEFF_p$ oraz efektywne wielkości prób badawczych w badaniach Europejskiego Sondażu Społecznego (ESS5-2010)

Nazwa kraju[19]	n_{resp} i)	$DEFF_p$ ii)	n_{eff_deffp} iii)	Nazwa kraju	n_{resp}	$DEFF_p$	n_{eff_deffp}
Belgia	1704	1,00	1704	Czechy	2386	1,20	1988
Dania	1576	1,00	1576	Holandia	1829	1,20	1524
Estonia	1793	1,00	1793	Wielka Brytania	2422	1,21	2001
Finlandia	1078	1,00	1078	Grecja	2715	1,22	2225
Norwegia	1548	1,00	1548	Bułgaria	2434	1,23	1978
Słowenia	1403	1,00	1403	Chorwacja	1649	1,24	1329
Szwajcaria	1506	1,00	1506	Izrael	2294	1,26	1820
Szwecja	1497	1,00	1497	Francja	1728	1,27	1360
Węgry	1561	1,00	1561	Irlandia	2576	1,27	2028
Polska	1751	1,01	1733	Rosja	2595	1,31	1980
Hiszpania	1885	1,12	1683	Portugalia	2150	1,34	1604
Niemcy	3031	1,12	2706	Słowacja	1856	1,85	1003
Cypr	1083	1,15	941	Ukraina	1931	1,97	980

Źródło: *European Social Survey 2012:* ESS5-2010, Documentation Report. Edition 2.0, Bergen, European Social Data Archive, Norwegian Social Science Data Service

i) Liczebność zrealizowanej próby badawczej.

ii) Miernik efektywności doboru próby z nierównymi prawdopodobieństwami selekcji.

iii) Efektywne wielkości prób badawczych (liczebność próby zrealizowanej umniejszona o efekt losowania z nierównymi prawdopodobieństwami selekcji).

nych (niezależnie od tego, czy próba była dzielona na warstwy, czy też nie), wartości wskaźników $DEFF_p$ były najmniejsze. Dotyczy to wszystkich dziewięciu krajów o wartościach $DEFF_p$ równych jedności, a także Polski (tutaj wartość miary tylko nieznacznie przekraczająca jeden) oraz Hiszpanii i Niemiec (w obu krajach $DEFF_p$ = 1,12). Natomiast we wszystkich pozostałych krajach, w których nie stosowano operatów imiennych, ale inne rejestry, umożliwiające na przykład wylosowanie ulic, budynków lub też gospodarstw domowych, wartości $DEFF_p$ były już zdecydowanie większe. W wielu krajach konieczne było sporządzanie dodatkowej listy mieszkań oraz określanie reguł wyboru gospodarstw domowych, nie mówiąc już o doborze konkretnych respondentów w oparciu o siatkę Kisha, bądź też datę najbliższych urodzin.

Wybór określonego schematu doboru próby pociąga więc za sobą pewne konkretne konsekwencje w postaci utraty precyzji wnioskowania statystycznego. Poza sytuacjami, dla których wartość wskaźnika $DEFF_p$ wynosi 1 (losowanie proste lub jakiś ekwiwalenty względem prostej próby losowej schemat doboru jednostek), wykorzystane w większości krajów procedury losowania obniżały efektywność prób badawczych. Szczególnie interesujące w tym kontekście wy-

[19] Dane uporządkowane zgodnie z wartością miary $DEFF_p$.

dają się przypadki Słowacji ($DEFF_p$ = 1,85) oraz Ukrainy ($DEFF_p$ = 1,97), w których odnotowano zdecydowanie najwyższe wartości miary $DEFF_p$ wśród wszystkich krajów biorących udział w ESS5-2010. Wykorzystując prostą formułę obliczeniową: $n_{\text{eff\_deffp}} = n_{\text{resp}} \times \text{DEFF}_p^{-1}$, można przeskalować liczebności zrealizowanych prób badawczych do równoważnych im liczebności prostych prób losowych. W przypadku Słowacji próba zrealizowana (licząca 1856 respondentów) miała efektywność prostej próby losowej mniejszej o ponad 850 osób, z kolei na Ukrainie badania przeprowadzone ze zbiorem 1931 osób miały efektywność próbkowania prostego o liczebności mniejszej o prawie 950 respondentów. Innymi słowy, gdyby badacze dysponowali operatami umożliwiającymi losowanie proste, to wystarczyłoby przebadanie 1003 osób na Słowacji oraz 980 osób na Ukrainie, aby otrzymać precyzję pomiaru taką, jak w próbach o znacznie większych liczebnościach. Wprawdzie nie jest niczym nadzwyczajnym wykorzystywanie odbiegających od losowania prostego schematów doboru respondentów, jednak w praktyce nie powinny one powodować tak dużego przyrostu wariancji. Warto jednak pamiętać, że na jakość procesu badawczego nie należy patrzeć z perspektywy pojedynczych źródeł uchybień, lecz całościowo, uwzględniając wzajemny wpływ wielu klas błędów.

II.1.3. Błąd pokrycia / błędy operatu losowania (*coverage / frame error*)

Błąd pokrycia jest konsekwencją rozbieżności występujących pomiędzy zbiorem wszystkich elementów należących do badanej populacji oraz możliwym do zastosowania operatem losowania, to jest rejestrem jednostek, na przykład osób lub też gospodarstw domowych, które można wykorzystać do wyboru reprezentantów owej populacji (por. Biemer 2010a: 33; Groves i in. 2004: 54). W książce *Survey Sampling* z 1965 roku L. Kish – rozważając potencjalne problemy, na które napotka badacz, decydując się na losowanie próby – przedstawia opis operatu „idealnego", uwypuklając jego główne cechy definicyjne:

> Operat jest idealny, jeżeli każdy [bez wyjątku – P.J.] element [populacji – P.J.] pojawia się w nim oddzielnie, jeden, i tylko jeden raz, oraz jeśli nie zawiera on nic innego [poza elementami należącymi do populacji – P.J.]. (Kish 1965: 53)

Każde odstępstwo od opisanej przez L. Kisha sytuacji „idealnej" może być przy tym źródłem poważnych błędów, skutkującym obniżeniem poziomu reprezentatywności próby badawczej. Dla przykładu, jeżeli badacz nie posiada kompletnej listy jednostek statystycznych wchodzących w skład populacji, to, ze względu na niepełne pokrycie oraz zerowe szanse doboru do próby, pewna część jednostek pozostaje niereprezentowana, co może skutkować wypacze-

niem wyników pomiaru. Chociaż problem ten jest przedmiotem szczegółowych analiz w odniesieniu do badań sondażowych realizowanych technikami wywiadów telefonicznych (CATI) lub ankiet internetowych (CAWI)[20], to jednak pozostaje też w obrębie zainteresowania badaczy skupiających uwagę na wywiadach prowadzonych tradycyjnymi metodami gromadzenia danych opartymi na bezpośrednim kontakcie z respondentem (PAPI, CAPI). Innymi słowy, choć rzeczywiście w przypadku wywiadów telefonicznych (por. Vicente i in. 2009: 105–111; Curtin i in. 2005: 90–95) oraz ankiet internetowych (por. Dever i in. 2008: 47–62; Heerwegh i in. 2008: 836–846; Couper i in. 2007: 131–148; Vehovar et al. 2002: 230–232; Best i in. 2002: 75–77) słabości operatów doboru prób są najbardziej widoczne, to jednak mogą one stanowić źródło problemów we wszystkich typach surveyów, niezależnie od zastosowanych procedur gromadzenia danych[21].

W odniesieniu do operatu losowania identyfikuje się także – poza niepełnym jej pokryciem – trzy inne źródła błędów wpływające na jakość rejestrów wykorzystywanych w doborze próby. Pierwszy z nich jest konsekwencją nadmiarowego pokrycia, to jest zawierania jednostek nienależących do badanej populacji. Na przykład, jeżeli sondaż ma być prowadzony na populacji mieszkańców jakiegoś miasta zameldowanych tam na pobyt stały, a badacz dysponuje operatem adresowym zawierającym również mieszkańców zameldowanych czasowo, to te wszystkie „dodatkowe" jednostki, które w istocie do badanej populacji nie należą, stanowią odstępstwo od sytuacji pożądanej. Ponieważ jednostki te nie leżą w kręgu zainteresowania badacza, to powinny być zidentyfikowane przed etapem doboru próby oraz usunięte z operatu. Najczęściej uda-

---

[20] Głównym problemem z wykorzystaniem w sondażach wywiadów telefonicznych oraz ankiet internetowych jest odpowiedni poziom telefonizacji oraz komputeryzacji populacji będącej przedmiotem badania, który wyklucza pewne kategorie z możliwości wzięcia udziału w badaniu. Dla przykładu, analiza przydatności techniki wywiadu telefonicznego do badań reprezentatywnych, przeprowadzona przez P.B. Sztabińskiego, jest jednoznacznie negatywna. Autor ten wskazuje, *że* „[p]roblem reprezentatywności prób do badań występuje […] nawet przy wyposażeniu w telefony przewodowe przekraczającym 90% gospodarstw domowych. Związany jest on z nierównomiernym ich wyposażeniem w zależności od regionu zamieszkiwania […]. W krajach, w których dysproporcje te są bardzo silne, jak na przykład w Polsce […] w ogóle nie jest możliwe prowadzenie badań na próbach generalnych ludności" (Sztabiński P. 2001: 67). Podobny wniosek można postawić w odniesieniu do techniki ankiety internetowej, gdyż zgodnie z danymi Diagnozy Społecznej za rok 2011, w dostęp do Internetu wyposażonych jest nieco ponad 51% gospodarstw domowych, przy czym jednocześnie z Internetu korzysta 60% Polaków (por. Batorski 2011: 299).

[21] W badaniach Europejskiego Sondażu Społecznego, opartego w większości na wywiadach bezpośrednich, wykorzystuje się kilka odmiennych rodzajów operatów losowania obejmujących rejestry (a) jednostek, (b) gospodarstw domowych, lub też (c) punktów adresowych. W każdym kraju biorącym udział w badaniach ESS wyboru odpowiedniego operatu losowania dokonuje się w oparciu o kryterium kompletności danych populacyjnych, tak aby wykorzystany rejestr możliwie najpełniej pokrywał badaną populację (por. *European Social Survey. ESS5-2010. Documentation Report ed. 2.0*).

je się je oznaczyć dopiero podczas kontaktu z respondentem, co nie wpływa mimo wszystko w żaden znaczący sposób na wyniki uzyskanego pomiaru[22]. Można jednak wyobrazić sobie sytuację, w której badaczowi (lub ankieterowi) nie uda się zidentyfikować jednostek błędnie przypisanych do populacji. W efekcie proces estymacji może być wypaczony przez tę „nadmiarową" część respondentów (por. na przykład Biemer i in. 2003: 65). Co ciekawe, niektórzy autorzy, na przykład Stoop i in. (2010), za Benthlehemem i in. (1986), lokują błąd niepełnego pokrycia w klasie błędów oddziałujących na reprezentatywność prób badawczych, ale błędy wynikające z nadmiarowego pokrycia populacji, zaliczają już do klasy błędów oddziałujących na dokładność prowadzonego pomiaru (por. Stoop i in. 2010: 4).

Drugie źródło problemów jest efektem multiplikowania szans doboru jednostek należących do populacji. Dla przykładu, gospodarstwa domowe posiadające więcej niż jedną linię telefoniczną będą miały większą szansę znalezienia się w próbie, niż gospodarstwa z dostępem do jednego tylko numeru. Podobnie, stosując operaty losowania zawierające zarówno abonentów telefonii stacjonarnej, jak i mobilnej, daje się większą szansę wyboru tym osobom, które posiadają telefon komórkowy, a ich gospodarstwo domowe podłączone jest do sieci telefonii stacjonarnej. Co ciekawe, w takich wypadkach owe nierówne prawdopodobieństwa doboru jednostek daje się wyznaczać oraz kompensować w postrealizacyjnym ważeniu danych (por. Lohr 2011: 197–213; Häder i in. 2010: 14–17; Haines i in. 2000: 121–129).

Wreszcie trzecie źródło problemów związanych z operatem doboru próby wynika z braku ekwiwalentności przedmiotowej pomiędzy pożądanymi jednostkami populacji oraz dostępnymi jednostkami losowania. Sytuacja taka występuje wtedy, gdy jednostka operatu grupuje kilka jednostek populacji. Typowym przykładem takiego zjawiska pozostają operaty numerów telefonicznych, a także adresowe próby gospodarstw domowych wykorzystywane do losowania jednostek indywidualnych[23]. We wszystkich takich przypadkach jednostki operatu nie przystają do jednostek populacji, a wybór konkretnego respondenta wymaga sporządzenia (kompletnej) listy jednostek wchodzących w jego skład oraz określenia zasad wewnątrzzespołowego doboru osób do wywiadu. Innymi słowy, problemy wynikające z niedoskonałości operatów mogą być powiązane ze schematami losowania prób badawczych, a nawet z problemami jednostek niedostępnych, czy też wreszcie z trafnością mierników oraz rzetel-

---

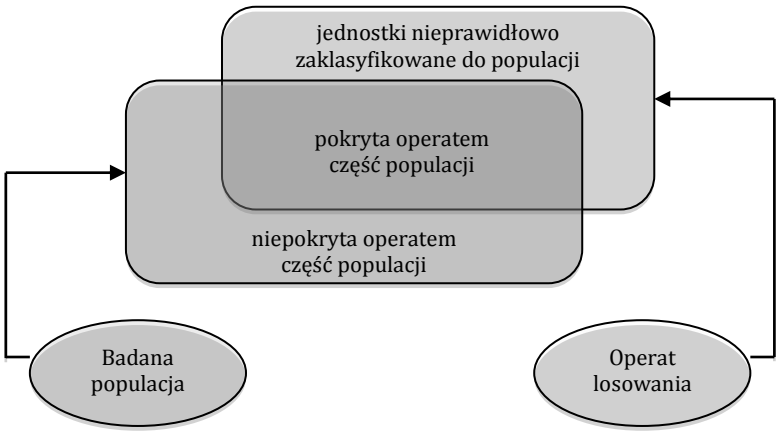
[22] Zmniejsza się jednak wielkość zrealizowanej próby badawczej, a zatem przyrasta poziom błędu statystycznego.

[23] Z sytuacją taką będzie się miało do czynienia wtedy, gdy populacją są pojedyncze osoby, a operat pozwala na dobór gospodarstwa domowego, numeru telefonicznego czy też adresu, czyli na wylosowanie grupy jednostek.

nością odpowiedzi udzielanych przez członków wylosowanych gospodarstw domowych (por. Biemer i in. 2003: 66).

W kontekście opisu i analizy źródeł błędów pokrycia należy zauważyć, że, pomimo iż w literaturze metodologicznej badacze identyfikują zazwyczaj (omówione powyżej) cztery typy błędu (systematycznego), to jest niepełne i nadmiarowe pokrycie, multiplikowanie jednostek oraz brak ekwiwalentności przedmiotowej operatu i populacji, to jednak wielkość błędu wyznaczana jest najczęściej tylko w odniesieniu do niepełnego pokrycia[24]. Tym samym pomija się drugie, trzecie oraz czwarte źródło błędów (por. Biemer 2010b: 840; Groves i in. 2004: 55; Biemer i in. 2003: 69). Jest to konsekwencją tego, że znaczna część problemów mających swe źródło w operatach losowania rozwiązywana jest przez wykorzystanie odpowiedniego schematu losowania próby badawczej lub też na etapie postsurveyowego ważenia danych. Niemniej jednak, mówiąc o systematycznym błędzie operatu, ma się najczęściej na myśli te ułomności rejestrów populacji, które wynikają z rozbieżności w jej pokryciu.

Wielkość błędu pokrycia definiowana jest przy tym jako różnica pomiędzy wartością parametru w części populacji objętej operatem losowania oraz wartością tego parametru w całej populacji. Jego wielkość zapisać można również



Ryc. II.1. Pokrycie populacji operatem losowania

Źródło: opracowanie własne na postawie Groves i in. 2004: 54

[24] Nie jest to jednak reguła, bowiem w monografii *Nonsampling Error in Survey* autorstwa Judith T. Lessler oraz Wiliama D. Kalsbeeka (1992: 16) odnajdujemy definicję błędu nadmiarowego pokrycia populacji. Wielkość błędu definiowana jest jako iloczyn proporcji elementów nadmiarowych oraz różnicy w wartościach parametrów odpowiadających warstwie jednostek z populacji oraz spoza populacji.

nieco inaczej: jako iloczyn frakcji jednostek populacji niepokrytych przez operat losowania oraz różnicy pomiędzy wartością parametru w pokrytej oraz niepokrytej przez operat części populacji. Jeżeli zatem przyjmie się, że populacja składa się z N jednostek oraz N_C jednostek jest zawartych w operacie losowania, a N_{NC} elementów jest pominiętych przez ten operat (przy czym $N = N_C + N_{NC}$), to dla parametru wskaźnika struktury wartość błędu pokrycia wynosi:

(II.12.) $$\mathrm{B}_{NC} \stackrel{\text{def}}{=} p_C - p$$

lub

(II.12'.) $$\mathrm{B}_{NC} \stackrel{\text{def}}{=} w_{NC}(p_C - p_{NC}),$$

z kolei dla parametru średniej arytmetycznej:

(II.13.) $$\mathrm{B}_{NC} \stackrel{\text{def}}{=} \mu_C - \mu$$

lub

(II.13'.) $$\mathrm{B}_{NC} \stackrel{\text{def}}{=} w_{NC}(\mu_C - \mu_{NC}).$$

W przedstawionych wzorach p_C oraz μ_C oznaczają wartości parametrów wskaźnika struktury oraz średniej arytmetycznej w pokrytej operatem losowania części populacji, p_{NC} oraz μ_{NC} odpowiadają wartościom parametrów w niepokrytej operatem losowania części populacji, z kolei p oraz μ odnoszą się do wartości obu parametrów w całej populacji (łącznie pokrytej oraz niepokrytej przez operat), natomiast $w_{NC} = N_{NC}/N$ wyznacza frakcję elementów niepokrytych przez operat losowania, a $w_C = N_C/N$ określa część populacji pokrytej operatem doboru próby.

Z powyższych definicji można łatwo odczytać, że jeśli badacz dysponuje operatem zawierającym prawie wszystkie jednostki populacji[25], to w zasadzie nie musi się zajmować błędem pokrycia, gdyż jego oddziaływanie na całkowity

[25] Niezwykle pouczające w tym względzie okazują się dane metodologiczne Europejskiego Sondażu Społecznego, zawierające szczegółowe informacje o krajowych operatach doboru prób badawczych. Wprawdzie dane o pokryciu populacji nie są dostępne w dokumentacji projektowej ESS dla wszystkich państw, jednak nawet z tych niepełnych danych można odczytać, że wykorzystane operaty (imienne oraz gospodarstw domowych) pokrywają w większości państw prawie całkowicie badane populacje. Dla przykładu w Chorwacji, gdzie operatem jest spis gospodarstw domowych, pokryte jest, wedle szacunków, od 94,8% do 97,9% populacji. Podobnie jest też na Słowacji, gdzie operat adresowy pozwala na dotarcie do co najmniej 97% mieszkańców kraju. Z kolei w Danii oraz Słowenii, gdzie losowanie respondentów odbywa się w oparciu o rejestry indywidualne (podobne do polskiego repozytorium PESEL), pokrywają one odpowiednio 99,9% oraz 99,0% populacji (por. *European Social Survey*...: ESS5-2010). Co ciekawe, o ile w odniesieniu do stopnia realizacji próby badawczej oraz efektywności schematu doboru próby wymaga się od krajowych koordynatorów tych badań przedstawienia szczegółowych danych zarówno o ilościowym udziale jednostek niedostępnych w próbie, jak też o nierównych prawdopodobieństwach selekcji pewnych kategorii jednostek, to dane o stopniu pokrycia populacji przez operat losowania nie muszą być już upubliczniane.

błąd pomiaru będzie niewielkie, i to nawet przy znacznej odmienności jednostek zawartych w populacji oraz w operacie losowania. Wszystko zależy więc zarówno od jakości wykorzystanego operatu, jak też od odsetka populacji, jaką on pokrywa. Warto także zwrócić uwagę na fakt, że w definiowaniu błędu pokrycia (inaczej niż w innych typach błędów) wykorzystuje się pojęcie parametru, a nie estymatora, co oznacza, że błąd ten pozostaje w zasadzie niezależny od procesu badawczego[26]. Komplikacje te niezwykle trafnie ujęli autorzy monografii *Survey Methodology*, którzy we fragmencie poświęconym błędowi pokrycia stwierdzają, iż:

> [i]stnieje on [niezależnie od – P.J.] […] doboru próby, a więc nie jest problemem […] [charakterystycznym dla – P.J.] badań reprezentatywnych. Błąd ten istniałby też, gdybyśmy chcieli przeprowadzić badania pełne całej populacji, używając w tym celu tego samego [rejestru jednostek populacji – P.J.]. (Groves i in. 2004: 54)

Stwierdzenie tego faktu prowadzi do oczywistych trudności związanych z możliwością ustalenia wartości błędu pokrycia. Wymaga ono wiedzy o wszystkich elementach składających się na populację, nawet o tych jednostkach, które znajdują się poza dostępnym badaczowi operatem. Trudności te doskonale obrazuje konstatacja R. Grovesa (1989), który mówi, iż:

> Nawet jeśli ograniczymy naszą uwagę do średniej arytmetycznej oraz [innych estymatorów punktowych – P.J.], to w praktyce surveyowej nadal będzie bardzo trudno wyznaczyć błąd pokrycia. Dane z badań sondażowych nie dostarczają, same w sobie, informacji o [niepokrytej operatem losowania części populacji – P.J.], ani też o różnicy pomiędzy wartością estymowanej wielkości parametru w pokrytej oraz niepokrytej operatem części populacji. (Groves 1989: 119–120)

Proponowane w literaturze procedury wyznaczania błędu pokrycia są zatem najczęściej metodami działania *nie wprost*, wymagającymi dostępu do zewnętrznych repozytoriów statystycznych (por. na przykład Biemer 2010b: 840). Być może dlatego F. Sztabiński stwierdza, że „[…] błędy pokrycia populacji są najrzadziej uwzględnianym rodzajem błędów zawartych w wynikach badania" (Sztabiński F. 2011: 50)[27].

---

[26] Zresztą błąd pokrycia populacji stanowi jedno z podstawowych źródeł błędów w ramach spisów powszechnych (por. Mulry 2007: 345–370; Renaud 2007: 199–210; Lachapelle i in. 2000: 43–52).

[27] Zresztą do podobnych wniosków prowadzi lektura artykułu Daniela Kasprzyka oraz Lee Giesbrechta (2003: 342–263). Autorzy ci, odwołując się do pracy Atkinsona i in. (1999: 321–349), prześledzili noty metodologiczne raportów z badań surveyowych pod kątem wykazywanych w tych raportach źródeł błędów. Studia te wykazały, że w połowie opracowań wspomniano o błędach operatów (nie podając ponadto żadnych dokładniejszych informacji), a w jednym na

II.1.4. Błąd wynikający z niepełnej realizacji próby badawczej (*unit & item nonresponse error*)

Błąd związany z brakiem odpowiedzi (na pojedyncze pytania kwestionariuszowe lub wszystkie pytania w wywiadzie) jest wprawdzie tylko jednym z wielu źródeł błędów systematycznych, jednak w przekonaniu wielu metodologów pozostaje głównym źródłem całkowitego błędu pomiaru (por. Biemer 2010b: 821–825). Nabiera on szczególnego znaczenia w kontekście malejących, z roku na rok, odsetków realizacji prób badawczych zarówno w Polsce, jak i na świecie (por. Grzeszkiewicz-Radulska 2009: 162, 167) oraz związanego z tym ryzyka zniekształcenia wyników pomiaru. Najlepszym tego potwierdzeniem jest opinia F. Sztabińskiego, który stwierdza, że „błędy niezrealizowania części próby mają najpoważniejsze konsekwencje dla wyników badania" (Sztabiński F. 2011: 51). Przeglądając literaturę badań sondażowych, można zatem zauważyć, że wysiłki badaczy dużo częściej koncentrują się właśnie na ograniczaniu błędów braków odpowiedzi, niż na redukcji błędów pokrycia populacji operatem losowania, czy też błędów wynikających z przyjęcia określonego schematu doboru próby. Parafrazując wspomniane stwierdzenie F. Sztabińskiego (2011: 50) o nikłym zainteresowaniu badaczy błędami pokrycia, można w odniesieniu do błędów braku odpowiedzi powiedzieć, że są one *najczęściej uwzględnianymi rodzajami błędów zawartych w wynikach badania.* Wynika to przede wszystkim z doniosłości problematyki jednostek niedostępnych oraz wpływu, jaki niezbadana część próby może (choć nie musi) wywierać na poziom jej reprezentatywności. Co więcej, Ineke Stoop (2005) w drugim rozdziale swojej znanej książki *The Hunt for the Last Respondent* pokazuje wyraźnie, że wskaźnik poziomu realizacji próby badawczej, który zawiera informacje o frakcji przebadanej oraz nieprzebadanej części jednostek wylosowanych do badania, traktowany jest często jako kryterium oceny profesjonalizmu oraz jakości pracy zespołu realizującego badania sondażowe. Z kolei F. Sztabiński uznaje wskaźnik realizowalności próby za „jedną z najważniejszych miar jakości realizacji badania" (Sztabiński F. 2011: 51).

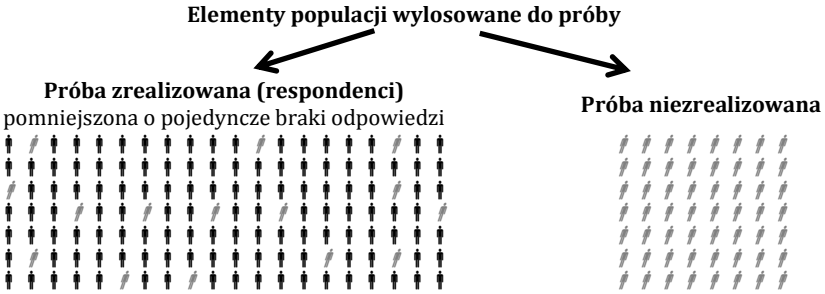
Osoby niedostępne nie stanowią – co oczywiste – kategorii homogenicznej. Przyczyn niezrealizowania wywiadów jest wiele, na niektóre z nich nie ma się wpływu (na przykład na błędne dane adresowe, zmianę miejsca zamieszkania wykluczającą respondenta z badanej populacji[28], wyjazd wylosowanych osób

---

sześć raportów pojawiała się dodatkowa informacja o stopniu pokrycia populacji operatem losowania, co w porównaniu z innymi źródłam błędów (zwłaszcza próbkowania oraz braków danych) stanowiło niewielki odsetek wskazań.

[28] W zasadzie należałoby uznać, że jednostki niedostępne to tylko i wyłącznie takie elementy próby – należące do populacji – z którymi nie udało się przeprowadzić wywiadu (por. Groves 1989: 137). Innymi słowy, wszystkie jednostki spoza populacji, które zostały do próby wylosowane, a w rzeczywistości nie powinny się w niej znaleźć, powinno klasyfikować się jako odrębną

poza miejsce zamieszkania w terminie przewidzianym na badanie, czy też na zgon lub chorobę respondenta), inne z kolei pozostają pod kontrolą badacza (na przykład odmowy udziału w badaniu, przerwanie wywiadu, brak kontaktu pomimo wielokrotnych prób dotarcia itp.). Co więcej, na wartość błędu braku odpowiedzi wpływają zarówno jednostki niedostępne (*unit nonresponse*), jak i sytuacje nieudzielenia przez respondenta odpowiedzi na pojedyncze pytania kwestionariuszowe (*item nonresponse*) (por. Biemer 2010a: 35; Grzeszkiewicz-Radulska 2009: 22–24; Mason i in. 2002: 150)[29]. Z pierwszą sytuacją ma się do czynienia wtedy, gdy wywiadu nie udaje się w ogóle przeprowadzić, z drugą natomiast, gdy zostaje on przeprowadzony tylko częściowo



Ryc. II.2. Jednostki niedostępne w próbie badawczej oraz pojedyncze braki odpowiedzi – próba wylosowana *vs.* próba zrealizowana

Źródło: opracowanie własne

W reprezentatywnych badaniach sondażowych jest zatem tak, że po zakończeniu okresu przewidzianego na realizację wywiadów wylosowaną do badania próbę można podzielić na dwie rozłączne kategorie: osób, z którymi udało się przeprowadzić wywiady w całości lub przynajmniej w części (respondenci), oraz osób, z którymi się to z różnych przyczyn nie udało (jednostki niedostępne)[30]. Ponieważ w tej pierwszej kategorii mieszczą się też braki odpowiedzi na pojedyncze pytania kwestionariuszowe, to wielkość próby zrealizowanej odnoszona powinna być *de facto* do każdego pytania z osobna. Innymi słowy, w *n*-elementowej próbie badawczej wylosowanej z populacji będącej przedmio-

kategorię jednostek, nie mającą wpływu na wskaźniki realizacji próby. Zagadnienia te omówione zostaną w ostatnim rozdziale pracy.

[29] W niektórych opracowaniach metodologicznych wskazuje się także na trzecią klasę jednostek niedostępnych wynikającą z tzw. częściowej niedostępności (oryg. *partial nonresponse*) (por. Brick i in. 2009: 163–164). Jest ona związana z niedostępnością jednostki losowanej w schemacie doboru dwustopniowego na przykład z operatu gospodarstw domowych lub niedostępnością pewnych jednostek w kolejnych odsłonach badania panelowego. W sensie substantywnym częściowa niedostępność jest jednak tym samym, co niedostępność na poziomie jednostki (tzn. *unit nonresponse*).

[30] W tym momencie założone będzie, że wszystkie jednostki niedostępne należą do populacji.

tem badania ma się zawsze n_R respondentów oraz n_{NR} jednostek niedostępnych (w tym pojedynczych braków odpowiedzi), przy czym $n = n_R + n_{NR}$[31]. Błąd braku danych definiuje się przeważnie jako różnicę pomiędzy wartością estymatora w warstwie respondentów oraz nieznaną wartością estymatora, jaką można byłoby wyznaczyć, gdyby badania udało się przeprowadzić ze wszystkimi tymi osobami, które należą do populacji oraz zostały wybrane do próby (por. Billiet i in. 2009: 6). Ponieważ różnica ta określa nielosowe i najczęściej systematyczne odstępstwo wyników obserwowanych od wyników rzeczywistych, to pozostaje źródłem systematycznego komponentu całkowitego błędu pomiaru.

Błąd wynikający z niepełnej realizacji próby można zapisać również nieco inaczej – jako funkcję frakcji wywiadów niezrealizowanych oraz różnicy w wartościach estymatorów w kategorii jednostek dostępnych i niedostępnych[32]. Odpowiednio zatem, dla estymatora parametru wskaźnika struktury błąd braku odpowiedzi przyjmuje postać dobrze znanych równoważnych sobie wzorów:

(II.14) $$\mathrm{B}_{NR} \stackrel{\text{def}}{=} \hat{p}_R - \hat{p}$$

(II.14'.) $$\mathrm{B}_{NR} \stackrel{\text{def}}{=} w_{NRR}(\hat{p}_R - \hat{p}_{NR}),$$

gdzie $\hat{p}$ jest symbolem oznaczającym nieznaną wartość estymatora dla całej wylosowanej n-elementowej próby zawierającej respondentów oraz jednostki niedostępne, $\hat{p}_R$ oznacza wartość estymatora wskaźnika struktury w warstwie respondentów, natomiast $\hat{p}_{NR}$ oznacza nieznaną wartość estymatora w warstwie jednostek niedostępnych (por. Jabkowski 2007: 73). Z kolei dla estymatora średniej arytmetycznej wielkość błędu można zapisać jako:

(II.15.) $$\mathrm{B}_{NR} \stackrel{\text{def}}{=} \hat{\mu}_R - \hat{\mu}$$

---

[31] Te dwie wielkości pozwalają ustalić wartość wskaźnika realizacji próby badawczej w jego najprostszej postaci zdefiniowanej wzorem $w_{RR} = n_R/n$ lub analogicznie, wartość wskaźnika frakcji niezrealizowanych wywiadów $w_{NRR} = n_{NR}/n$. Przyjęte w metodologii sposoby definiowania wskaźników realizacji próby badawczej (*response rate*) omówione zostaną w rozdziale poświęconym jednostkom niedostępnym.

[32] Przyjęty został tutaj deterministyczny model błędu braku danych (odpowiedni na poziomie próby), w którym zakłada się, że zbiór jednostek próby można podzielić na warstwę osób dostępnych i niedostępnych. Innym sposobem definiowania błędu braku odpowiedzi jest model probabilistyczny (odpowiedni dla poziomu populacji), w ramach którego przyjmuje się, iż każda jednostka ma określone – choć nieobserwowalne – prawdopodobieństwo zaklasyfikowania jej do grupy respondentów lub jednostek niedostępnych, pod warunkiem jej wylosowania. W takim przypadku błąd systematyczny, będący efektem niedostępności, definiuje się jako iloraz dwóch czynników, tj. (1) współczynnika kowariancji pomiędzy prawdopodobieństwem udziału w badaniu oraz wartościami analizowanej zmiennej, a także (2) oczekiwanego prawdopodobieństwa udziału w badaniu (por. Stoop i in 2010: 31; Brick i in. 2009: 170; Groves 2006: 648; Groves i in. 2004: 27; Bethlehem 2002: 276). Jeżeli zatem skłonność do udziału w badaniu będzie skorelowana w sposób istotny z wartościami mierzonych zmiennych, wówczas błąd systematyczny będzie znaczny. Kwestie te podjęte będą w rozdziale poświęconym jednostkom niedostępnym.

lub też równoważnie:

(II.15'.) $$B_{NR} \stackrel{\text{def}}{=} w_{NRR}(\hat{\mu}_R - \hat{\mu}_{NR}),$$

gdzie $\hat{\mu}$ oznacza nieznaną średnią w całej próbie włącznie z osobami nieprzebadanymi, $\hat{\mu}_R$ jest symbolem oznaczającym średnią ustaloną na podstawie przebadanej części próby, natomiast $\hat{\mu}_{NR}$ jest symbolem wartości średniej w nieprzebadanej części próby wylosowanej do badań (por. Groves 1989: 648; Lissowski 1971: 10).

Formalnie zatem rzecz ujmując, z równań (II.14'.) i (II.15'.) można wywnioskować, że błąd braku danych nie występuje (równa się zeru) wtedy i tylko wtedy, gdy zachodzi jeden z dwóch przypadków:

1°: w_{NRR} = 0, a to występuje jedynie wtedy, gdy n_{NR}= 0, czyli wówczas, kiedy udało się dokonać pomiaru wszystkich jednostek wybranych do badania. Sytuacja taka może być rozpatrywana jedynie w teorii statystycznej, natomiast w badaniach społecznych problem jednostek niedostępnych jest w zasadzie powszechny. Ten trywialny przypadek nie będzie zatem przedmiotem dalszych dociekań. Co ciekawe, chociaż wielkość błędu braku odpowiedzi zależy od udziału jednostek niedostępnych w próbie badawczej, to jednak – jak wskazuje R. Groves w artykule z 2006 roku, poświęconym zależności błędu braku odpowiedzi od stopnia zrealizowania próby badawczej – „nie ma minimalnej wartości wskaźnika realizacji próby, poniżej którego wartości estymatorów będą z pewnością wypaczone [przez błąd braku odpowiedzi – P.J.]" (Groves 2006: 650). Zbyt optymistycznie brzmią zatem słowa F. Sztabińskiego, iż „[m]inimalizacja błędu braku odpowiedzi jest bardzo ważna, ponieważ wysoki odsetek realizacji próby zapewnia reprezentatywność badania" (Sztabiński F. 2011: 51). Studia metodologiczne ukazują raczej, że znaczny odsetek realizacji próby badawczej nie jest gwarantem eliminacji błędu systematycznego[33] (por. Kohler 2007: 55–67; Merkle i in. 2002: 243–258; Curtin i in. 2000: 413–428; Keeter i in. 2000: 125–148), nie jest też – siłą rzeczy – dobrym predykatorem poziomu błędu systematycznego będącego skutkiem braku odpowiedzi (por. Groves i in. 2006: 721). Wystarczy przywołać ustalenia Ulricha Kohlera, który w artykule *Surveys from Inside: An Assessment of Unit Nonresponse Bias with Internal Criteria* podkreśla jednoznacznie:

> to, iż wysoki poziom odsetka realizacji próby nie koreluje z niskim poziomem błędu wypaczenia, powstałego na skutek braków odpowiedzi, nie jest zbyt wielkim zaskoczeniem. Fakt ten potwierdza jednak, że słuszne są argumenty

[33] A zarazem wskazują, że niska wartość wskaźnika realizacji próby badawczej nie oznacza koniecznie wypaczenia wyników (por. Johnston 2006: 300; Merkle i in. 2002: 243–258; Krosnick 1999: 537–567; Dillman 1991: 225–249).

> tych, którzy podają w wątpliwość zasadność traktowania odsetka realizacji próby jako wskaźnika jakości badania (Kohler 2007: 63),

na co zwracał zresztą uwagę również R. Groves, mówiąc, iż „współczynnik niezrealizowanej próby badawczej jest często błędnie uważany za miarę jakości statystyk sondażowych" (Groves 1989: 240).

2°: $\hat{p}_R - \hat{p}_{NR} = 0$ lub $\hat{\mu}_R - \hat{\mu}_{NR} = 0$, a to zachodzi wówczas, gdy estymatory wskaźników struktury wśród respondentów oraz jednostek niedostępnych są sobie równe. A zatem, wielkość błędu zależy również od różnicy pomiędzy wartością statystyki dla respondentów oraz jednostek niedostępnych (por. Jabkowski 2007: 93). Oznacza to, że „gdyby jednostki niedostępne były losowo rozrzucone w populacji, wypaczenie wyników nie byłoby w ogóle problemem nawet przy wysokich odsetkach niezrealizowania próby badawczej" (Goldberg 2003: 41). W takich przypadkach nie trzeba byłoby się w ogóle przejmować poziomem realizacji próby, bowiem wielkość ta miałaby wpływ jedynie na błąd losowy (który w sondażach jest akurat najmniejszym problemem), nie będąc w gruncie rzeczy powiązana z błędem systematycznym wypaczającym wyniki badań. Trudno jednak wskazać na jakieś poważne studia empiryczne potwierdzające na tyle zasadność tej tezy, by można było w ogóle uchylić założenie o różnicach pomiędzy respondentami oraz jednostkami niedostępnymi. Zdecydowana większość badań pokazuje, że podział próby na część dostępną i niedostępną jest silnie skorelowany z badanymi cechami wylosowanych osób, a jednostki niedostępne różnią się od osób niezbadanych zarówno jeśli chodzi o rozkłady opinii, jak i cechy społeczno-demograficzne (por. Domański 1999: 72–78; Groves 2006: 657–662; Sztabiński i in. 2007: 31–37). Dobrym tego przykładem jest podsumowanie artykułu Henryka Domańskiego, który, konkludując wyniki analiz opartych na dodatkowych wywiadach wśród jednostek niedostępnych, dochodzi do następującego wniosku:

> [jednostki niedostępne – P.J.] różniły się od respondentów, z którymi przeprowadzono wywiady – różnice te wyszły przy porównywaniu rozkładów kilkudziesięciu zmiennych [...] jednak nie znalazło to odzwierciedlenia w sile i zależnościach kształtujących [wybrane zmienne społeczno-demograficzne P.J.]. (Domański 1999: 89)

Wniosek sformułowany przez H. Domańskiego nie jest zatem jednoznaczny. Z jednej strony rozkłady odpowiedzi jednostek niedostępnych w badaniu różniły się od rozkładów odpowiedzi respondentów, z drugiej jednak strony nie miało to wpływu na formułowane wnioski w odniesieniu do różnic zaobserwowanych ze względu na cechy społeczno-demograficzne badanych osób. H. Domański wskazuje jednak na ograniczenia swoich analiz, podając, że udało mu się prze-

prowadzić wywiady ze 125 osobami z grupy 391 jednostek niedostępnych (por. Domański 1999: 90–91) – nadal zatem nie wiadomo, jakie byłyby odpowiedzi pozostałych 266 osób. Owych 125 osób niedostępnych było, używając terminologii Pawła B. Sztabińskiego, tzw. „miękkimi nie-respondentami", a więc ich pierwotna odmowa udziału w badaniu mogła być spowodowana czynnikami, które daje się dość łatwo zniwelować w procesie badawczym (por. Sztabiński P.B. 2006: 19–24). P.B. Sztabiński wskazuje równocześnie, że rozkłady odpowiedzi udzielane przez „nie-respondentów" znacznie odbiegają od odpowiedzi respondentów (co jest zgodne z pierwszą częścią wniosku H. Domańskiego). Podsumowując, można wskazać, że ignorowanie błędu nielosowego wynikającego z niezrealizowania części próby rodzi poważne błędy w oszacowaniu parametrów w populacji poprzez obniżenie poziomu reprezentatywności prób badawczych. Oznacza to, iż należy podjąć działania zmierzające do ustalenia oraz zminimalizowania wpływu tego typu błędu na całkowity błąd pomiaru.

Zauważyć można jednak, że definiowanie błędu braku odpowiedzi jako różnicy pomiędzy estymatorami w warstwie jednostek dostępnych oraz niedostępnych prowadzi do trudności w oszacowaniu wielkości tego błędu. Głównym i oczywistym ograniczeniem jest utrudniony (choć oczywiście możliwy) dostęp do osób niedostępnych w badaniu. W efekcie wielkość błędu pozostaje zawsze niewiadoma, gdyż nieznane są dokładne wartości estymatorów w warstwie jednostek niedostępnych. O ile bowiem poziom realizowalności próby jest łatwy do wyznaczenia, o tyle niewiele wiadomo o osobach, z którymi wywiadów nie udaje się zrealizować (por. Domański 1999: 67). Co prawda zaproponowano wiele procedur, które pozwalają zwiększyć pole manewru w tym zakresie, lecz nie są one wolne od różnorakich wad, w tym od dyskusyjnych i problematycznych założeń, które tkwią u ich podstaw (por. Grzeszkiewicz-Radulska 2009: 47). Kwestie te omówione będą dokładniej w rozdziale poświęconym w całości zagadnieniom jednostek niedostępnych oraz ich oddziaływaniu na poziom reprezentatywności prób badawczych.

II.1.5. Zmiana precyzji wnioskowania wynikająca z ważenia danych

Wprawdzie adjustacja danych prowadzona jest przez badaczy w ramach postsurveyowego przetwarzania wyników pomiaru, to jednak najbardziej zasadne wydaje się jej rozpatrywanie w kontekście zagadnień związanych z reprezentatywnością próby. Przemawiają za tym dwa główne argumenty. Po pierwsze, analizy literaturowe ukazują, że metodolodzy badań społecznych identyfikują potrzebę ważenia danych zazwyczaj jako odpowiedź na: (a) ko-

nieczność rekompensacji nierównych szans selekcji pewnych jednostek (lub kategorii jednostek) do próby badawczej[34], (b) niepełne pokrycie populacji operatem losowania lub (c) niepełną realizację sondażowej próby badawczej (por. Brick i in. 2009: 174–176; Biemer i in. 2003: 245; Kalton i in. 2003: 81–97). Oznacza to, oczywiście, że podstawowym celem ważenia jest redukcja błędu systematycznego powstałego na skutek pojawienia się tych źródeł błędów, które przyporządkowuje się do klasy błędów związanych z reprezentatywnością prób badawczych. Po drugie, chociaż ważenie pozwala zmniejszyć – w niektórych przypadkach – poziom błędu nielosowego[35], to jednak zazwyczaj wpływa przede wszystkim na wzrost wariancji, powodując utratę precyzji estymacji skutkującą obniżeniem poziomu reprezentatywności próby (por. Billiet i in. 2009: 12–13; Kalton i in. 2003: 83).

To przekonanie o przyroście wariancji w zbiorze ważonych wyników pomiaru bierze się z ustaleń L. Kisha, przedstawionych w 1965 roku w publikacji *Survey Sampling*, w której autor definiuje miarę mnożnika wariancji (*coefficient of variance*) (Kish 1965: 47), tj. komponent wykorzystany następnie w pracach na przykład Billieta i in. (2010: 13), Laaksonena (2007: 126), Vehovara (2007: 340–343), Kaltona i in. (2003: 83), Duncan i in. (2001: 125) do określenia wskaźnika *VIF* (z ang. *variance inflation factor*), czyli miary służącej ocenie stopnia przyrostu wariancji estymatorów w próbie ważonej (relatywnie do poziomu wariancji w próbie nieważonej). Pamiętać należy jednak, iż współczynnik *VIF* w rozumieniu zaprezentowanym w przywołanych opracowaniach jest miernikiem odnoszącym się do wszystkich zmiennych w ogóle[36], a zatem pozwala określić jedynie poziom maksymalnego przyrostu wariancji. Innymi słowy, opisuje on przypadek najmniej korzystny, który może, chociaż wcale nie musi wystąpić. Co więcej, dla pewnych estymatorów, ważenie danych może

---

[34] Przyporządkowywanie wag wyrównujących nierówne prawdopodobieństwa selekcji jednostek populacji do prób badawczych, będące efektem zastosowania określonych schematów doboru prób badawczych, rozpatrywane jest zazwyczaj w kontekście efektu schematu doboru próby. Uwzględnia się je zatem w omówionym wcześniej wskaźniku DEFF.

[35] Oczywiste jest jednak, że procedura ważenia danych pozwala ograniczyć błąd systematyczny pomiaru jakiejś zmiennej, powstały na skutek błędów operatu losowania lub błędów braku odpowiedzi, tylko wówczas, gdy rozkłady wartości takiej zmiennej w obrębie jednostek niedostępnych będą zbliżone do tych rozkładów, które ustalono na podstawie przebadanej części próby (por. Biemer i in. 2003: 246). W literaturze metodologicznej wskazuje się przy tym na trzy główne typy relacji wiążące cechy jednostek niedostępnych z wartościami zmiennych będących przedmiotem pomiaru (por. Akacha i in. 2011: 1073; Pokropek 2011: 82-86; Wood i in. 2004: 525-526; Little i in. 2002: 13-17). Wprowadza się przy tym rozróżnienie wywodzące się od Donalda B. Rubina (1976) na: (a) braki całkowicie losowe (*missing completely at random*), (b) braki losowe (*missing at random*) oraz (c) braki nielosowe (*missing not at random*).

[36] Dla pojedynczych zmiennych precyzję estymatorów ważonych, ujętą relatywnie do precyzji estymatorów nieważonych, można wyznaczyć z danych wynikowych jako kwadrat ilorazu błędu standardowego estymatora ważonego oraz nieważonego.

nawet zwiększyć, a nie zmniejszyć, tak jak by wynikało ze wskaźnika *VIF*, precyzję wnioskowania statystycznego[37] (por. Little i in. 2005: 162).

Mimo wszystko miernik *VIF* jest niezwykle przydatny w praktyce badawczej, zwłaszcza jeśli chce się wyznaczyć efekt losowania jednostek z nierównymi prawdopodobieństwami doboru. Dla estymatora wskaźnika struktury (maksymalną) wariancję wyników w próbie ważonej można zapisać jako:

(II.16.) $$Var(\hat{p}_w) = Var(\hat{p}) \times VIF,$$

natomiast dla estymatora średniej arytmetycznej:

(II.17.) $$Var(\bar{y}_w) = Var(\bar{y}) \times VIF,$$

przy czym *VIF* (w wersji konserwatywnej) wyraża formuła (por. Billiet i in. 2009: 13; Park i in. 2004: 187):

(II.18.) $$VIF = 1 + \frac{s_w^2}{\bar{w}^2},$$ gdzie:

$\hat{p}_w$ oraz $\bar{y}_w$ oznaczają wartości estymatorów wskaźnika struktury oraz średniej arytmetycznej w próbie ważonej, $\hat{p}$ oraz $\bar{y}$ odpowiednie wartości tych estymatorów w próbie nieważonej, s_w^2 wariancję wartości wag przypisanych kolejnym jednostkom w próbie badawczej, natomiast $\bar{w}^2$ oznacza kwadrat średniej arytmetycznej z wag przyporządkowanych kolejnym elementom próby. Ów przyrost wariancji, na który wskazuje *VIF* jest więc konsekwencją dodatkowego zróżnicowania w zbiorze wyników pomiaru wynikającego z wariancji wag. Zauważyć można, że po przekształceniach wzorów (II.16.) oraz (II.17.) wskaźnik *VIF* jest ilorazem wariancji wyników ważonych oraz nieważonych, przyjmuje zatem postać analogiczną do zdefiniowanej wcześniej miary DEFF (por. wzór nr (II.7.)), ukazującej skalę utraty precyzji estymacji w skutek zastosowania schematu doboru próby badawczej odbiegającego od prostego doboru losowego. Jeżeli zatem wariancja w zbiorze wartości zmiennych jest niezerowa, to dla dowolnego estymatora $\hat{\theta}$ wartości parametru θ wskaźnik *VIF* można określić wzorem:

(II.18'.) $$VIF \stackrel{\text{def}}{=} \frac{Var(\hat{\theta}_w)}{Var_{SAMPL}(\hat{\theta})}.$$

Miernik *VIF* nie jest oczywiście jedynym możliwym sposobem ujmowania relacji zachodzących pomiędzy ważeniem danych, błędem systematycznym oraz wariancją estymatorów. Dla przykładu, Sergey Dorofeev i Peter Grant definiują tzw. miernik efektu ważenia (*weighting effect*), którego wartość wyzna-

[37] Dla pewnych specyficznych zmiennych ważenie danych może oczywiście poprawić precyzję wnioskowania statystycznego. Taka sytuacja nastąpi na przykład wtedy, gdy wagi będą skorelowane z wartościami zmiennych wynikowych (por. Little i in. 2005: 161–162; Groves 1989: 258).

czana jest jako iloraz liczebności próby zrealizowanej oraz tzw. liczebności skalibrowanej[38] (por. Dorofeev i in. 2006: 105). Owa liczebność skalibrowana jest niczym innym jak przeskalowaną wielkością próby zrealizowanej do realnej efektywnej wielkości próby ważonej[39]. Chociaż S. Dorofeev i P. Grant określają miarę efektu ważenia nieco inaczej, niż definiuje się wskaźnik *VIF*, to jednak interpretacja wartości obu wskaźników pozostaje taka sama, tzn. główną konkluzją jest opinia o utracie precyzji estymacji pod wpływem ważenia danych (por. Dorofeev i in. 2006: 106).

Przegląd literatury ukazał również, że choć znaczna część metodologów zajmujących się problematyką błędów badań sondażowych wyraża przekonanie o wzroście wariancji jako wyraźnym efekcie ubocznym ważenia wyników pomiaru, to jednocześnie zdecydowana większość z nich nie mówi w ogóle o skali oddziaływania takich przekształceń na precyzję estymacji. W pracy R. Grovesa (1989) odnaleźć można tylko fragmenty wspominające o tym problemie. Dla przykładu, w szóstym rozdziale książki *Survey Errors and Survey Costs* jej autor podkreśla, że „próba ważona cechuje się zazwyczaj wyższą wariancją od próby nieważonej, z wyjątkiem sytuacji, w której poszczególne jednostki są dobierane tak, że zachodzi korelacja pomiędzy wagami a wariancją indywidualnych obserwacji" (Groves 1989: 258)[40]. Podobnie czyni P. Biemer, wskazując – co prawda – na pewne trywialne źródła uchybień wynikające z ważenia danych, na przykład na pomyłki obliczeniowe (por. Biemer 2010a: 35, Biemier 2010b: 825), ale nie wspomina już o obniżeniu precyzji pomiaru. Wprawdzie w monografii Biemera i in. (2003: 347) odnaleźć można wzmiankę o wzroście wariancji na skutek adjustacji danych, lecz poza krótką notą, problematyka ta nie jest już dalej podejmowana. W zamian tego autorzy wskazują, że „w przypadku złożonego charakteru wag istnieje ryzyko, że będą one wyznaczone nieprawidłowo i mogą […] spowodować wzrost [całkowitego błędu pomiaru – P.J.]" (Biemer i in. 2003: 246). Należy bowiem pamiętać, że w wielu przypadkach wagi same w sobie są estymatorami (por. Rässler i in. 2008: 375) i podlegają tym samym ograniczeniom, na jakie natrafia się w pomiarze sonda-

---

[38] Liczebność skalibrowana wyznaczana jest na podstawie wartości wag ($w_1, w_2, \ldots, w_n$) przypisanych kolejnym respondentom w n-elementowej próbie badawczej za pomocą następującego wzoru: $n_c = \frac{(\sum_{i=1}^{n} w_i)^2}{\sum_{i=1}^{n} w_i^2}$ (por. Dorofeev i in. 2006: 105).

[39] Efektywność mierzona jest utratą precyzji estymacji, czyli wzrostem wariancji w próbie ważonej. Jednoznacznie świadczą o tym zaprezentowane przez autorów książki cechy miary efektu ważenia, a dokładniej pierwsza z siedmiu omawianych właściwości: „[w]ielkość próby skalibrowanej nie może być większa od wielkości próby nieważonej (i stąd wartość miary efektu ważenia, nie może być nigdy mniejsza od 1). Są one sobie równe wtedy i tylko wtedy, gdy wagi mają stałe wartości" (por. Dorofeev i in. 2006: 107).

[40] Ta szczególna sytuacja, o której wspomina R. Groves, wynika z zastosowania specyficznego schematu doboru próby, który może być bardziej efektywny od prostej próby losowej (na przykład w losowaniu warstwowym proporcjonalnym).

żowym. Jeżeli zatem procedury ważenia danych oparte zostaną na zmiennych obarczonych błędami nielosowymi, to w konsekwencji utraci się nie tylko precyzję, ale także w żaden sposób nie zredukuje się błędu systematycznego (por. Billiet i in. 2010: 21).

Przeglądając literaturę metodologiczną, odnaleźć można także takie studia empiryczne, których autorzy wskazują, iż przyrost wariancji nie jest koniecznym następstwem ważenia danych (por. Liao i in. 2012: 53–62; Cervantes i in. 2009: 4642–4655; Pike 2008: 153–171; Little i in. 2005: 161–168; Little 1986: 139–157; Kalton i in. 1986: 1–16; Oh i in. 1983: 143–184; Holt i in. 1979: 22–46). Jedną z ciekawszych propozycji w tym zakresie zaprezentowali autorzy artykułu *Does Weighting for Nonresponse Increase the Variance of Survey Means?* (por. Little i in. 2005: 161–168). Badacze ci ukazują, że wskaźnik *VIF* jest właściwym oszacowaniem przyrostu wariancji tylko i wyłącznie w sytuacji słabego skorelowania wag ze zmiennymi poddawanymi pomiarowi. Okazuje się bowiem, że przy pewnych warunkach brzegowych oraz specyficznym sposobie wyznaczania wartości wag próba ważona może cechować się nie tylko niższym poziomem błędu systematycznego wynikającego z niepełnej realizacji próby (będzie tak, gdy wartości wag będą jednocześnie skorelowane z prawdopodobieństwem udzielenia odpowiedzi oraz z wartościami badanej zmiennej), ale także charakteryzować się może niższym poziomem wariancji wyników (im skorelowanie wag z wartościami zmiennej będzie większe, tym większe będą korzyści z ważenia danych dla precyzji estymacji).

Tabela II.2. Wpływ ważenia danych na losowy oraz systematyczny komponent całkowitego błędu pomiaru

Związek wag z niedostępnością	**Związek wag z analizowaną zmienną**	
	Niski	**Wysoki**
Niski	Błąd syst.: --- Wariancja: ---	Błąd syst.: --- Wariancja: spadek
Wysoki	Błąd syst.: --- Wariancja: przyrost	Błąd syst.: spadek Wariancja: spadek

Źródło: Little i in. (2005: 164)

W przywołanej pracy odnaleźć można również alternatywny sposób mierzenia efektywności procedur ważenia danych (por. wzory (11), (12), (13) oraz (14) w: Little i in. 2005: 164), którego idea polega na porównaniu błędu średniokwadratowego w próbie ważonej oraz nieważonej (w ocenie efektywności uwzględnia się więc zarówno losowy, jak i też systematyczny komponent *TSE*). Autorzy ukazują, że ważenie danych jest najbardziej efektywne w redukcji błę-

du systematycznego powstałego na skutek niepełnej realizacji próby, w odniesieniu do zmiennych pozostających w silnej korelacji z mechanizmem niedostępności oraz z wartościami wag. Oznacza to, że wagi prowadzące do efektywnej redukcji błędu systematycznego będą skutkować poprawą, a nie spadkiem precyzji wnioskowania, na co według założeń wskazywałby wskaźnik *VIF*. Jeżeli zatem ważenie danych stosuje się w celu zminimalizowania wielkości błędu związanego z brakiem obserwacji, to ma ono sens tylko i wyłącznie w odniesieniu do tych zmiennych, które pozostają w silnej korelacji z wartościami wag. Nawet jeśli redukcja błędu systematycznego nie jest w takiej sytuacji pewna (nie zawsze wiadomo, czy wagi są powiązane z procesem niedostępności jednostek), to umożliwia poprawę precyzji estymacji. W przeciwnym przypadku ważenie powoduje wzrost wariancji bez żadnej rekompensaty w postaci ograniczenia wielkości błędu systematycznego.

II.2. Błędy związane z pomiarem

II.2.1. Etap konceptualizacji oraz operacjonalizacji problematyki badawczej

Pomimo że konceptualizacja oraz operacjonalizacja problematyki badawczej (w tym zagadnienia doboru wskaźników) pozostają jednym z najważniejszych etapów każdego projektu badawczego, to jednak refleksja nad jakością formułowanych pytań kwestionariuszowych, a w zasadzie nad jakością dobieranych wskaźników, podejmowana jest najczęściej w ramach analizy błędów pomiarowych, wśród których błąd doboru wskaźnika stanowi tylko jedno z wielu potencjalnych źródeł błędów (por. na przykład Weisberg 2005: 92–129; Groves i in. 2004: 201–237). Nie pochylając się w ogóle nad zasadnością takiego podejścia (dobór wskaźników jest jednak czymś istotnie odmiennym od przełożenia pytań badawczych na pytania kwestionariuszowe), można zauważyć, że metodologowie rozpatrujący problemy operacjonalizacji w ramach oceny jakości pomiaru, koncentrują swoją uwagę bardziej wokół kwestii takich, jak zależność otrzymywanych odpowiedzi od: (a) sposobów sformułowania pytań kwestionariuszowych, (b) przedstawianych respondentom zestawów kategorii odpowiedzi, czy też (c) statycznej oraz dynamicznej struktury kwestionariusza (por. Sztabiński F. 2011: 54–58), niż na analizach poświęconych jakości wskaźników. Choć oczywiście problemy te są niezwykle ważne, a w paradygmacie całkowitego błędu pomiaru zupełnie fundamentalne, to jednak trzeba stwierdzić, że na etapie opracowywania pytań kwestionariuszowych chodzi w mniejszym stopniu o metodologiczne dywagacje nad tym, czy zastosowanie dwóch alternatywnych zestawów pytań prowadzi do odmiennych odpowiedzi respon-

dentów, ale bardziej o to, które z pytań precyzyjniej mierzy to, co powinno mierzyć, lub czy w ogóle mierzy to, co interesuje badacza. Chociaż tematyka ta jest doskonale rozpoznana w literaturze metodologicznej (zwłaszcza w badaniach konstruktów psychometrycznych), to jednak w paradygmacie całkowitego błędu pomiaru badań sondażowych nie doczekała się systematycznych implementacji. Co prawda zagadnienia te podejmowali P. Biemer i L. Lyberg, definiując specjalną kategorię błędu nazwaną *błędem specyfikacji* (por. Biemer i in. 2003: 28–29), a także R. Groves, który w klasycznym dziele *Survey Errors and Survey Costs* odwoływał się do psychometrycznych miar trafności oraz rzetelności pomiaru (por. Groves 1989: 18–47), jednak na ogół błędy popełniane przy okazji doboru wskaźników analizowane są w kontekście błędów narzędzi badawczych.

W tej sekcji rozdziału uwaga skoncentrowana będzie na dwóch sposobach identyfikacji ułomności procesu doboru wskaźników. Wykazane zostanie, że o wiele bardziej użytecznym rozwiązaniem (choć nie pozbawionym wad, na przykład nie dającym się przełożyć na terminologię błędów w paradygmacie *TSE*) jest rozpatrywanie jakości sformułowanych pytań w ramach oceny ich trafności oraz rzetelności, a nie próby posługiwania się nieprecyzyjnie sformułowanym pojęciem błędu specyfikacji. Wynika to zapewne z faktu, iż komplikacje wynikające z doboru wskaźników, a także ich wpływ na jakość badań, dotyczą zdecydowanie bardziej zjawisk nieobserwowalnych (to znaczy opinii, przekonań, odczuć itd.), dla których poszukuje się zastępczych miar pozwalających badaczom na ich zidentyfikowanie, mniej natomiast dotyczą pomiaru tych zjawisk, których sens empiryczny jest jednoznaczny, a koncepcja *TSE* ma proste przełożenie na systematyczne oraz losowe odstępstwa wartości estymatorów od rzeczywistych wartości szacowanych parametrów[41].

Rozpocząć warto od prezentacji błędu specyfikacji. W kontekście tego błędu niezwykle ważne jest to, że pojawia się on (jako jedno ze źródeł całkowitego błędu pomiaru) tylko i wyłącznie w koncepcji P. Biemera oraz L. Lyberga, a dokładnie w drugim rozdziale ich wspólnej publikacji *Introduction to Survey Quality* (2003: 28–29, 38–40). Pierwsza wzmianka o tej kategorii błędu jest zamieszczona w ramach charakterystyki początkowej fazy procesu badawczego związanej z formułowaniem pytań badawczych. Autorzy stwierdzają bowiem, że „błąd specyfikacji występuje, gdy pytania kwestionariuszowe nie uwzględ-

[41] Prowadzone tu rozważania pozostają zatem rozwinięciem – rozpatrywanych już w rozdziale I – kontrowersji wynikających z definiowania całkowitego błędu pomiaru w odniesieniu do konstruktów dla których wartość prawdziwa pozostaje co do zasady umowna i stanowi co najwyżej wartość ukrytą. Przywołane w rozdziale pierwszym konstatacje R. Grovesa (1989: 9) oraz F. Sztabińskiego (2011: 48) – którzy uznawali za zasadne włączanie w obręb analiz jakości pomiaru również psychometryczne miary jakości badań – odnoszą się właśnie do sytuacji związanych z analizą trafności oraz rzetelności wskaźników i skal pomiarowych.

niają wszystkiego tego, co jest niezbędne do udzielenia odpowiedzi na pytania badawcze" (Biemer i in. 2003: 28–29). W przywołanej monografii nie jest to jednak dominujący sposób rozumienia błędu specyfikacji. Dopiero we fragmencie opisującym pięć głównych źródeł błędów systematycznych, które w badaniach sondażowych wpływają na dokładność pomiaru, autorzy przedstawiają właściwą definicję błędu specyfikacji jako takiego, „który pojawia się wtedy, gdy mierzony pytaniem kwestionariuszowym konstrukt różni się od tego, który powinien być zmierzony" (Biemer i in. 2003: 38). W późniejszym swoim tekście P. Biemer (2010a) doprecyzowuje tę definicję poprzez uwypuklenie skutków błędu specyfikacji:

> kiedy to nastąpi [tj. pojawi się błąd specyfikacji – P.J.] wówczas w konsekwencji zły parametr zostanie wyestymowany, co może prowadzić do nieprawidłowych wniosków. (Biemer 2010a: 822)

Struktura narracji Biemera i Lyberga (2003), a także samego Biemera (2010a, 2010b) jest taka, że po krótkiej charakterystyce błędu specyfikacji przywoływane są przykłady obrazujące potencjalne źródło zniekształceń wyników na skutek tego błędu. Egzemplifikacje błędów są w gruncie rzeczy proste i bardzo do siebie podobne, pozostają jednak istotne o tyle, że ukazują, w jaki sposób i do czego autorzy odnoszą ten typ błędu. Z przykładów tych można wywnioskować, że definiowany przez nich błąd specyfikacji obejmuje sytuację, w której badacz próbuje oszacować prawdziwe wartości pewnego parametru, lecz zapomni, lub po prostu nie uwzględni, jakiegoś komponentu (na przykład składnika miary), mającego istotne znaczenie dla ustalenia owej prawdziwej wartości. Dla przykładu, gdyby celem badania było oszacowanie dochodu respondentów, którzy proszeni byliby o wskazanie właściwej kwoty w odniesieniu do różnych źródeł pozyskiwania dochodu, a badacz zapomniałby na przykład zapytać o dochody osiągane z umów o dzieło, to określona w ten sposób miara nie doszacowałaby wielkości parametru populacyjnego. W takim ujęciu błąd specyfikacji nie jest więc tym, czym psychometryczna trafność, która w metodologii – nie tylko badań psychologicznych (por. Brzezińska A. i in. 2011: 385–386; Hornowska 2007: 80–100; Brzeziński i in. 1984), ale także tej poświęconej błędom badań sondażowych (por. Groves 1989: 18–47; Sztabiński F. 2011: 62–71) – odnoszona jest do przypadku pomiaru zapośredniczonego. Pomimo tych wyraźnych rozbieżności w sposobie definiowania obu pojęć, P. Biemer przyrównuje błąd specyfikacji do trafności wskaźników:

> błąd specyfikacji jest powiązany z koncepcją trafności wskaźników, tzn. możliwością pomiaru badanego konstruktu przy pomocy pytań kwestionariuszowych. Pomimo tego, że nietrafność jest spowodowana także poprzez błędy pomiaru, to błąd specyfikacji jest z definicji różny od błędu pomiaru. Błąd

> specyfikacji odnosi się do pomiaru złego konstruktu, a nie słabego pomiaru dobrego konstruktu. (Biemer 2010b: 31)

W sumie jednak ani krótkiej definicji opisowej, ani też prostym przykładom nie towarzyszy żadna poważniejsza refleksja nad źródłem tego błędu oraz sposobami jego ograniczania, może z wyjątkiem bardzo ogólnego stwierdzenia, że „wykrycie błędu specyfikacji wymaga zazwyczaj prześledzenia każdego pytania w kwestionariuszu wywiadu przez badaczy, którzy powinni określić, czy pytania mierzą to, co miało być zmierzone" (Biemer i in. 2003: 40).

Niezwykle symptomatyczne jest to, że w literaturze metodologicznej odnaleźć można już tylko nieliczne prace, których autorzy odwołują się do zdefiniowanego przez Biemera i Lyberga pojęcia błędu specyfikacji. Jeśliby wyłączyć autocytowanie Biemera w jego dwóch niemal identycznych publikacjach z 2010 roku (2010a: 822, 2010b: 31–32), powielających w znacznej części fragmenty wspólnej książki z Lybergiem (2003), to określenie „błąd specyfikacji" wykorzystują jeszcze Frauke Kreuter, Gerrit Muller oraz Mark Trappmann (2010: 821), a także M. Fuchs (2008: 898), który, zamieszczając definicję tego błędu, zwraca uwagę na fakt, że:

> jeżeli jakiś istotny aspekt badanego konstruktu został pominięty […] wówczas trafność operacjonalizacji tego konstruktu byłaby zagrożona […], a błąd specyfikacji mógłby się pojawić. Mogłoby to pociągnąć za sobą poważne wypaczenie danych, gdyż wartości estymatorów […] nie odpowiadałyby rzeczywistym [wartościom parametrów – P.J.] w badanej populacji. (Fuchs 2008: 898)

Definicja M. Fuchsa pozostaje nie tyle zbieżna z propozycją Biemera i Lyberga, co w sposób oczywisty wyrasta na jej bazie, czego jednoznacznym potwierdzeniem jest zamieszczony spis literatury (por. Fuchs 2008: 902) zawierający *explicite* odwołanie do monografii *Introduction to Survey Quality*.

Choć celem tej publikacji nie jest dochodzenie przyczyn tego, że pojęcie błędu specyfikacji – w formie zaproponowanej przez Biemera i Lyberga – nie spotkało się z szerszym zainteresowaniem metodologów badań sondażowych, to jednak wydaje się, że dwie przyczyny miały znaczenie decydujące. Po pierwsze, pomimo określenia tego źródła błędu w sposób opisowy, P. Biemer i L. Lyberg nie przedstawiają już definicji formalnej. Innymi słowy, stwierdzają, że błąd specyfikacji wypacza wyniki, ale nie pokazują już, w jaki sposób kierunek tego wypaczenia wyznaczyć. Swoją drogą, takie działanie byłoby chyba bezcelowe. Po drugie, to niewielkie zainteresowanie metodologów błędem specyfikacji w formie zaproponowanej przez obu autorów wynikać może też z faktu (na który zresztą zwrócili uwagę oni sami), że termin wykorzystany przez nich do nazwania błędu doboru wskaźników ma swoje silne zakorzenienie w literaturze ekonometrycznej, gdzie oznacza zupełnie co innego, bowiem od-

nosi się do niewłaściwego doboru zmiennych w modelach regresyjnych czy też eksperymentalnych, a dokładniej oznacza włączenie do analiz zmiennych nieistotnych dla budowy modelu (niepowiązanych substantywnie ze zmienną zależną) lub pominięcie takich zmiennych, których obecność byłaby znacząca dla opisu i wyjaśnienia badanych zjawisk (por. Biemer i in. 2003: 40).

Chociaż Biemer i Lyberg podkreślają jednoznacznie, że w ich rozumieniu błąd specyfikacji nie odnosi się w ogóle do modeli analitycznych, ale do pytań w kwestionariuszu, to jednak ekonometryczne rozumienie tego typu błędu okazało się dominujące. Zresztą nie tylko w pozycjach ekonometrycznych, ale również tych z zakresu badań sondażowych, odnaleźć można wyraźne odniesienia do błędu specyfikacji jako do efektu niewłaściwego doboru zmiennych w analizach statystycznych. Świetnym przykładem takiego podejścia są rozważania H. Weisberga (2005), w ramach opisu błędów pojawiających się już po fazie pomiaru autor ten używa bowiem określenia „błąd specyfikacji" na opis sytuacji „pominięcia istotnych zmiennych niezależnych w analizie ich oddziaływania na zmienną zależną, skutkującego stronniczością szacunków kierunku wpływu predykatorów" (Weisberg 2005: 274–275). W takim samym znaczeniu pojęcie to pojawia się również w opublikowanym w 2004 roku artykule Ulfa H. Olssona, Trona Fossa oraz Einara Brevika (2004: 453–500), zamieszczonym w znanym metodologicznym czasopiśmie „Sociological Methods & Research". Innymi słowy, pojęcie błędu specyfikacji utrwaliło się w literaturze raczej w rozumieniu ekonometrycznym, a nie tym zaproponowanym przez P. Biemera i L. Lyberga.

Dla oceny jakości narzędzi pomiarowych niezwykle użyteczne wydaje się natomiast posługiwanie miarami trafności oraz rzetelności wskaźników. Wprawdzie nie wnoszą one wiele w oszacowanie wielkości całkowitego błędu pomiaru badań sondażowych, gdyż nie mają bezpośredniego przełożenia na koncepcję błędu, to jednak stosuje się je w ocenie pomiaru specyficznych konstruktów teoretycznych, w odniesieniu do których estymacja wielkości błędu pomiarowego jest jeszcze bardziej kłopotliwa, niż ma to miejsce w badaniu cech obserwowalnych. Innymi słowy, chociaż mierniki trafności oraz rzetelności pozostają w znacznym stopniu niezależne od kryteriów oceny danych przyjmowanych w świetle koncepcji *TSE* (por. Sztabiński F. 2011: 61–76; Groves 1989: 18–47), to jednak nic nie stoi na przeszkodzie, aby tam, gdzie jest to zasadne, oprzeć się na dodatkowych standardach ewaluacji badań, co może – jeśli nawet nie bezpośrednio, to przynajmniej pośrednio – wpłynąć na ograniczenie wielkości całkowitego błędu pomiaru.

Charakteryzując trafność wskaźników, Robert Groves (1989) odwołuje się do znanego artykułu Georga W. Bohrnstedta (1983: 70–122) pod tytułem *Measurement*. W opracowaniu tym przedstawiono rozróżnienie na *trafność teoretyczną* oraz *trafność empiryczną*. To pierwsze pojęcie odnosi się do „kore-

lacji pomiędzy prawdziwym wynikiem oraz odpowiedzią respondenta uzyskaną we wszystkich różnych powtórzeniach pomiaru" (Groves 1989: 22), drugie natomiast – do sposobu empirycznej oceny trafności miar pewnych konstruktów poprzez ich zwielokrotniony wielowskaźnikowy pomiar. W sensie formalnym trafność teoretyczną pewnej miary Y można zapisać za pomocą formuły (por. Groves 1989: 42; Groves i in. 2004: 254):

(II.19.) $$\text{Trafność}(Y) \stackrel{\text{def}}{=} \frac{\sum_{t,i}(Y_{ti}-\bar{Y})(\mu_i-\bar{\mu})}{\sqrt{\sum_{t,i}(Y_{ti}-\bar{Y})^2 \sum_{t,i}(\mu_i-\bar{\mu})^2}},$$

gdzie:

- $t \in \{1, 2, \ldots, T\}$ oznacza kolejne replikacje pomiaru,
- $i \in \{1, 2, \ldots, N_{pomiar}\}$ jest oznaczeniem kolejnych badanych osób,
- Y_{ti} oznacza wartość miary *Y* dla *i*-tej osoby w kolejnej replikacji pomiaru,
- $\bar{Y}$ jest średnią z wartości Y_{ti} dla wszystkich *t* oraz *i*,
- μ_i jest wartością prawdziwą dla *i*-tej osoby,
- $\bar{\mu}$ jest średnią z wartości μ_i.

Oznacza to, że co do istoty miara trafności jest współczynnikiem korelacji pomiędzy prawdziwą wartością pomiaru a tą, którą zaobserwowano za pomocą wskaźnika *Y*.

Określony w ten sposób miernik trafności ma jednak wyraźne niedostatki, z których bodaj najpoważniejszym jest to, że możliwość wyznaczania trafności wskaźników ograniczona jest do przypadków pomiaru tych konstruktów, dla których badacz jest w stanie przyrównać otrzymywane wyniki pomiaru do tak zwanych „złotych standardów", czyli zewnętrznych danych o znanych prawdziwych wartościach. Wartości takie są oczywiście niedostępne dla pomiaru opinii, odczuć, przekonań itd., czyli wszędzie tam, gdzie stosuje się pomiar pośredni. Innymi słowy, trafność teoretyczna może zostać jedynie oszacowana, a nie dokładnie wyznaczona. Jedną z takich procedur estymacji jest wielokrotny pomiar tego samego konstruktu za pomocą różnych wskaźników, co prowadzi do wspomnianego już pojęcia trafności empirycznej. Zgodnie z propozycją G.W. Bohrnstedta (1983), R. Groves definiuje trafność empiryczną jako

> współczynnik korelacji pomiędzy miarą [wskaźnikiem – P.J.] oraz inną obserwowalną zmienną, która jest postrzegana jako wskaźnik tego samego konstruktu, stąd trafność empiryczna pewnego indikatora jest zawsze określana w relacji do innej miary. (Groves 1989: 23)

Następnie omawia dwie metody wyznaczania trafności empirycznej, to znaczy tak zwaną trafność prognostyczną (oryg. *predictive validity*) oraz trafność równoczesną (oryg. *concurent validity*). Tę pierwszą oblicza się jako współczynnik korelacji pomiędzy pomiarem konstruktu za pomocą określonego wskaźnika

oraz przeprowadzonym w późniejszym czasie pomiarem tego samego konstruktu za pomocą innej miary, stanowiącej punkt odniesienia dla oceny trafności wybranego wskaźnika. Z kolei druga metoda opiera się na wyliczeniu współczynnika korelacji pomiędzy pomiarem konstruktu za pomocą wybranego wskaźnika oraz przeprowadzonym w tym samym czasie pomiarem innego wskaźnika, stanowiącego punkt odniesienia (wzorzec) w ocenie trafności. W obu podejściach spełnione musi być założenie, że porównywane wskaźniki, tj. oceniany oraz wzorcowy, mierzą w istocie ten sam konstrukt (por. Groves 1989: 23). Podejście to jest niezwykle rozpowszechnione w analizie trafności wskaźników, czego przykładem są publikacje Franciszka Sztabińskiego (2011: 67–68), Duane F. Alwina (2007: 23), czy też Chawy Frankfort-Nachmias i in. (2001: 181–182).

W publikacjach tych odnaleźć można także inną ciekawą metodę badania trafności wskaźników, opracowaną przez Donalda T. Campbella oraz Donalda W. Fiskego w 1959 roku. Opiera się ona na wielowskaźnikowym pomiarze tego samego konstruktu poprzez dwa kryteria oceny, tzw. trafność zbieżną (oryg. *convergent validity*) oraz trafność różnicującą (oryg. *discriminant validity*). Procedura ta znana jest w literaturze metodologicznej także pod inną nazwą, tak zwanej *analizy wielu cech, wielu metod* (por. Sztabiński F. 2011: 69–71; Frankfort-Nachmias i in. 2001: 184), a u jej podstawy leży założenie, że różne wskaźniki tego samego konstruktu powinny dawać podobne rezultaty pomiaru, z kolei wskaźniki różnych konstruktów – prowadzić do wyników odmiennych. W sensie obliczeniowym trafność zbieżna wyznaczana jest jako współczynnik korelacji pomiędzy wskaźnikami tego samego konstruktu, a trafność dyskryminacyjna – jako współczynnik korelacji pomiędzy wskaźnikami różnych konstruktów. W metodzie tej ważne jest to, aby przestudiować oba kryteria trafności jednocześnie. Pomiar będzie można uznać za trafny, jeżeli różne pomiary tej samej cechy będą ze sobą skorelowane nawet wtedy, gdy pomiaru dokonywano odmiennymi metodami a także, jeżeli różne pomiary odmiennych właściwości nie będą ze sobą skorelowane w sposób istotny, nawet wtedy, gdy w pomiarach wykorzystano tę samą metodę zbierania danych. Innymi słowy, w przypadku trafności zbieżnej współczynniki korelacji powinny być wysokie/dodatnie, a w przypadku trafności dyskryminacyjnej – niskie/ujemne (por. Campbell i in. 1959: 81–105).

Omówione powyżej procedury nie wyczerpują oczywiście wszystkich opisanych w literaturze przedmiotu metod wyznaczania trafności wskaźników[42], jednak w świetle prowadzonych tutaj rozważań nie jest istotne to, by je wszystkie wymienić i opisać, ale by zauważyć, że zdecydowana większość z tych propozycji, a przynajmniej te najbardziej znane, opierają się na idei wyliczenia

[42] W przywoływanych pracach odnaleźć można przynajmniej kilkanaście alternatywnych sposobów definiowania empirycznych miar trafności wskaźników.

miernika korelacji pomiędzy wynikami pomiaru uzyskanymi dzięki zastosowaniu określonego wskaźnika (lub wskaźników) a jakąś inną miarą (lub miarami) stanowiącymi punkt odniesienia. To właśnie w tym sposobie definiowania miar trafności empirycznej ujawnia się zasadnicza różnica pomiędzy psychometryczną oraz statystyczną koncepcją oceny jakości pomiaru, co sprawia, że miar tych nie da się przełożyć na terminologię błędów w koncepcji *TSE*.

Powiedziano już wiele o psychometrycznym kryterium trafności, niewiele natomiast o drugim mierniku oceny jakości wskaźników, czyli rzetelności przeprowadzonego pomiaru. Ogólnie rzecz biorąc, trafność pomiaru informuje badacza o tym, czy narzędzie badawcze mierzy to, co ma rzeczywiście mierzyć, rzetelność z kolei o tym, z jaką dokładnością narzędzie pomierzyło to, co w istocie zmierzyło. Ta gra słów wprowadzona przez F. Sztabińskiego (2011: 62, 71) świetnie oddaje istotę i różnicę obu pojęć. Trafność odnosi się bowiem do zamierzeń badacza i określa „zdolność" narzędzia badawczego do pomiaru konstruktów, które pierwotnie miały być zmierzone (możliwość uzyskiwania wyników, które są adekwatnym odbiciem mierzonego pojęcia), natomiast rzetelność określa empiryczną precyzję pomiaru (otrzymywanie spójnych wyników za pomocą tego samego narzędzia), niezależnie oczywiście od tego, co pomierzono lub inaczej, czy w ogóle zmierzono to, co było oryginalnym zamierzeniem badacza (por. Sztabiński F. 2011: 75). Innymi słowy, rzetelność jest miarą zróżnicowania wartości wskaźników otrzymanych po przebadaniu tych samych osób w różnych replikacjach pomiaru i w pewnym sensie ukazuje, w jakim zakresie pomiar jest wolny od błędu losowego.

Taki sposób rozumienia rzetelności zakłada, że respondenci są stali w swoich odpowiedziach (nie zmienia się wartość indicatum), a ewentualne różnice w odnotowywanych wartościach są efektem nierzetelności narzędzia[43]. Robert Groves (1989: 22), odwołując się do przywołanej już pracy G. W. Bohrnstedta (1983), definiuje rzetelność jako stosunek wariancji wartości prawdziwych do wariancji obserwowanych wyników. Ponieważ wariancję wyniku prawdziwego σ_{μ}^2 można zapisać jako różnicę wariancji wyniku obserwowanego σ_Y^2 oraz wariancji błędu σ_{ε}^2 (por. Frankfort-Nachmias i in. 2001: 186), to miara rzetelności przyjmuje postać następującego ilorazu:

(II.20.) $$\text{Rzetelność}(Y) \stackrel{\text{def}}{=} \frac{\sigma_{\mu}^2}{\sigma_Y^2} = \frac{\sigma_Y^2 - \sigma_{\varepsilon}^2}{\sigma_Y^2}.$$

[43] Rzetelność pomiaru niezwykle łatwo zobrazować przykładem pomiaru wagi ciała lub wzrostu. Rzetelne narzędzie pomiarowe powinno dawać w każdej replikacji pomiaru tej samej jednostki tę samą wartość, przy czym ewentualne rozbieżności będą świadczyć o nierzetelności narzędzia. Oczywiście w pomiarze cech fizykalnych sprawa nie jest wielce skomplikowana (najczęściej każdy pomiar jest niezależny od siebie, to znaczy pierwszy nie wpływa na drugi itd.), ale wtedy, gdy przedmiotem pomiaru są opinie, przekonania itp., spełnienie warunku niezależności kolejnych replikacji pomiaru staje się poważnym wyzwaniem dla badacza.

W pewnym sensie rzetelność jest zatem podobna do losowego komponentu całkowitego błędu pomiaru, chociaż punktem odniesienia nie jest prosta próba losowa, ale zróżnicowanie prawdziwych wartości w grupie badanych osób. Zdefiniowany za pomocą wzoru (II.20.) współczynnik rzetelności przyjmuje wartość od 0 do 1, przy czym jeżeli cała zmienność wynika z błędu, tj. gdy $\sigma_Y^2 = \sigma_\varepsilon^2$, to wówczas Rzetelność $(Y) = 0$. Z drugiej strony, gdy błąd zostaje całkowicie wyeliminowany z pomiaru, tj. $\sigma_\varepsilon^2 = 0$, lub inaczej $\sigma_\mu^2 = \sigma_Y^2$, to Rzetelność $(Y) = 1$. Ponieważ rzetelność pomiaru odnosi się zarówno do regularności otrzymywanych wartości tego samego wskaźnika w różnych pomiarach, jak i do homogeniczności różnych wskaźników tego samego *indicatum*, to zasadniczo stosuje się dwie procedury jej empirycznego oszacowania. Pierwsza opiera się na powtórzonych pomiarach tego samego wskaźnika na tych samych osobach, druga natomiast na jednokrotnym, ale wielowskaźnikowym pomiarze tego samego konstruktu (por. Groves i in. 2004: 262–265). Obie te metody wymagają jednak spełnienia pewnych znaczących i problematycznych założeń. W odniesieniu do procedury powtórzonego pomiaru przyjmuje się bowiem, że: (a) pomiędzy dwoma pomiarami nie zachodzi zasadnicza zmiana *indicatum*, tj. zmiana prawdziwych wartości pewnego konstruktu w odniesieniu do każdej badanej osoby, (b) wszystkie istotne czynniki związane z przebiegiem pomiaru pozostają takie same, w taki sam sposób oddziałują na wynik oraz (c) pierwszy pomiar nie ma wpływu na drugi, to jest: uzyskiwane wyniki są niezależne. Chociaż spełnienie tych warunków jest niezwykle problematyczne, to jednak opracowano wiele statystyk pozwalających na oszacowanie rzetelności przeprowadzonego w ten sposób pomiaru (por. Groves i in. 2004: 262–263). Podobnie zresztą empiryczna miara rzetelności pomiaru wielowskaźnikowego opiera się na kilku istotnych założeniach, między innymi na tym, że: (a) wszystkie pytania są wskaźnikami tego samego konstruktu (ich wartości oczekiwane są jednakowe), (b) wszystkie wskaźniki charakteryzują się podobnym rozrzutem odpowiedzi, (c) pomiar poszczególnych wskaźników jest niezależny od siebie, tj. odpowiedzi na jedno pytanie nie mają wpływu na udzielenie odpowiedzi na inne (por. Groves i in. 2004: 264). Reguły te wydają się zdecydowanie mniej restrykcyjne od tych, które sformułowano w procedurze powtórzonego pomiaru. Zresztą to właśnie ten drugi sposób wyznaczania rzetelności jest zdecydowanie bardziej rozpowszechniony w praktyce badawczej. Miarą oceny rzetelności jest doskonale znany w literaturze metodologicznej współczynnik α-Cronbacha, zaprezentowany po raz pierwszy w 1951 roku przez amerykańskiego psychologa Lee J. Cronbacha. Jednak w kontekście przydatności tego wskaźnika dla oceny błędu pomiaru niezwykle doniośle brzmi stwierdzenie autorów monografii *Survey Methodology*, którzy, analizując właściwości miary α-Cronbacha, zwracają uwagę na następującą właściwość tej miary rzetelności:

> wysoka wartość współczynnika α-Cronbacha może oznaczać zarówno wysoką rzetelność, jak i niewielkie zróżnicowanie odpowiedzi. Niestety, może to również oznaczać, że odpowiedzi na jedno pytanie wpływały na odpowiedzi udzielane na inne, co skutkuje znaczną dodatnią korelacją. Niska wartość współczynnika może oznaczać niską rzetelność, ale także to, że wskaźniki nie mierzą tego samego konstruktu. (Groves i in. 2004: 265)

Innymi słowy, zarówno trafność, jak i rzetelność wskaźników nie mają żadnego bezpośredniego przełożenia na koncepcję błędów badań sondażowych, niemniej jednak oba kryteria oceny wydają się niezwykle wartościowe i przydatne we wszystkich dążeniach ukierunkowanych na poprawę jakości prowadzonego pomiaru. Świetnie świadczy o tym zresztą częstość ich wykorzystywania w praktyce surveyowej.

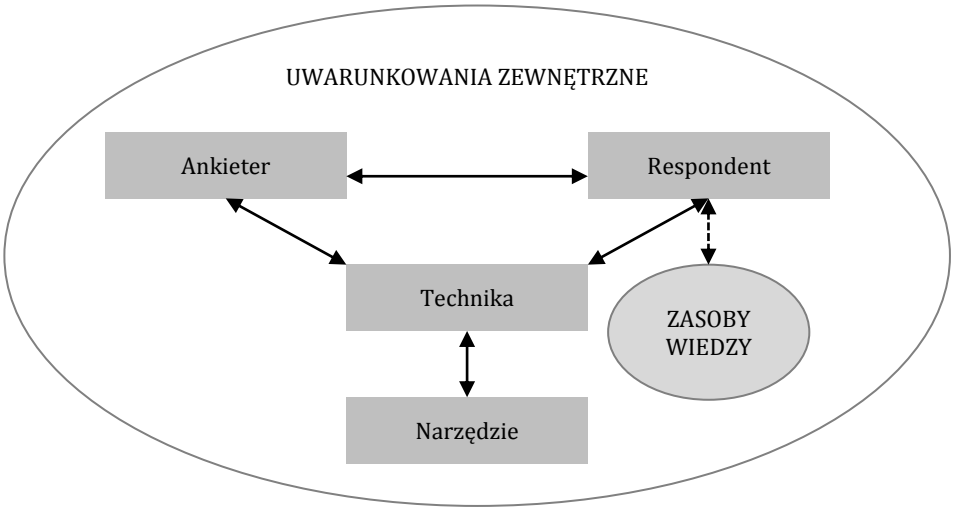
II.2.2. Błędy pomiaru (*measurement errors*)

Literatura z zakresu problematyki pomiaru jest równie bogata, jeśli nawet nie obszerniejsza od tej, która odnosi się do próbkowania czy też zagadnień jednostek niedostępnych. Wynika to w sposób oczywisty z faktu kluczowego znaczenia procesu pomiaru dla całości badania. To w nim w dużej mierze rozstrzygają się kwestie jakości estymacji oraz dokładności formułowanych sądów. Zwraca na to uwagę wielu badaczy, czego typowym przykładem jest opinia Paula Biemera i Larsa Lyberga wyrażona we wprowadzeniu do IV rozdziału monografii *Introduction to Survey Quality*, poświęconego zagadnieniom pomiaru, iż „różne [komponenty procesu pomiaru – P.J.] stanowią, być może, główne źródło błędu nielosowego w sondażach" (Biemer i in. 2003: 116).

Ponieważ jednak szczegółowa charakterystyka tego procesu, zawiłości związanych z różnymi jego aspektami, analiza technik zbierania danych, konsekwencje łączenia różnego rodzaju procedur pomiarowych, zasady koordynacji oraz organizacji badań terenowych, a także inne zagadnienia związane z uzyskiwaniem danych, wykraczają daleko poza temat i cele pracy, to w tej części rozdziału scharakteryzowane zostaną tylko te wątki tematyczne, które konceptualnie powiązane są z omówioną w rozdziale I perspektywą całkowitego błędu pomiaru. Przywołane będą zatem takie pozycje literaturowe, w których pomiar rozpatrywany jest przez pryzmat błędów, w tym do monografii o zasięgu światowym (por. Biemer 2011; Weisberg 2005; Biemer i in. 2003; Groves i in. 2004; Groves 1989), jak i krajowym (por. Sztabiński F. 2011; Sztabiński P.B. i in. 2005; Sztabiński P.B. 1997). Chociaż zakres przedstawionych zagadnień pozostanie uboższy w stosunku do zakresu tematycznego całej dostępnej literatury przedmiotu, to jednak, mimo wszystko, charakterystyka ta będzie posiadała interesującą wartość poznawczą, gdyż zbierze te główne wątki literaturowe,

w których definicje błędów oraz powiązane z nimi procedury wyznaczania ich wielkości odgrywają istotną rolę.

Zacząć wypada od choćby krótkiej charakterystyki procesu pomiaru. W literaturze wyodrębnia się kilka kluczowych komponentów konstytuujących pomiar oraz różnego rodzaju powiązania między nimi. Dla przykładu, w niezwykle syntetycznym opisie przedstawionym przez Biemera oraz Lyberga (2003) w trzech następujących po sobie rozdziałach przywołanej już monografii *Introduction to Survey Quality* pomiar definiowany jest jako proces obejmujący siedem czynników, w tym: (1) metodę oraz technikę zbierania danych, (2) narzędzie badawcze (kwestionariusze wywiadów i ankiet), (3) ankieterów (jeśli w danej technice występują), (4) respondentów, (5) system informacji, tj. zasoby wiedzy wykorzystywane przez respondentów do formułowania udzielanych odpowiedzi, (6) ogół okoliczności zewnętrznych oddziałujących na przebieg pomiaru, a także (7) wzajemne powiązania pomiędzy tymi czynnikami (por. Biemer i in. 2003: 116–119).



Ryc. II.3. Źródła błędów pomiarowych

Źródło: Biemer i in. 2003: 117

Co istotne, taki sposób ujmowania pomiaru nie jest kontestowany w literaturze przedmiotu i pojawia się, w mniej lub bardziej zmienionej formie, w wielu innych opracowaniach, między innymi w: Sztabiński F. (2011: 53–60), Weisberg (2005: rozdziały 4–6 oraz 12), Groves i in. (2004: rozdziały 7–9), Sztabiński P.B. (1997: 31–33), Biemer i in. (1991: sekcje A–D), Groves (1989: rozdziały

7–11), wyznaczając ich autorom zakres zagadnień podejmowanych w ramach analizy różnorakich kwestii związanych z poszczególnymi elementami procesu zbierania danych. Chociaż wymienieni tu autorzy poświęcają każdemu z tych komponentów wiele stron analiz, wyczerpująco opisując potencjalne źródła błędów oraz mechanizmy ich powstawania, to jednak wielkość błędu definiują w sposób formalny wyłącznie dla szczególnej konsekwencji uchybień w pracy ankieterów, która przejawia się w skorelowaniu uzyskiwanych odpowiedzi z osobą przeprowadzającą wywiad. Jest to z całą pewnością skutkiem złożonej natury procesu pomiaru, ale też trudności, na jakie narażone są wszystkie próby wyznaczania wielkości błędu w odniesieniu do całego procesu gromadzenia danych. Wystarczy zauważyć, że oszacowanie błędu wypaczenia danych wymaga od badacza dostępu do praktycznie niedostępnych informacji o rzeczywistych wartościach pomiaru, a ponadto takich, które charakteryzują wszystkich badanych lub przynajmniej jakąś ich losową pod-próbę. Wówczas błąd wypaczenia zdefiniować można za pomocą formuły zaprezentowanej przez P. Biemera (2010b: 843):

(II.21.) $$B_{pomiar} \stackrel{\text{def}}{=} \hat{\theta}_{pomiar} - \hat{\theta}'_{pomiar},$$

gdzie:

- $\hat{\theta}_{pomiar}$ jest wielkością estymatora ustaloną na podstawie pomiaru,
- $\hat{\theta}'_{pomiar}$ jest rzeczywistą wartością estymatora w grupie przebadanych osób, ustaloną na podstawie „złotych standardów", a więc takich danych, dla których znane są prawdziwe wartości pomiaru indywidualnie dla każdej przebadanej osoby.

Ponieważ jednak wielkość $\hat{\theta}'_{pomiar}$ znana jest wyłącznie w nielicznych przypadkach i to takich, które rzadko interesują badacza, a nawet jeśli jest osiągalna, to wyłącznie dla tak wąskiej grupy respondentów, że trudno uznać ją zazwyczaj za reprezentację całej próby, to wyznaczanie wielkości systematycznego błędu pomiaru staje się w istocie zadaniem prawie beznadziejnym. Siłą rzeczy badacze skupiają się na tych źródłach błędów, które pozostają mierzalne, w tym na wspomnianych już konsekwencjach wynikających z efektu ankieterskiego (por. Weisberg 2005; Groves i in. 2004; Biemer i in. 2003). Owej specyficznej kombinacji błędu systematycznego oraz losowego poświęcona będzie dalsza część tej sekcji rozdziału.

Warto najpierw wskazać, że w polskiej literaturze metodologicznej wyjątkowo interesującej analizy źródeł błędów związanych z pracą ankietera, zachowaniami respondenta oraz mechanizmami powstawania tych błędów dostarcza publikacja Pawła B. Sztabińskiego *Ankieterzy i ich respondenci* wydana w 1997 roku nakładem Wydawnictwa IFiS PAN w Warszawie, która zbiera wszystkie najważniejsze wątki teoretyczne, metodologiczne oraz empiryczne

w zakresie tej problematyki. Część empiryczna pracy oparta jest na bardzo pouczających oraz na gruncie polskiej metodologii pionierskich badaniach: *Ankieter jako źródło zniekształceń w procesie badawczym* (por. Sztabiński P.B. 1995), zrealizowanych przez jej autora w połowie lat 90. XX wieku. W przywoływanej monografii P.B. Sztabiński wychodzi od krótkiej charakterystyki źródeł błędów popełnianych w badaniach sondażowych, co pozwala mu ulokować błędy związane z osobą ankietera i respondenta w szerszym kontekście zagadnień związanych z pomiarem, następnie dokonuje, skądinąd uzasadnionego, założenia o konieczności łącznego rozpatrywania obu źródeł błędów, a także przeprowadza krótką deskryptywną rekonstrukcję dwóch koncepcji teoretyczno-metodologicznych charakteryzujących proces powstawania błędów ankietera oraz respondenta, czego efektem jest schemat zamieszczony w drugim rozdziale tej publikacji, stanowiący podstawę teoretyczną dla analiz empirycznych przeprowadzonych w ramach autorskiego projektu badawczego (por. Sztabiński P.B. 1997: 40).

Paweł B. Sztabiński przeciwstawia dwa sposoby ujmowania błędów wynikających z pracy ankietera. Pierwszy z nich wywodzi z tez zawartych w opracowaniu Roberta L. Kahna oraz Charlesa F. Cannella *The Dynamic of Interviewing: theory, technique and cases* z 1957 roku, drugi natomiast opiera na propozycji Roberta M. Grovesa (1989) z wielokrotnie już przywoływanej pracy *Survey Errors and Survey Costs*. Nie wchodząc w tym momencie w szczegóły tych dwóch koncepcji metodologicznych, można zauważyć, że P.B. Sztabiński wyróżnia „błędy w pracy ankietera", czyli sposoby oddziaływania na respondentów, oraz „błędy ankietera", czyli rezultaty tego oddziaływania (Sztabiński P.B. 1997: 39). Innymi słowy, zdaniem autora, „o wystąpieniu błędu [ankietera – P.J.] […] można mówić wówczas, gdy stwierdziliśmy, że istotnie ankieterzy oddziaływali na wynik pomiaru. Z kolei [błąd w pracy ankietera – P.J.] […] nie musi prowadzić do uzyskania błędnej odpowiedzi […]. Jeżeli ankieter sugeruje [respondentowi – P.J.] w jakiś sposób swoje własne przekonania w danej sprawie, to sugestii tej respondent może się podporządkować lub też ją odrzucić" (Sztabiński P.B. 1997: 37). Jednocześnie Paweł B. Sztabiński podkreśla, że wprowadzone przez niego rozróżnienie pozostaje w znacznym stopniu zbieżne ze znanym w polskiej metodologii badań sondażowych pojęciem „efektu ankieterskiego" oraz „wpływu ankieterskiego" (por. Lutyńska 1978: 147)[44]. W od-

[44] Opis tych dwóch terminów odnajdujemy również w późniejszym opracowaniu autorki z 1993 roku pt. *Surveye w Polsce* (por. Lutyńska 1993: 95–108) oraz artykule *Wpływ ankieterski w pierwszej fazie badań kwestionariuszowych* (por. Lutyńska 1997: 53–71). Lutyńska definiuje „wpływ" jako pozytywne lub negatywne oddziaływanie ankietera na wyniki badań, które dokonuje się poprzez jego cechy psychiczne, wygląd zewnętrzny, sposób komunikowania się oraz zapisywania odpowiedzi respondentów. Miarą wpływu ankieterów na badanie jest na przykład liczba informacji, których nie udało się uzyskać ankieterowi, czy też liczba informacji błędnych. Z kolei

różnieniu od Krystyny Lutyńskiej, Paweł B. Sztabiński nie uważa jednak za stosowne używania przymiotników *pozytywny* oraz *negatywny* w odniesieniu do pojęcia *wpływu ankieterskiego*, gdyż w jego opinii oddziaływanie to może być wyłącznie negatywne[45] (por. Sztabiński P.B., 1997: 39).

Wynika to w sposób jednoznaczny z przyjętej definicji błędu, który rozpatrywany jest jako zanieczyszczenie „prawdziwych" wyników pomiaru. Paweł B. Sztabiński mówi zatem, iż „[j]eśli przyjąć, że odpowiedzią 'prawdziwą' respondenta jest ta, której udzieliłby sam sobie na zawarte w kwestionariuszu pytanie" (Sztabiński P.B. 1997: 36), to „błędem […] jest każdy wpływ ankietera na uzyskaną i zanotowaną w kwestionariuszu informację" (Sztabiński P.B. 1997: 44). Taki sposób rozumienia błędu jako „zanieczyszczenia" wyników pomiaru, a także swoiste definiowanie wartości „prawdziwej" jako tej, którą respondent sam sobie by udzielił, wywodzi się z przyjmowanej przez autora perspektywy psychometrycznej, przeciwstawianej ujęciu statystycznemu, w którym błędem byłoby zanotowanie informacji niezgodnej ze stanem obiektywnym. Nie wchodząc w polemikę z P.B. Sztabińskim, można jedynie się zastanowić, czy respondent udzieliłby sam sobie odpowiedzi na pytanie, którego pewnie nigdy by sobie sam nie zadał; z drugiej jednak strony, każde badanie jest sytuacją sztuczną, mającą na celu wywołanie pewnych odpowiedzi i jako taką należy ją rozpatrywać, można zatem w schemacie idealizacyjnym przyjąć, że respondent niepoddany żadnemu oddziaływaniu ze strony ankietera lub samego badacza udziela sobie prawdziwej odpowiedzi na przedstawione mu jakoś w „czystej" postaci pytanie kwestionariuszowe[46]. Pomijając sytuacje, w których przedmiotem badania są rzeczywiście zjawiska wykraczające poza pomiar bezpośredni, to przyjęte przez P.B. Sztabińskiego rozumienie wartości prawdziwej uznać można za niezwykle wygodne w tym oczywiście sensie, że pozwala uniknąć problemów wynikających z tego, iż odpowiedź badanej osoby, jakiej udzieliłaby ona sama sobie, może pozostawać – pomimo szczerych intencji respondenta – nadal niewłaściwa. Wystarczy przywołać trywialny przykład, w którym odpowiedź „nie wiem" na pytanie o wielkość dochodów przypadających na jedną osobę w danym gospodarstwie domowym, wynika z rzeczywistej niewiedzy respondenta, ale nadal pozostaje niezgodna ze stanem faktycznym.

„efekt" definiuje ona jako różnicę między wynikami otrzymywanymi przez poszczególnych ankieterów, wynikającą z ich wpływu na badanie. Zakłada się przy tym, że skoro respondenci przydzielani byli ankieterom w sposób losowy, to nie powinny zachodzić zależności pomiędzy uzyskiwanymi odpowiedziami a ankieterami, którzy wywiad przeprowadzali.

[45] Dla przykładu, w opinii P.B. Sztabińskiego stymulowanie respondenta do odpowiedzi nie jest pozytywnym oddziaływaniem, gdyż właśnie na tym polega rola ankietera, natomiast brak takiego zachęcania jest wpływem negatywnym, w odróżnieniu od Lutyńskiej, dla której nie jest to ani wpływ pozytywny, ani negatywny.

[46] Na przykład za pomocą ankiety, tak jak sugerują autorzy monografii *Survey Methodology* (por. Groves i in. 2004: 270).

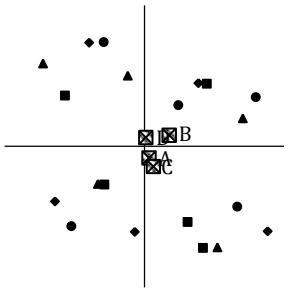
Kwestie związane z pamięcią oraz wiedzą respondentów próbowali włączyć do analiz błędów pomiarowych oraz oceny jakości danych uzyskiwanych w badaniach sondażowych między innymi F. Sztabiński (2011: 59), Groves i in. (2004: 213–218), jak również P. Biemer i in. (2003: 116–119).

Wracając do ustaleń Pawła B. Sztabińskiego, warto wskazać, że w ramach analizy potencjalnych czynników odpowiedzialnych za pojawianie się błędów będących efektem interakcji zachodzącej pomiędzy ankieterem oraz osobą przez niego indagowaną, autor wyróżnia: (1) charakterystyki obserwowalne, podzielone dodatkowo na: (a) społeczno-demograficzne, takie jak: płeć, wiek itd. oraz (b) „pozaspołeczne", takie jak: ogólne wrażenie, sposób zachowania, stosunek do sytuacji wywiadu itp., a także (2) cechy nieobserwowalne, określone w pracy P.B. Sztabińskiego mianem psychologicznych. Wśród nich autor wymienia: (a) charakterystyki osobowościowe, (b) system wartości oraz związane z nimi wzory reagowania (zachowywania się) w określonych sytuacjach, (c) opinie ankietera na tematy poruszane w wywiadzie i ocenę ich doniosłości z punktu widzenia ankietera i respondenta, (d) wyobrażenia oraz stereotypy, a także (e) elementy bezpośrednio związane z wywiadem, między innymi motywacje do uczestnictwa w wywiadzie, a w przypadku ankietera – jego umiejętności i wiedzę. Wszystko to zaś przekłada się na określone zachowania ankieterów (na przykład sposoby zadawania pytań, notowanie odpowiedzi, zachowania niewerbalne, atmosferę wywiadu i inne) oraz zachowania respondentów (na przykład sposób udzielania odpowiedzi lub atmosferę panującą pomiędzy aktorami wywiadu) (por. Sztabiński P.B. 1997: 39–92). Chociaż przedstawiony przez P.B. Sztabińskiego zestaw czynników stanowiących źródło błędu ankieterskiego oraz błędu respondenta posiada w polskiej metodologii badań sondażowych niekwestionowaną wartość poznawczą, to jednak, co warte podkreślenia, wpisuje się (z niewielkimi modyfikacjami) w schematy i standardy wypracowane w literaturze światowej. Ponieważ jednak, jak wspomniano na wstępie, dokładna charakterystyka wpływu określonych zachowań a także oddziaływanie obserwowalnych oraz nieobserwowalnych cech ankietera i respondenta na uzyskiwane wyniki wykracza daleko poza cele tej pracy, to wystarczy poprzestać na stwierdzeniu, że poszczególne czynniki pozostają ze sobą powiązane, prowadząc do uzyskania odpowiedzi obarczonych błędem (por. Sztabiński P. B. 1997: 185–211).

Koncentracja uwagi na ankieterach jako źródle błędu pomiaru bierze się stąd, że „odgrywają oni główną rolę w całym procesie badawczym, […] pochłaniając znaczną część kosztów oraz istotnie oddziałując na jakość danych" (Groves i in. 2004: 269). Dla przykładu, w świetle paradygmatu całkowitego błędu pomiaru badań sondażowych niezwykle ciekawą konsekwencją błędu ankietera jest to, że udzielane przez respondentów odpowiedzi pozostają, w mniejszym lub większym stopniu, skorelowane z osobą przeprowadzającą

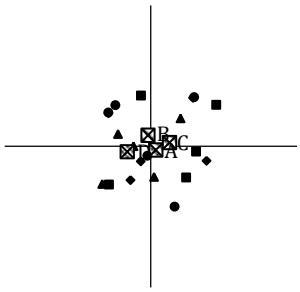
wywiad, czego skutkiem jest przyrost wariancji i/lub systematyczne wypaczanie wartości przeprowadzonego pomiaru. Rozważyć wystarczy przykład pytania, w ramach którego respondent proszony jest o podanie przeciętnego dochodu w gospodarstwie domowym przypadającego na 1 domownika. Sformułowane w ten sposób pytanie wymaga od respondenta wiedzy nie tylko o sumie dochodów wszystkich osób w jego gospodarstwie domowym, co już stanowić może wyzwanie dla znacznej części badanych, ale także przeliczenia łącznej sumy dochodu przez liczbę członków gospodarstwa domowego. Nietrudno sobie wyobrazić, że udzielenie poprawnej odpowiedzi wymagać będzie uszczegółowienia sposobu, w jaki należy wyznaczyć poszukiwaną wielkość (pojawić może się wiele wątpliwości i pytań związanych na przykład z różnymi źródłami dochodów, liczbą osób koniecznych do uwzględnienia w obliczeniach itd.). Co oczywiste, wszystkie takie działania wspomagające respondenta powinny być precyzyjnie określone w przygotowanej przez badacza instrukcji ankieterskiej, zawierającej opis wystandaryzowanych działań zwiększających szanse na udzielenie przez badaną osobę odpowiedzi zgodnej ze stanem faktycznym.

Jeżeli zatem ankieterzy postępowaliby zgodnie z przygotowaną przez badacza instrukcją, a zarazem – używając terminologii P.B. Sztabińskiego – inne cechy pozaspołeczne lub nieobserwowalne nie wpływałyby na odpowiedzi respondentów, to wartości pomiaru powinny być zbliżone do rzeczywistych wartości parametrów, co oznaczałoby, że ankieter (i ewentualnie respondent) nie był źródłem błędu. Jeżeli jednak ankieterzy nie przywiązywaliby szczególnej uwagi do przygotowanej instrukcji i na ewentualne wątpliwości ze strony respondenta o to, w jak sposób wyznaczyć wielkość przeciętnego dochodu, prosiliby o jego oszacowanie, a nie o dokładną wartość, to zapewne część respondentów zawyżyłaby, a część zaniżyłaby wielkość dochodu. Niemniej jednak, pomimo tej nieścisłej instrukcji, wartość estymatora mogłaby nadal pozostać zbliżona do rzeczywistej wielkości parametru, bowiem wynikiem wpływu ankieterskiego byłoby zwiększenie wariancji wyników, ale raczej nie ich systematyczne wypaczenie. Z kierunkowym zniekształceniem wyników miałoby się jednak do czynienia wtedy, gdyby ankieterzy z jakichś powodów sugerowali zawyżanie lub zaniżanie wielkości dochodów. Zresztą nie musiałoby się to odbywać w sposób bezpośredni, bowiem cechy ankieterów, na przykład sposób wyrażania się lub ubierania, mogłyby skłaniać respondentów do zawyżania lub zaniżania wielkości swoich dochodów. Wreszcie bodaj najciekawszą sytuacją byłby przypadek, w którym kierunkowe przesunięcie wyników pomiaru nie byłoby jednakowe dla wszystkich ankieterów. Część z nich mogłaby postępować zgodnie z instrukcją, część prosić o oszacowanie, jeśli już nie dokładne wyznaczenie wielkości dochodu, inni ankieterzy mogliby sugerować zawyżenie, wreszcie pozostali zaniżenie poziomu dochodów. Innymi słowy, systematyczne



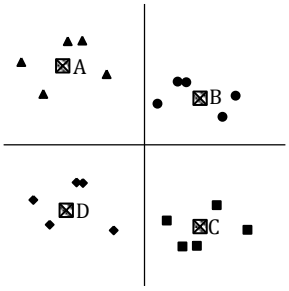
(a) niewielkie skorelowanie wyników, znaczna wariancja, niewielki błąd systematyczny,

tj. $\rho_{ANK} \cong 0, Var(\hat{\theta}) > 0, |B(\hat{\theta})| \cong 0$



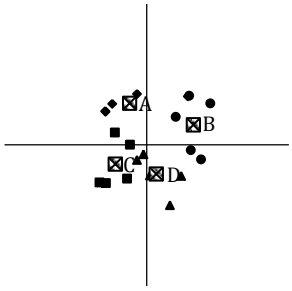
(b) niewielkie skorelowanie wyników, mała wariancja, niewielki błąd systematyczny,

tj. $\rho_{ANK} \cong 0, Var(\hat{\theta}) \cong 0, |B(\hat{\theta})| \cong 0$



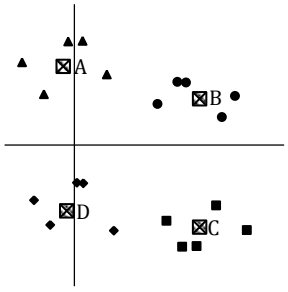
(c) znaczne skorelowanie wyników, znaczna wariancja, niewielki błąd systematyczny,

tj. $\rho_{ANK} > 0, Var(\hat{\theta}) > 0, |B(\hat{\theta})| \cong 0$



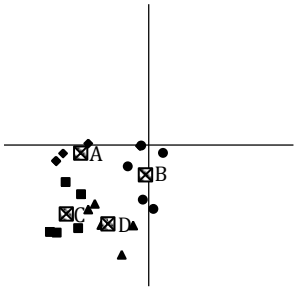
(d) znaczne skorelowanie wyników, mała wariancja, niewielki błąd systematyczny,

tj. $\rho_{ANK} > 0, Var(\hat{\theta}) \cong 0, |B(\hat{\theta})| \cong 0$



(e) znaczne skorelowanie wyników, znaczna wariancja, znaczny błąd systematyczny,

tj. $\rho_{ANK} > 0, Var(\hat{\theta}) > 0, |B(\hat{\theta})| > 0$



(f) znaczne skorelowanie wyników, mała wariancja, znaczny błąd systematyczny,

tj. $\rho_{ANK} > 0, Var(\hat{\theta}) \cong 0, |B(\hat{\theta})| > 0$

Ryc. II.4. Schemat skorelowania wyników pomiaru z osobą ankietera

Źródło: opracowanie własne

wypaczenie wyników mogłoby występować u części ankieterów, a u innych nie, ponadto jeśli by się już pojawiło, to mogłoby być niejednorodne. Ten fenomen występowania różnic pomiędzy przeciętnymi wartościami pomiaru uzyskiwanymi przez poszczególnych ankieterów, tj. różnic w kierunku systematycznych wypaczeń wyników, znany jest w literaturze metodologicznej poświęconej błędom pomiarowym pod pojęciem „wariancji ankieterskiej”[47] (por. West i in. 2010: 1004–1026; Weisberg 2005: 53–63; Biemer i in. 2003: 156–157; Groves 1989: 67), której oddziaływanie na wielkość całkowitego błędu pomiaru badań sondażowych zostało w precyzyjny sposób określone w postaci miernika liczbowego.

Zanim jednak wielkość ta zdefiniowana zostanie w sposób formalny, należy rozpatrzyć najpierw możliwe schematy oddziaływania ankieterów na uzyskiwane wyniki. Kolejne diagramy na rycinie II.4. symbolizują pomiary różniące się skorelowaniem wyników z osobą ankietera (tj. poziomem wariancji ankieterskiej ρ_{ANK}), poziomem wariancji estymatorów ($Var(\hat{\theta})$) oraz wielkością błędu systematycznego ($B(\hat{\theta})$). Dla uproszczenia przyjęto, że badania realizowało czterech ankieterów oraz że każdy z nich przeprowadził po pięć wywiadów. Odnotowywane przez ankieterów odpowiedzi oznaczono jako punkty, z kolei kwadrat z przyporządkowanym symbolem A, B, C lub D określa przeciętną wartość z odpowiedzi uzyskanych przez ankietera.

W pierwszej kolejności można zauważyć, że w obrębie każdej pary pomiary (a) i (b), (c) i (d) oraz (e) i (f) różnią się wariancją wyników. Ponadto, (a) i (b), podobnie jak (c) i (d), symbolizują pomiary o niewielkim błędzie systematycznym, inaczej niż (e) oraz (f), dla których przeciętna wartość pomiaru (tj. wartość estymatora) nie pokrywa wartości prawdziwej. W świetle problemów będących skutkiem efektu ankieterskiego o wiele bardziej interesujące staje się jednak prześledzenie skorelowania wyników z osobą ankietera. Wymaga to postawienia dodatkowych założeń oraz krótkiej charakterystyki sposobów przydziału jednostek obserwacji do kolejnych ankieterów.

Przede wszystkim trzeba pamiętać o tym, że poziom skorelowania wyników pomiaru z osobą ankietera będzie możliwy do wyznaczenia tylko i wyłącznie wtedy, gdy spełniony pozostanie warunek losowego przyporządkowywania respondentów do ankieterów (por. Biemer 2010a: 842; Biemer i in. 2003: 167–159). Nie stanowi to szczególnego problemu przy realizacji badań w oparciu o wywiady telefoniczne (por. Groves i in. 2004: 361–362), ale staje się poważnym wyzwaniem w wywiadach bezpośrednich. P. Biemer oraz L. Lyberg (2003: 157) zauważają bowiem, że typowy przydział jednostek badawczych odbywa się najczęściej z uwzględnieniem kryterium minimalizacji kosztów

[47] W literaturze zamiennie używa się pojęć: *interviewer variability*, *interviewer variance*, *interviewer error*, *interviewer effect, interviewer design effect*.

podróży oraz czasu koniecznego na dotarcie do miejsca zamieszkania respondenta, stąd respondentów przypisuje się do odpowiednich jednostek terytorialnych, a następnie przydziela do tych jednostek odpowiednich ankieterów, zamieszkujących w pobliżu określonego obszaru. Przy takim sposobie organizacji badań terenowych poziom skorelowania wyników z osobą ankietera pozostanie tak długo trafną miarą ich wpływu na uzyskiwane wyniki, jak długo rozkład mierzonych cech jest niezależny od ich rozkładu w strukturze przestrzennej (por. Weisberg 2005: 55). Zatem, jeżeli charakterystyki populacyjne osób zamieszkujących poszczególne obszary pozostaną homogeniczne w ich obrębie, a heterogeniczne pomiędzy nimi, to porównywanie wyników otrzymywanych przez kolejnych ankieterów nie będzie miało sensu. Innymi słowy, we wszystkich takich przypadkach otrzymane różnice będą pochodną korelacji przestrzennej i nie będą już wynikać z różnic (lub precyzyjniej, nie da się ich wytłumaczyć na postawie różnic) w oddziaływaniu ankieterów na wynik pomiaru. Na tę niedogodność zwracają uwagę Groves i in. (2004), wyodrębniając dwie możliwe przyczyny odnotowywania przez ankieterów pewnych specyficznych, w porównaniu z innymi ankieterami, wyników pomiaru. Po pierwsze, jednym z możliwych wyjaśnień tego fenomenu jest to, że ankieterzy w istotny sposób wpływali na odpowiedzi udzielane przez respondentów, drugim natomiast, że to jednak przypisani kolejnym ankieterom respondenci różnili się wartościami mierzonych zmiennych. Inaczej rzecz ujmując, autorzy ci zwracają uwagę na to, że „w celu oszacowana czystego [tj. nieobarczonego innymi czynnikami – P.J.] wpływu wywieranego przez ankieterów na respondentów zachodzi konieczność wyeliminowania efektu będącego skutkiem występowania rzeczywistych różnic pomiędzy respondentami" (Groves i in. 2004: 274–275). Powracając do przywołanego przykładu związanego z pomiarem dochodu przypadającego na jednego członka gospodarstwa domowego, może się na przykład okazać, że pewna grupa ankieterów odnotowywała większy lub mniejszy poziom dochodów tylko dlatego, że przydzieleni im respondenci z jakiegoś określonego obszaru charakteryzowali się systematycznie wyższym lub niższym poziomem dochodów od respondentów zamieszkujących inne obszary. Wtedy różnice te nie będą świadczyły o odmiennym oddziaływaniu ankieterów na wyniki pomiaru, ale będą konsekwencją realnego zróżnicowania cech respondentów.

W odniesieniu do badań realizowanych techniką wywiadów kwestionariuszowych znaczna część autorów, chociażby Biemer (2011: 332–333), Gillikin (2008: 359–360), Weisberg (2005: 55), oraz Groves (2004: 360–364), odwołuje się do jednej z metod losowego przydziału respondentów do ankieterów o trudnej do przełożenia na język polski nazwie *interpenetrated sample assignment* (w skrócie *ISA*). Stanowi ona zbiór wytycznych, opracowanych w 1946 roku przez hinduskiego statystyka Prasanta C. Mahalanobisa, zawierający ze-

staw rekomendacji określających zasady zrandomizowanego przyporządkowania jednostek obserwacji kolejnym ankieterom. Chodzi o to, aby przydzielani respondenci stanowili swego rodzaju reprezentatywną (pod)próbę całej badanej populacji (por. Biemer 2011: 332), co w założeniu powinno umożliwić oszacowanie efektu ankieterskiego, z wyłączeniem zróżnicowania wynikającego z występowania rzeczywistych różnic pomiędzy respondentami (por. Groves i in. 2004: 274). W najbardziej ogólnym sensie dąży się do tego, aby w oparciu o liczebność próby badawczej oraz liczbę ankieterów zaangażowanych w realizację badań terenowych, przyporządkować każdemu ankieterowi w sposób zrandomizowany taką samą część wywiadów. Takie losowe przyporządkowanie nie jest problemem w badaniach realizowanych na względnie małych obszarach, staje się jednak poważnym wyzwaniem w projektach badawczych o większym zasięgu przestrzennym (trudno sobie wyobrazić, aby ankieter realizował wywiady rozproszone na przykład w całej Polsce). We wszystkich takich przypadkach rekomendacja *ISA* przyjmuje nieco inną postać, a mianowicie losowy przydział wywiadów do ankieterów powinien odbywać się w obrębie mniejszych jednostek terytorialnych, co nie pozwala oczywiście na wyznaczenie efektu oddziaływania ankieterów w odniesieniu do całej próby badawczej, ale przynajmniej umożliwia jego oszacowanie w obrębie ankieterów pracujących na określonym obszarze (por. Groves i in. 2004: 275).

Przyjmując zatem, że poszczególni respondenci przydzielani byli ankieterom w sposób losowy, można powrócić do analizy schematów, które zaprezentowane zostały na rycinie II.4. Ponieważ pomiary (c) i (e) oraz (d) i (f) różnią się poziomem błędu systematycznego, ale już nie stopniem skorelowania wyników z osobą ankietera, to rozważone zostaną jedynie różnice zachodzące w parach (a) – (c) oraz (b) – (d). Na rycinach (a) i (b) widać wyraźnie, że wyniki pomiarów uzyskane przez kolejnych ankieterów nie tworzą żadnych odseparowanych od siebie klastrów i choć cechują się mniejszą lub większą wariancją, to jednak uśrednione dane są do siebie zbliżone. Przeciwieństwem tej sytuacji są pomiary zaprezentowane na rycinach (c) oraz (d). Widać na nich, że wartości zanotowane przez każdego ankietera są odmienne od wyników innych ankieterów, stąd można wnioskować, że ankieterzy w różny sposób wpływali na pomiar, co w efekcie przełożyło się na wzrost wariancji w zbiorze surowych wyników oraz obniżyło precyzję estymacji (por. Weisberg 2005: 55, Biemer i in. 2003: 157).

Można już teraz przejść do zdefiniowania miernika opisującego skalę oddziaływania ankieterów na poziom precyzji pomiaru. Choć istnieje wiele miar, które można tu zastosować, to jednak najbardziej znanym oraz najczęściej wykorzystywanym miernikiem jest – zdefiniowany przez L. Kisha (1962: 92–115) – współczynnik korelacji wewnątrzklasowej (z ang. *intraclass correlation coefficent*, oznaczany tu symbolem ρ_{ANK}) (por. Biemer i in. 2003: 162). Współczynnik

ten stanowi element składowy innej miary, tzw. $DEFF_{ANK}$, która określa skalę przyrostu wariancji w zbiorze wyników obarczonym efektem ankieterskim (por. na przykład Biemer 2011: 332; Biemer 2010b: 44; West i in. 2010: 1005; Weisberg 2005: 55–56; Groves i in. 2004: 275; Groves 1989: 364). Co oczywiste, wyznaczenie wartości współczynnika $DEFF_{ANK}$ pozostanie zasadne jedynie pod warunkiem losowego przydziału respondentów do ankieterów i umożliwia odróżnienie sytuacji pomiaru (a) od (c) oraz (e) i odpowiednio (b) od (d) oraz (f), ale już nie (c) od (e), czy też (d) od (f). Jest to konsekwencją tego, że współczynnik korelacji wewnątrzklasowej pozostaje miarą zróżnicowania wyników pomiędzy grupami odpowiedzi odnotowanymi przez poszczególnych ankieterów, ale już nie jest miernikiem ich systematycznego wypaczenia[48]. Innymi słowy, wielkość miary $DEFF_{ANK}$ będzie taka sama dla schematów (c) oraz (e), jak i dla (d) oraz (f). W sensie obliczeniowym $DEFF_{ANK}$ jest wielkością wyznaczaną w ten sam sposób, w jaki ustala się przyrost wariancji w próbie dobranej zgodnie ze schematem losowania zespołowego/wiązek respondentów (por. Biemer 2011: 332; Park i in. 2004: 184), co oznacza, że przyjmuje postać określoną wzorem:

(II.22.) $DEFF_{ANK} \stackrel{\text{def}}{=} 1 + (\overline{m} - 1)\rho_{ANK}$ (por. Biemer 2011: 332),

gdzie:

- $\overline{m}$ jest przeciętną ważoną liczbą wywiadów zrealizowanych przez każdego z M ankieterów, przy czym $\overline{m} = \frac{1}{M-1}\left(N_{pomiar} - \sum_{m=1}^{M}\frac{n_m^2}{N_{pomiar}}\right)$, N_{pomiar} oznacza liczebność próby zrealizowanej, a n_m, gdzie $n_1+n_2+\ldots+n_M = N_{pomiar}$, jest liczbą wywiadów zrealizowanych przez m-tego ankietera (m = 1, 2, …, M);
- ρ_{ANK} odpowiada wspominanemu wcześniej wewnątrzklasowemu współczynnikowi korelacji. Dla danych ciągłych[49] jego wartość można oszacować, wykorzystując model jednoczynnikowej analizy wariancji (por. Groves 1989: 363–364). Przyjmuje ona wówczas postać wyrażenia $\hat{\rho}_{ANK} = \frac{MSB-MSW}{MSB+(\overline{m}-1)MSW}$ (por. Gabler i in. 2008: 196; Ukoumunne 2002:

[48] Możliwość wyznaczenia poziomu wypaczenia danych wymaga dodatkowych studiów, których idea polega na porównaniu odpowiedzi uzyskanych przez ankieterów z odpowiedziami tych samych respondentów, które zostały przez nich samodzielnie zapisane w ankiecie. W opinii Grovesa i in. (1989: 270) różnice w odpowiedziach świadczą o pojawieniu się błędu (którego rozumienie jest takie same jak u P.B. Sztabińskiego (1997), czyli oznacza zapisanie odpowiedzi innej niż ta, której respondent udzieliłby sam sobie). Dodatkowo, gdy zaobserwuje się stałą kierunkową deformację wyników, to różnice takie są wskaźnikiem błędu systematycznego.

[49] W ciekawym artykule autorstwa Sandry M. Eldridge, Obioha C. Ukoumunne oraz Johna B. Carlina zamieszczonym w czasopiśmie „International Statistical Review" odnajdujemy definicję współczynnika korelacji wewnątrzklasowej dla przypadku danych dychotomicznych (por. Eldridge i in. 2009: 382–385).

3760), gdzie *MSB* oznacza średni kwadrat zróżnicowania międzyklasowego[50] (tj. między ankieterami), a *MSW* odpowiada średniemu kwadratowi zróżnicowania wewnątrzklasowego[51] (tj. w obrębie wyników otrzymywanych przez kolejnych ankieterów).

Co oczywiste, wielkość ρ_{ANK} wyznacza się podobnie jak wielkość całkowitego błędu pomiaru w odniesieniu do konkretnego estymatora, nie zaś dla pomiaru wszystkich zmiennych jednocześnie. Jeżeli zatem, w ramach pewnego pytania, udzielane odpowiedzi byłyby bardziej do siebie podobne w zbiorze osób indagowanych przez konkretnego ankietera niż wśród respondentów badanych przez różnych ankieterów, to wielkość miary $\hat{\rho}_{ANK}$ byłaby znaczna, osiągając nawet wartość bliską jedności – czyli teoretycznej wielkości maksymalnej. Jeżeli jednak odpowiedzi uzyskane przez danego ankietera pozostają zbliżone do odpowiedzi uzyskiwanych przez wszystkich innych ankieterów, to wówczas wartość wskaźnika $\hat{\rho}_{ANK}$ zbliża się do zera. Na uwadze należy mieć jednak fakt, na który wskazuje wielu autorów (por. na przykład Biemer 2011: 332; Groves i in. 2004: 227), że nawet bliskie zeru wartości współczynnika korelacji wewnątrzankieterskiej mogą mieć znaczący wpływ na przyrost wariancji w zbiorze uzyskanych wyników pomiaru. Dla przykładu R. Groves oraz inni współautorzy monografii *Survey Methodology*, powołując się na wcześniejsze analizy opublikowane samodzielnie przez Grovesa w 1989 roku (w których prześledził on wielkości współczynników $\hat{\rho}_{ANK}$ w sondażach różnego typu), ukazują, że nawet przy przeciętnej wielkości $\hat{\rho}_{ANK} = 0{,}01$, uzyskanej *notabene* w badaniach realizowanych techniką wywiadów telefonicznych[52], przyrost

---

[50] $MSB = \frac{\sum n_m(\bar{X}_{m\cdot} - \bar{X}_{\cdot\cdot})}{M-1}$, gdzie $\bar{X}_{m\cdot}$ jest średnią z wartości odnotowanych przez *m*-tego ankietera, natomiast $\bar{X}_{\cdot\cdot}$ jest średnią ze wszystkich wyników pomiaru.

[51] $MSW = \frac{\sum\sum(X_{mi} - \bar{X}_{m\cdot})}{N_{pomiar} - M}$, gdzie X_{mi} jest kolejnym *i*-tym wynikiem odnotowanym przez *m*-tego ankietera, przy czym i=1, 2, …, n_m, dla m=1, 2, …, M.

[52] P. Biemer oraz L. Lyberg podkreślają, że w badaniach realizowanych za pomocą wywiadów osobistych, przeciętna wielkość $\hat{\rho}_{ANK}$ była nieco wyższa i osiągnęła wielkość 0,03 (por. Biemer i in. 2003: 168). Stąd, jak podkreślają, technika wywiadów telefonicznych wydaje się mniej podatna na skorelowanie wyników, niż jest to widoczne w odniesieniu do wywiadów *PAPI* lub *CAPI*. Zdaniem autorów monografii *Introduction to Survey Quality* wynik ten nie jest zaskakujący, gdyż jednym z podstawowych udogodnień oferowanych w ramach wywiadów telefonicznych pozostaje możliwość scentralizowanego monitorowania zachowań teleankieterskich, w tym prowadzenia nadzoru w trakcie wywiadów. Co oczywiste, w wywiadach realizowanych w oparciu o kontakt bezpośredni możliwość bieżącego monitorowania jest wydatnie ograniczona, może z wyłączeniem sytuacji, w których w trakcie realizacji wywiadu obecna jest osoba nadzorująca poprawność jego przeprowadzenia. Pewnym rozwiązaniem tego problemu wydaje się zastosowanie elektronicznych narzędzi badawczych, które dają możliwość „odsłuchania" przebiegu wywiadu, wymaga to jednak zgody respondenta na nagrywanie (por. Sawiński 2005a: 234–235). Więcej o konsekwencjach wynikających z obecności obserwatora w trakcie realizacji wywiadu przeczytać można między innymi w artykule F. Sztabińskiego (1995: 49–60). Z kolei Zbigniew Sawiński w 29. rozdziale podręcznika

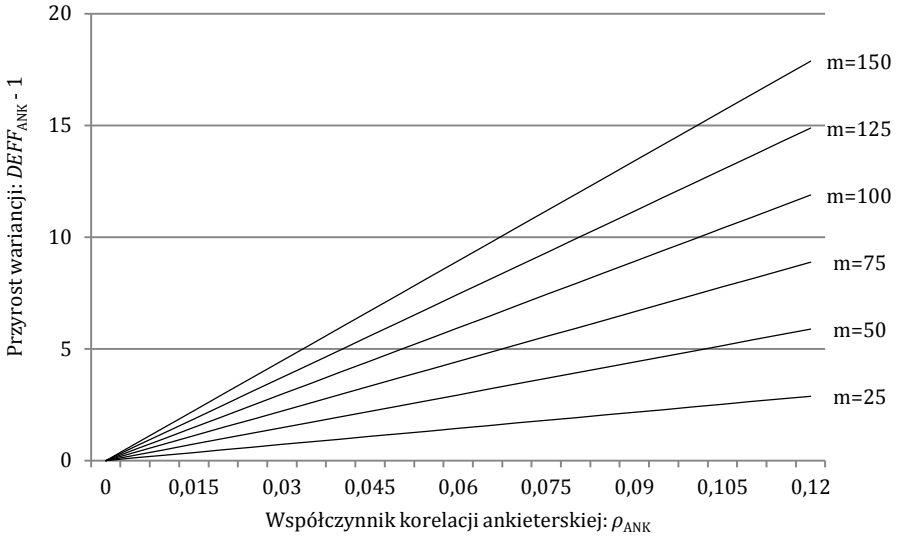
wariancji (określony wartością miary $DEFF_{ANK}$) może być – w najlepszym przypadku – mniej więcej 40-procentowym, co w efekcie przełoży się na znaczne obniżenie precyzji estymacji oraz redukcję efektywnej wielkość próby badawczej (por. Groves i in. 2004: 227–228).

Jednak, pomimo że w wywiadach telefonicznych wartość $\hat{\rho}_{ANK}$ jest zasadniczo mniejsza niż w wywiadach osobistych, to wyniki uzyskiwane przy wykorzystaniu techniki CATI są znacznie bardziej narażone na przyrost wariancji. Innymi słowy, miara $DEFF_{ANK}$ może osiągać znacznie większe wartości w wywiadach telefonicznych, niż ma to miejsce w przypadku wywiadów bezpośrednich (por. Biemer 2011: 169). Powód jest niezwykle prosty: liczba wywiadów zrealizowanych przez jednego ankietera w badaniach telefonicznych jest generalnie znacznie wyższa niż w wywiadach bezpośrednich. Jednocześnie ze wzoru (II.22.) można łatwo odczytać, że $(DEFF_{ANK} - 1) \times 100\%$, czyli wielkość ukazująca procentowa skalę przyrostu wariancji, jest funkcją liniową argumentu $\hat{\rho}_{ANK}$ z parametrem $(\overline{m} - 1)$. Oznacza to, że wielkość $DEFF_{ANK}$ nie zależy wyłącznie od skorelowania wyników uzyskiwanych przez ankieterów, ale także – a może przede wszystkim – od przeciętnej liczby przeprowadzonych przez nich wywiadów. Przywołując skrajny przykład, można zauważyć, że ankieterzy nie będą w ogóle źródłem przyrostu wariancji wyników w sytuacji pełnego skorelowania wartości pomiaru z osobą ankietera. Ten zaskakujący wniosek wynika z tego, że jeżeli powierzyłoby się każdemu ankieterowi realizację tylko i wyłącznie jednego wywiadu, to wewnątrzankieterska wariancja wyników stanowiłoby 0% całkowitej zmienności, inaczej niż międzyankieterska, na podstawie której dałoby się wyjaśnić 100% zmienność uzyskanych wyników. W takim przypadku $\hat{\rho}_{ANK}$ byłaby wprawdzie równa wartości maksymalnej 1, ale jednocześnie, ponieważ $\overline{m} = 1$, to $DEFF_{ANK}$ osiągnęłoby wartość 1, oznaczającą brak wpływu ankietera na przyrost zróżnicowania wyników. Te zależności pomiędzy wielkościami $\hat{\rho}_{ANK}$, $\overline{m}$ oraz $DEFF_{ANK}$ obrazuje rycina II.5.

Przedstawione dane ukazują, że przyrost wariancji w zbiorze pomiaru może być w dużo większym stopniu zależny od przeciętnej liczby wywiadów zrealizowanych przez każdego z ankieterów niż od wielkości współczynnika korelacji wewnątrzankieterskiej. Chociaż w praktyce miara $\hat{\rho}_{ANK}$ zawiera się dla większości pytań kwestionariuszowych w przedziale pomiędzy 0,0 a 0,05 (por. Biemer 2011: 164), to nawet przy tak niewielkim skorelowaniu wyników (przyjmując $\hat{\rho}_{ANK}$=0,05) i przy przeciętnej liczbie 50 wywiadów przypadających na ankietera, poziom wariancji wzrośnie o 245 procent, a przy

---

Fieldwork jest sztuką wydanym wspólnie z Pawłem B. Sztabińskim oraz Franciszkiem Sztabińskim (2005) opisuje projekt badawczy *Program Kontroli Jakości Pracy Ankieterów*, obejmujący zestaw ujednoliconych rekomendacji określających zasady pracy ankieterów oraz sposoby jej kontroli przez instytucje badawcze (por. Sawiński 2005b: 401–410).



Ryc. II.5. Skala przyrostu wariancji na skutek skorelowania wyników pomiaru z osobą ankietera

Źródło: opracowanie własne na podstawie Biemer i in. 2003: 164

liczbie 75 wywiadów na osobę – o ponad 370 procent! Zważywszy na niewielki stopień skorelowania wyników, oznacza to oczywiście olbrzymi przyrost wariancji w zbiorze wyników pomiaru. Obserwacja tych zależności doprowadziła zresztą współautorów monografii *Survey Methodology* do sformułowania następującej rekomendacji, że:

> zaangażowanie większej liczby ankieterów oraz w konsekwencji zmniejszenie liczby zrealizowanych przez każdego z nich wywiadów, jest jednym ze sposobów ograniczenia skutków oddziaływania ankieterów na wielkości błędów standardowych. (Groves i in. 2004: 296)

Nie jest to oczywiście jedyny, ani nawet najważniejszy, sposób ograniczania negatywnych skutków błędu ankieterskiego. W literaturze metodologicznej panuje konsensus co do tego, że kluczową rolę w redukcji błędów związanych z pracą ankieterów odgrywa kontrola jakości ich pracy, a przede wszystkim standaryzacja procesu badawczego (por. Sztabiński P.B. 2005: 50–54), w tym między innymi: procedur doboru osób do badania, sposobów zadawania pytań zgodnie z zapisem w kwestionariuszach, stosowania się do instrukcji ankieterskich opracowanych przez badacza, (samo)kontrola werbalnych i niewerbalnych zachowań ankietera, które mogą wpłynąć na odpowiedzi udzielone przez respondenta, a także dokładne zapisywanie udzielonych odpowiedzi (por. Groves i in. 2004: 278–296).

II.3. Błędy związane z przetwarzaniem danych (*processing errors*)

W literaturze badań surveyowych odnaleźć można tylko nieliczne prace, których autorzy podejmują się analizy błędów przetwarzania danych w ramach teoretycznych wyznaczonych przez paradygmat całkowitego błędu pomiaru. Wystarczy przypomnieć, że nawet w tak przełomowym dziele, jakim dla popularyzacji teorii całkowitego błędu pomiaru była monografia *Survey Errors and Survey Costs*, Robert M. Groves wykluczył z pola swojego zainteresowania błędy procedowania danych. Abstrahując w tym momencie od przedstawionej przez niego argumentacji uzasadniającej rzeczowość takiego podejścia, można uznać, że wbrew swoim deklaracjom nie traktował on błędów będących skutkiem uchybień w przetwarzaniu danych zbyt poważnie. Ponieważ kluczowym zagadnieniem, wokół którego Groves koncentrował swoją uwagę, były nakłady/koszty ponoszone w ramach poszczególnych przedsięwzięć badawczych, a zarazem wiadomo, iż „operacje przeprowadzane na zbiorach danych stanowią znaczną część całkowitego budżetu badawczego […] pochłaniając około 40% kosztów" (por. Biemer i in. 2003: 220), to autor przywoływanej pracy nie powinien, mimo wszystko, tak pochopnie decydować się na wyłączenie ze swoich rozważań błędów procedowania danych[53]. Jest to zresztą przypadłość charakteryzująca większość opracowań z zakresu metodologii badań sondażowych, nie tylko tych poświęconych zagadnieniom błędów pomiarowych. Zwrócili na to uwagę autorzy monografii *Introduction to Survey Quality*, stwierdzając, że:

> literatura dotycząca zagadnień błędów procedowania danych i ich kontroli jest relatywnie dużo mniejsza od tej poświęconej błędom pomiarowym […] oraz jednostkom niedostępnym. […] Pomimo swojego potencjalnego wpływu na wyniki surveyu, błędy procedowania danych traktowane są przez wielu metodologów jako mniej interesujące. […] Chociaż istnieje wiele dowodów na ich wpływ […] na wyniki badań sondażowych, to związane z nimi struktury błędów są zasadniczo nieznane i niezbadane. (Biemer i in. 2003: 219)

---

[53] W wydanej pięć lat później monografii *Survey Methodology* R. Groves i in. (2004: 303–343) poświęcają cały rozdział tzw. postterenowemu etapowi przetwarzania danych, koncentrując się na następujących jego fazach: kodowaniu, wprowadzaniu danych i ich edycji, imputacji braków danych, konstruowaniu wag oraz estymacji poziomu precyzji pomiaru. Główna uwaga autorów skupia się na dwóch fazach tego procesu, tj. na kodowaniu danych oraz konstruowaniu wag, które mogą stanowić poważne źródło przyrostu wariancji oraz wypaczenia w zbiorze danych wynikowych. Ponieważ pozostałe etapy omówione są skrótowo, a charakterystyka procedury kodowania pozostaje tożsama z zaprezentowanym podejściem Biemera i Lyberga (2003), to charakterystyka Grovesa i in. (2004) przywołana będzie tylko fragmentarycznie przy okazji definicji wskaźnika przyrostu wariancji wynikającej z kodowania odpowiedzi respondentów na zadane im pytania „otwarte".

W opinii autorów tej pracy mniejsze zainteresowanie błędami procedowania danych wynika z faktu, że „inaczej, niż w przypadku braków odpowiedzi, czy też projektowania narzędzi badawczych, nie ma modeli teoretycznych opisujących oraz wyjaśniających etap badań związany z przetwarzaniem danych" (Biemer i in. 2003: 219). Bardziej prawdopodobne wydaje się jednak, że ponieważ przetwarzanie danych związane jest z „działaniem analityków oraz osób zarządzających bazami danych" (Groves 1989: 12), to wielu badaczy może być po prostu przekonanych o możliwości skutecznego kontrolowania procesu obróbki danych oraz w efekcie o możliwości wyeliminowania wszelkich potencjalnych źródeł błędów pojawiających się na tym etapie badania. Chociaż jest w tym wiele racji, to jednak, jak stwierdza Herbert F. Weisberg, „staranność dokładana przez badaczy w fazie zbierania danych nie jest często przez nich kontynuowana na etapie przetwarzania danych" (Weisberg 2005: 262). Biemer i Lyberg dodają do tego, że „operacje [na zbiorach danych – P.J.] przeprowadzane są bez żadnej kontroli ich jakości, a co za tym idzie, wpływ tych źródeł błędów na całkowitą wielkość MSE jest zazwyczaj nieznany" (Biemer i in. 2003: 215), poprzedzając przywołaną konstatację dosyć karykaturalnym stwierdzeniem, że „wiedza o procesie przetwarzania danych oraz błędach powiązanych z tym procesem jest bardzo niewielka [nawet – P.J.] w instytucjach sondażowych" (Biemer i in. 2003: 215).

Odwołując się do dwóch monografii, tj. publikacji P. Biemera oraz L. Lyberga (2003), a także opracowania *The Total Survey Error Approach. A Guide to the New Science of Survey Research* autorstwa H. Weisberga (2005), można wskazać, że błędy związane z przetwarzaniem danych, lub inaczej mówiąc błędy postsurveyowe, ulokować należy w obrębie tych wszystkich przedsięwzięć badawczych, które podejmowane są od momentu zakończenia badań terenowych, aż po publikację raportów badawczych. Niezwykle trafnie proces ten scharakteryzowany został w książce *Introduction to Survey Quality*, gdzie we wprowadzeniu do rozdziału siódmego poświęconego w całości błędom procedowania danych oraz procedurom wykorzystywanym do ograniczania ich negatywnych konsekwencji, odnaleźć można następującą definicję:

> [p]rzetwarzanie danych jest zbiorem czynności mających na celu przekształcenie wyników badań sondażowych, uzyskanych w trakcie zbierania danych, z ich surowego stanu do takiej postaci […], która może być wykorzystana w analizie, prezentacji i upowszechnianiu wyników. Podczas tego procesu dane mogą być przetwarzane na wiele sposobów, […] czego celem jest poprawa ich dokładności. Dane mogą być zatem sprawdzane, porównywane, poprawiane, wprowadzane do baz danych, kodowane itd., do czasu, aż [nie osiągną postaci – P.J.] [...], którą uznać można za 'dopasowaną do wymagań' [badaczy – P.J.]. (Biemer i in. 2003: 215)

W podobnym tonie wypowiada się H. Weisberg, który w pierwszym zdaniu wprowadzenia do rozdziału poświęconego błędom popełnianym w fazie przetwarzania danych definiuje błędy postsurveyowe „jako takie, które pojawiają się po zrealizowaniu wywiadów" (Weisberg 2005: 261), a także F. Sztabiński, który, mówiąc o błędzie opracowania, „[ma – P.J.] na myśli sytuację, w której w fazie przetwarzania danych dokonano nieintencjonalnej zamiany zapisu oryginalnych informacji uzyskanych w badaniu na informacje inne co do wartości" (Sztabiński F. 2011: 60). Ponieważ F. Sztabińskiego interesują bardziej kwestie związane bezpośrednio z pomiarem, a nie z opracowywaniem danych (tytułowa *ocena jakości danych w badaniach sondażowych* analizowana jest pod kątem oceny jakości ich pozyskiwania), to tym drugim źródłom błędów oraz samej fazie opracowywania danych autor nie poświęca wiele uwagi. Bardziej szczegółową charakterystykę procesu przetwarzania danych odnaleźć można za to w opracowaniu P. Biemera i in. (2003: 215–257), jak też w monografii H. Weisberga (2005: 261–277). W obu publikacjach ich autorzy prezentują zbieżny co do istoty zestaw działań podejmowanych w ramach postsurveyowego opracowywania wyników badań oraz powiązane z tymi działaniami błędy[54].

W świetle podejmowanych w tej pracy zagadnień ważne jest to, że przekształcenia przeprowadzane na zbiorach danych mogą przyczynić się zarówno do systematycznego zniekształcenia danych wynikowych, jak i do obniżenia ich precyzji, ponieważ oddziałują one odpowiednio na systematyczny oraz losowy komponent całkowitego błędu pomiaru. Dla przykładu H. Weisberg ukazuje, że „błędy związane z wprowadzaniem danych są zazwyczaj losowe, zatem nie będą wypaczać wyników, ale obniżać rzetelność danych" (Weisberg 2005: 266), ale już te związane z edycją danych (przetwarzaniem zbiorów wynikowych), czy też z kodowaniem pytań otwartych, mogą dodatkowo powodować systematyczne wypaczenie danych (por. Weisberg 2005: 267; Biemer i in. 2003: 219). Ustalenia Biemera i Lyberga ukazują również, że automatyzacja procesu przetwarzania danych ma tendencję do generowania błędów systematycznych, podczas gdy operacje wykonywane „ręcznie" rodzą zarówno błędy systematyczne, jak i losowe (por. Biemer i in. 2003: 219).

W sensie formalnym dla dowolnego estymatora $\hat{\theta}$ parametru θ wielkość błędu systematycznego, powstałego na skutek niewłaściwego przetwarzania zbioru danych wynikowych, można zapisać w postaci następującego wzoru:

(II.23.) $$B_{DP} \stackrel{\text{def}}{=} \hat{\theta}_{DP} - \hat{\theta}_{pomiar},$$

[54] Działania te obejmują: (1) wstępną edycję danych, (2) wpisywanie wyników (tworzenie bazy danych), (3) „czyszczenie" baz danych (4) kodowanie danych (w tym pytań otwartych), (5) przygotowanie zbioru danych do obliczeń, (6) analizę danych, (7) raportowanie (por. Biemer i in. 2003: 215–257; Weisberg 2005: 261–277).

gdzie:

- $\hat{\theta}_{DP}$ jest wielkością estymatora, wyznaczoną na podstawie zbioru wyników powstałego po obróbce „surowych" danych;
- $\hat{\theta}_{pomiar}$ jest wielkością estymatora, którą dałoby się wyznaczyć ze zbioru wyników „surowych", gdyby nie trzeba było go poddawać żadnej dalszej obróbce.

Z kolei wielkość błędu losowego można zapisać jako miarę przyrostu wariancji, analogicznie jak dla wskaźnika $DEFF_{TOTAL}$ czy też miernika *VIF*, tzn. jej wielkość w przekształconym zbiorze danych odnieść do do poziomu wariancji w zbiorze „surowych" wyników pomiaru:

(II.24.) $$DEFF_{DP} \stackrel{\text{def}}{=} \frac{\mathrm{Var}(\hat{\theta}_{DP})}{\mathrm{Var}(\hat{\theta}_{pomiar})},$$

gdzie:

- $\mathrm{Var}(\hat{\theta}_{DP})$ jest wariancją w zbiorze danych powstałym po fazie obróbki „surowych" danych wynikowych;
- $\mathrm{Var}(\hat{\theta}_{pomiar})$ jest wariancją w zbiorze wyników niepoddanych przekształceniom, tj. otrzymanych bezpośrednio po przeprowadzeniu terenowej fazy badań.

Należy nadmienić, że chociaż P. Biemer, L. Lyberg oraz H. Weisberg poświęcają błędom procedowania obszerne fragmenty swoich książek, to jednocześnie (inaczej niż w przypadku innych źródeł błędów, może z wyjątkiem błędu specyfikacji określonego przez Biemera oraz Lyberga wyłącznie w sposób opisowy) w ogóle nie definiują formalnie (tj. za pomocą wyrażenia matematycznego) ich wielkości. W obu monografiach uwaga skupia się bowiem bardziej na wypracowaniu standardów edycji baz danych (czego efektem ma być minimalizacja ryzyka wystąpienia błędów), niż na próbie oszacowania wielkości samych błędów. Postulat wypracowania uniwersalnych standardów kodowania pytań oraz edycji baz danych wspomniany będzie jeszcze na końcu tego rozdziału, w tym momencie należy wskazać, że formuły II.23. oraz II.24. podano wyłącznie dla zachowania jednakowego porządku narracji w pracy, bowiem możliwość ich praktycznej implementacji jest dalece ograniczona. Po pierwsze, wielkości błędów procedowania nie powinny być wyznaczane łącznie dla całego etapu postsurveyowej obróbki danych, ale dla każdej czynności z osobna, podobnie zresztą, jak nie wyznacza się ogólnej wielkości błędu całkowitego, tylko estymuje się ich poszczególne komponenty. Trudność polega również na tym, że operacje prowadzone na zbiorach danych wynikowych obejmują na tyle szerokie spektrum działań potencjalnie narażonych na ryzyko błędu, że nie sposób jest ich wszystkich określić. Poza tym próba wyczerpującego ustalenia wszystkich, bez wyjątku, źródeł błędów postsurveyowych pozbawiona byłaby sensu, podobnie zresztą jak bezcelowe są wszelkie próby drobiazgowego wyliczenia

błędów pojawiających się na innych etapach realizacji badań sondażowych. Schemat postępowania powinien być zatem podobny do tego, jaki przyjmowany jest w analizie błędów związanych z próbą badawczą czy też z procesem pomiaru, a zatem uwaga badaczy skupiona powinna być na tych źródłach błędów, których oddziaływanie na całkowity błąd pomiaru jest najbardziej znaczące.

W tym miejscu trzeba wskazać, że poza etapem ważenia danych (opisanym w sekcji II.1.5. tego rozdziału), innym, niezwykle podatnym na błędy momentem w procesie przetwarzania wyników badań jest kodowanie pytań otwartych (kategoryzacja odpowiedzi). Nie jest to wprawdzie proces występujący we wszystkich badaniach sondażowych, jeśli jednak już występuje, to staje się istotnym źródłem błędów (por. Biemer i in. 2003: 234). W sekcji 7.5. rozdziału poświęconego procedowaniu danych P. Biemer oraz L. Lyberg podają opisową definicję błędu kodowania, stwierdzając, iż „występuje on wtedy, gdy [wypowiedzi respondenta – P.J.] przypisano inny kod, niż właściwy”[55] (Biemer i in. 2003: 236), poprzedzając ją założeniem o istnieniu owego „właściwego”, tj. odpowiadającego rzeczywistości oraz intencjom badacza, kodu charakteryzującego wypowiedź respondenta[56] (por. Biemer i in. 2003: 236). Formalne zdefiniowanie tego typu błędu wymaga jednak, choćby krótkiej, charakterystyki procedury kodowania pytań otwartych.

P. Biemer oraz L. Lyberg traktują kodowanie jako proces, którego elementy składowe stanowią: (1) odpowiedzi respondentów w postaci swobodnej wypowiedzi na pytanie otwarte, (2) określone wcześniej klucze kodowe, tzn. zbiory liczb (lub innych symboli) wraz z przypisanymi im kategoriami zmiennej wynikowej oraz opisową charakterystyką każdej kategorii, a także (3) instrukcje kodowe, które zawierają zbiór reguł pozwalających osobie zaangażowanej w kodowanie pytań na powiązanie (1) z (2), tj. udzielonych odpowiedzi z odpowiednimi dla nich wartościami zmiennej wynikowej (por. Biemer i in. 2003:

---

[55] Poza domenę błędów kodowania wyłączone są te wszystkie przypadki, które obejmują sytuacje przypisania właściwego kodu numerycznego lub alfanumerycznego dla określonej odpowiedzi respondenta, udzielonej jednak niezgodnie ze stanem rzeczywistym. W konsekwencji, używając statystycznej terminologii błędów pomiarowych, nadany kod nie oddaje wartości prawdziwej, co w oczywisty sposób przesuwa pole zainteresowania w kierunku błędów pomiarowych.

[56] Biemer i in. (2003: 236) uwypuklają przy tym trudności, na jakie napotyka badacz w związku z koniecznością spełnienia założenia o istnieniu tego „właściwego” kodu. Problemy mogą się pojawić zarówno w sytuacji udzielenia przez respondenta niejasnej lub niejednoznacznej odpowiedzi, ale także wtedy, gdy jest ona precyzyjna. Autorzy publikacji *Introduction to Survey Quality*, odwołując się do studiów empirycznych, zwracają uwagę na znaczne różnice w kodach nadawanych tym samym odpowiedziom przez różne osoby. W konsekwencji badacze stają przed koniecznością wyznaczenia dodatkowych reguł umożliwiających jednoznaczne wyodrębnienie „właściwego” kodu odpowiedzi. Przykładem takich reguł jest między innymi omówiona dalej w tym rozdziale tak zwana *uproszczona weryfikacja kodowania* oraz *dwustopniowa niezależna weryfikacja kodowania z głosem rozstrzygającym*.

234–235). Kodowanie polega zatem na przyporządkowywaniu swobodnych wypowiedzi respondentów do ustalonych kategorii, z których każda wyklucza inną, jednak łącznie wypełniają przestrzeń wszystkich możliwych sytuacji[57]. Innymi słowy, „surowe" wypowiedzi badanych osób zamieniane są, zgodnie z regułami określonymi przez badacza, na predefiniowane kody, tak aby usystematyzować wypowiedzi respondentów w postaci rozkładów częstości lub też wykorzystać je jako wyróżnik grup porównawczych w parametrycznych lub nieparametrycznych analizach statystycznych. Doskonałym przykładem procesu kodowania odpowiedzi pozostają kwestie związane z pytaniami o wykonywany zawód. Predefiniowane kategorie obejmować mogą przy tym nawet dziesiątki kodów numerycznych, czego świetną ilustracją jest operacjonalizacja pozycji jednostek w strukturze społecznej, zaprezentowana w wydanej w 2007 roku nakładem Wydawnictwa Instytutu Filozofii i Socjologii PAN publikacji H. Domańskiego, Z. Sawińskiego oraz Kazimierza M. Słomczyńskiego *Nowe klasyfikacje i skale zawodów*. Kodowanie może przybierać różne formy, począwszy od „zdecentralizowanego" lub „scentralizowanego"[58] kodowania ręcznego, poprzez działania wspomagane komputerowo, a kończąc na automatycznym kodowaniu tekstu z ręcznym uzupełnianiem nietypowych przypadków (por. Biemer i in. 2003: 235; Weisberg 2005: 265–266).

W świetle zagadnień związanych z estymacją miary całkowitego błędu pomiaru badań sondażowych niezwykle istotną konsekwencją błędu kodowania jest to, że jego wielkość może być skorelowana z osobą kodera (zupełnie tak samo, jak wyniki wywiadu mogą być skorelowane z osoba ankietera). Dzieje się tak na skutek tego, iż osoby przeprowadzające kodowanie odpowiedzi respondentów mogą w różny sposób interpretować opracowane klucze kodowe (por. Brill 2008: 101–102; Weisberg 2005: 265; Funkhouser i in. 1968: 122–128), czego efektem jest przyrost wariancji w zbiorze wyników pomiaru (por. Biemer

---

[57] Na te dwa warunki, jakie spełniać musi klucz kodowy, tj. (1) rozłączność oraz (2) zupełność, uwagę zwrócił między innymi H. Weisberg (2005: 265), omawiając zasady opracowywania schematów kodowania. Pierwszy z nich oznacza, że każdej odpowiedzi respondenta daje się jednoznacznie przypisać jeden (i tylko jeden) kod. Biemer i in. (2003: 236) zwracają jednak dodatkowo uwagę na pewne problemy wynikające z konieczności spełnienia tego warunku: w praktyce odpowiedź respondenta może być nieścisła, co w konsekwencji przełoży się na przyporządkowanie jednego z wielu potencjalnie „pasujących" kodów, w zależności od przyjętej interpretacji odpowiedzi respondenta. Kryterium zupełności oznacza natomiast, że zdefiniowane kody da się przyporządkować wszystkim (bez wyjątku) odpowiedziom respondentów. Kryterium to spełniane jest zazwyczaj poprzez dołączenie kodu „inne", na oznaczenie tych wszystkich odpowiedzi, którym nie da się przypisać – używając terminologii Weisberga – „kodów substantywnych" (por. Weisberg 2005: 265).

[58] Rozróżnienie na te dwie formy kodowania ręcznego odnajdujemy w pracy Biemera i in. (2003: 235–236). Kodowanie „scentralizowane" przeprowadzane jest wewnątrz instytucji sondażowej przez mniej lub bardziej wyspecjalizowanych w tym zakresie pracowników, z kolei kodowanie „zdecentralizowane" przeprowadzane jest przez ankieterów lub samych respondentów w trakcie realizacji wywiadu.

2010a: 45–46). Statystyką mierzącą wpływ tego zjawiska jest miara $DEFF_{KOD}$, oparta na współczynniku korelacji wewnątrzklasowej (por. Kish 1965: 161–164; 170–172), która definiowana jest według podobnych zasad, co przedstawiona w sekcji II.2.3. miara $DEFF_{ANK}$. Tym samym, miarę $DEFF_{KOD}$ można zapisać jako:

(II.25.) $$DEFF_{KOD} \stackrel{\text{def}}{=} 1 + \left(\bar{k} - 1\right)\rho_{KOD},$$

lub opcjonalnie (uwzględniając dodatkowo wskaźnik rzetelności kodowania oznaczony symbolem *r*) w postaci przedstawionej przez Grovesa i in. (2004: 317) oraz Weisberga (2005: 265, 343):

(II.25'.) $$DEFF_{KOD} \stackrel{\text{def}}{=} 1 + \left(\bar{k} - 1\right)(1 - r)\rho_{KOD},$$

gdzie:

- $\bar{k}$ jest średnią ważoną liczbą odpowiedzi przekazanych do kodowania każdej z K osób, przy czym $\bar{k} = \frac{1}{K-1}\left(N_{pomiar} - \sum_{k=1}^{K} \frac{n_k^2}{N_{pomiar}}\right)$, N_{pomiar} oznacza liczebność próby zrealizowanej a n_k, gdzie $n_1 + n_2 + \ldots + n_K = N_{pomiar}$, jest liczbą pytań przydzieloną do każdego (k = 1, 2, …, K) kodera;
- ρ_{KOD} jest współczynnikiem korelacji wewnątrzklasowej, którego oszacowaniem może być wielkość $\hat{\rho}_{KOD} = \frac{MSB - MSW}{MSB + (\bar{k}-1)MSW}$, oparta na analizie wariancji (analogicznie do współczynnika $\hat{\rho}_{ANK}$ zdefiniowanego w poprzedniej sekcji tego rozdziału), przy czym *MSB* oznacza średni kwadrat zróżnicowania międzyklasowego (tj. między koderami), a *MSW* odpowiada średniemu kwadratowi zróżnicowania wewnątrzklasowego (tj. w obrębie kodów przypisanych przez daną osobę).

Miara $DEFF_{KOD}$ podlega tym samym ograniczeniom co miernik $DEFF_{ANK}$, z czego bodaj najważniejszym jest to, że każda z obserwacji musi być przypisana do poszczególnych osób kodujących odpowiedzi respondentów w sposób losowy. Spełnienie tego warunku sprawia, że miara $DEFF_{KOD}$ jest w istocie miarą oddziaływania koderów na uzyskane wyniki, bowiem przy zrandomizowanym przydziale obserwacji należy się spodziewać, że zaobserwowane dysproporcje wynikać będą z różnic w interpretacji kluczy kodowych, a nie na przykład ze specyfiki respondentów, których odpowiedzi przypisano poszczególnym koderom (por. Weisberg 2005: 55).

Jedną z bardziej skutecznych metod, umożliwiających wyeliminowanie błędu skorelowania wyników kodowania z osobą kodera, jest multiplikowanie liczby osób zaangażowanych w prace nad przetwarzaniem wypowiedzi respondentów. Wiąże się to z przyjęciem zasady opracowywania tych samych przypadków (wypowiedzi) przez co najmniej dwóch koderów, co pociąga za sobą wzrost kosztów oraz wydłużenie czasu potrzebnego na opracowanie wy-

ników, jednakże, jeżeli kodowanie odbywa się w sposób niezależny[59], to możliwe staje się określenie rzetelności (ang. *intercoder reliability*) oraz niezgodności[60] (ang. *coder disagreement rate*) kodowania, czyli miar świadczących o homogeniczności przeprowadzonej operacji. Pierwszy wskaźnik (element składowy miary $DEFF_{KOD}$ oznaczony we wzorze II.25'. jako *r*) liczony jest jako współczynnik korelacji (odpowiednio dla poziomu pomiaru zmiennej) pomiędzy kodami przyporządkowywanymi przez pary koderów, drugi natomiast podaje odsetek przypadków, w których pary koderów przypisały odmienne wartości do tej samej wypowiedzi respondenta. Warto jednak podkreślić, że choć wysoka wartość współczynnika rzetelności kodowania oraz niska wartość wskaźnika niepodobieństwa świadczy o spójności przeprowadzonego działania, to nadal nie jest dowodem na brak popełnionego w tym zakresie błędu (por. Weisberg 2005: 265). Innymi słowy, poszczególne przyporządkowania wypowiedzi respondentów do odpowiednich kategorii kodowych przez kolejnych koderów mogą być zgodne, ale nadal pozostawać przypisane niewłaściwe, czyli (przyjmując opisową definicję błędu kodowania) niezgodnie z rzeczywistością oraz intencjami badacza. Mocno brzmią zatem słowa P. Biemera oraz L. Lyberga, mówiące o tym, że „jeżeli kodowanie pozostaje poza kontrolą [badacza – P.J.], to współczynniki poziomu błędów są wysokie, co z kolei może prowadzić do wzrostu [całkowitego – P.J.] błędu pomiaru" (Biemer i in. 2003: 238). Autorzy ci przedstawiają przy tym ciekawą odmianę uproszczonej weryfikacji kodowania, opierającej się na idei kontroli dwustopniowej z tzw. głosem rozstrzygającym, która wydaje się efektywnym narzędziem ograniczającym błąd kodowania[61].

---

[59] Niezależny sposób kodowania oznacza, że pary koderów przyporządkowują kody nie znając wzajemnie przyporządkowywanych przez siebie wartości. Biemer i in. (2003: 238–239) omawiają także przykład kodowania zależnego, które polega na tym, że przypisane przez pierwszego kodera kody przekazywane są do drugiego kodera, który, znając przyporządkowany kod, nadaje własny i weryfikuje zgodność kodów, a w przypadku niezgodności decyduje o kodzie wyjściowym. P. Biemer i L. Lyberg ukazują jednak, że taki schemat kontroli procedury kodowania, nazywanej uproszczoną weryfikacją kodowania, nie jest efektywny; winę za to ponosi prosty mechanizm, który sprawia, że na ocenę „weryfikatora" znaczny wpływ wywiera kod przyporządkowany już wcześniej. Co więcej, procedura ta umożliwia jedynie wychwycenie oczywistych i jednoznacznych odstępstw od instrukcji kodowych, te mniej istotne pozostaną niepoprawione.

[60] Wskaźnik niezgodności kodowania (oznaczony symbolem *CD*) wyznaczyć można, stosując formułę określoną wzorem $CD = \left(1 - \frac{\sum_{i=1}^{N_{resp}} c_i}{N_{resp}}\right) 100\%$, gdzie: $\overset{\forall}{i \in \{1, 2, \ldots, N_{resp}\}} c_i = \begin{cases} 1; \text{dla } x_i^{\mathrm{A}} = x_i^{\mathrm{B}} \\ 0; \text{dla } x_i^{\mathrm{A}} \neq x_i^{\mathrm{B}} \end{cases}$, natomiast x_i^{A} oraz x_i^{B} oznaczają kody przyporządkowane wypowiedzi *i*-tego respondenta przez osoby A i B.

[61] Polega ona na tym, że osoby A oraz B kodują odpowiedzi niezależnie od siebie, przyporządkowując wszystkie obserwacje – wypowiedzi respondentów – do odpowiednich kategorii zmiennej wynikowej. Jeżeli pary kodów są zgodne, to przyjmuje się, że ustalono kod wyjściowy. W sytuacji, gdy osoby A oraz B przypisują odmienne kody, trzeci koder podaje, niezależnie od A i B, swoją propozycję kodu. Jeśli okaże się, że koder C przypisał kod równy A lub B, to obowiązującą warto-

Wydaje się rzeczą oczywistą, że nie tylko kodowanie pytań otwartych, ale także inne działania wykonywane w ramach postsurveyowej obróbki danych wymagają od badaczy wzmożonej kontroli jakości przeprowadzanych operacji. Chodzi przede wszystkim o to, że błędy przetwarzania danych mają swoje główne źródło w osobie badacza (lub w personelu badawczym), a zatem zachowanie wysokich standardów edycji baz danych może okazać się jedną ze skuteczniejszych metod ograniczających ryzyko pojawienia się błędów przetwarzania danych. Doskonale ujęli to autorzy monografii *Introduction to Survey Quality*, którzy w rozdziale siódmym, w części poświęconej naturze błędów procedowania danych, stwierdzają, iż „bardziej efektywnym podejściem jest […] lepszy nadzór, szkolenia prewencyjne oraz wypracowanie procedur umożliwiających redukcję lub wyeliminowanie [błędów procedowania – P.J.] niż estymowanie ich wielkości" (Biemer i in. 2003: 219). Podobnie, jednym z ważniejszych postulatów sformułowanych przez H. Weisberga (2005: 261–263) jest wniosek o podjęcie dyskusji nad wypracowaniem uniwersalnych standardów edycji baz danych, podobnych do tych, jakie wypracowano w zakresie losowania prób badawczych, konstruowania narzędzi pomiarowych, prowadzenia wywiadów, kontroli pracy ankieterów itd.. Wszystko to zaś służyć ma poprawie jakości procesu przetwarzania danych, a w kontekście paradygmatu całkowitego błędu pomiaru – przyczynić się do ograniczenia jego wielkości.

II.4. Uwagi końcowe

Wprawdzie przedstawione w tym rozdziale zagadnienia wykraczają w części dotyczącej pomiaru oraz procedowania danych poza tematykę pracy (skoncentrowanej wokół problemów reprezentatywności sondażowych prób badawczych), to jednak pozostają istotnym dopełnieniem zarysowanego w rozdziale pierwszym paradygmatu całkowitego błędu pomiaru. Koncepcja ta porządkuje i w znacznym stopniu wyznacza podejmowane dalej zagadnienia. Zresztą nietrudno zauważyć, że opisane w tym rozdziale definicje błędów (systematycznych oraz losowych) pojawiających się na kolejnych etapach realizacji badań – począwszy od doboru i realizacji próby badawczej, poprzez pomiar, a na postsurveyowej obróbce danych kończąc – są ze sobą jednoznacznie po-

---

ścią jest ta przypisana przez osobę C. W przeciwnym przypadku czwarty koder (najlepiej sam badacz) rozstrzyga ostatecznie o tym, jaki powinien być kod wyjściowy, jednak nie musi być nim wcale żaden z kodów przyporządkowany wcześniej przez osobę A, B, czy też C. Zaprezentowany schemat weryfikacji kodowania posiada w opinii Biemera i Lyberga zdecydowanie wyższą efektywność od kodowania zależnego, tzn. pozwala wykryć większą liczbę błędnie podjętych decyzji, a także ma podobną efektywność w porównaniu z innymi bardziej złożonymi i kosztownymi schematami weryfikacyjnymi (por. Biemer i in. 2003: 239–241).

wiązane i chociaż nie opisują wszystkich możliwych źródeł błędów badań sondażowych, to jednak charakteryzują te najbardziej znaczące.

W kolejnych rozdziałach monografii uwaga skoncentrowana będzie na zagadnieniach reprezentatywności sondażowych prób badawczych. Nie zabraknie jednak odniesień do procesu zbierania danych oraz do ich przetwarzania. Podjęta zostanie problematyka związana z wykorzystaniem schematów doboru prób badawczych, ułomnościami operatów losowania, a także z uciążliwym dla badaczy problemem jednostek niedostępnych oraz z procedurami ważenia danych. Charakterystyka najważniejszych ustaleń teoretyczno-metodologicznych w tym zakresie zobrazowana zostanie autorskimi analizami empirycznymi. Szczególnie pouczające okażą się dane metodologiczne Europejskiego Sondażu Społecznego, zawierające szczegółowe informacje o komplikacjach związanych z realizacją jednego z najbardziej wyrafinowanych metodologicznie projektów badawczych.

ROZDZIAŁ III

Losowanie jednostek – operaty doboru próby

Losowanie próby może wydawać się przedsięwzięciem stosunkowo mało problematycznym, zwłaszcza gdy spojrzy się na skalę trudności, przed którymi stają badacze w trakcie realizacji badań terenowych, jednakże to właśnie na etapie wyboru jednostek do próby rozgrywa się już często batalia o jakość całego przedsięwzięcia badawczego. Wysiłki ukierunkowane na wylosowanie z populacji możliwie najbardziej reprezentatywnego zbioru jednostek są w znacznej mierze zdeterminowane przez dostępne oraz możliwe do zastosowania rejestry, tj. spisy jednostek lub zespołów jednostek, służące doborowi elementów z populacji do próby badawczej. To właśnie w wyniku ułomności operatów doboru próby (lub po prostu wskutek braku dostępu do odpowiednich rejestrów), badacze rezygnują z doboru prostego na rzecz innych – mniej efektywnych – schematów losowania jednostek.

Oczywiście możliwość zastosowania „operatu idealnego" – czyli takiego, w którym każda jednostka z populacji pojawia się jako osobny element tylko i wyłącznie jeden raz oraz żadne inne „obce" jednostki spoza populacji nie wchodzą w skład operatu (por. Kish 1965: 53) – nie daje jeszcze gwarancji ani wylosowania, ani, tym bardziej, zrealizowania próby spełniającej definicyjne kryteria losowania prostego. Po pierwsze, decyzja o zastosowaniu jakiegoś specyficznego schematu doboru respondentów z takiego idealnego operatu (na przykład losowania zespołowego, wielostopniowego lub też warstwowego) może w sposób zupełnie oczywisty zróżnicować szanse selekcji jednostek do próby, prowadząc tym samym do odrzucenia fundamentalnego założenia definicyjnego prostej próby losowej, ujętego w postaci kryterium równości prawdopodobieństw selekcji. Po drugie, nawet jeśli wylosowana reprezentacja populacji spełniać będzie warunki losowania prostego, to terenowa jej realizacja już takich wymogów nie wypełni (głównie z uwagi na pojawienie się jednostek niedostępnych). W tym kontekście niezwykle słuszna wydaje się uwaga autorów monografii *Introduction to Survey Quality* z wprowadzenia do rozdziału III, ukazująca, że:

błąd operatu doboru próby oraz błąd braku odpowiedzi oddziałują w bardzo podobny sposób na błąd średniokwadratowy [tj. całkowity błąd pomiaru – P.J.]. Niektórzy mogliby nawet uznać, że jednostki pominięte w operatach losowania są jednym z typów nie–respondentów, gdyż w obu przypadkach informacje o jednostkach nie są znane. Pomimo to metody ograniczania konsekwencji tych dwóch źródeł błędów są zupełnie odmienne. (Biemer i in. 2003: 63)

Warto zaznaczyć, że owa w gruncie rzeczy fatalistyczna perspektywa związana z brakiem możliwości pełnej realizacji prób badawczych (które nawet dobrane w sposób przystający do kryteriów losowania prostego i tak nie zapewnią równie reprezentatywnych wyników pomiaru) nie powinna jednak, mimo wszystko, w żaden sposób skłaniać do rezygnacji z wysiłków na rzecz konstruowania oraz wykorzystywania operatów o parametrach zbliżonych do tych, które – używając terminologii L. Kisha – można nazwać idealnymi.

III.1. Typologia populacji – rozróżnienia pojęciowe

Zanim przeanalizowane zostaną zagadnienia związane z operatami losowania, warto odnieść się jeszcze raz do ustaleń poczynionych w drugim rozdziale monografii. Zdefiniowano już tam pojęcie operatu oraz scharakteryzowano błędy, jakie w związku z niedoskonałościami operatów mogą pojawić się w badaniach sondażowych. Przypomnieć można zatem, że operat losowania określony został jako dostępny rejestr jednostek (na przykład imienny wykaz osób) lub zespół jednostek (na przykład wykaz gospodarstw domowych, baza punktów adresowych, listy numerów telefonicznych abonentów telefonii stacjonarnej lub mobilnej, itp.) należących do badanej populacji, natomiast błąd operatu zdefiniowano jako efekt rozbieżności (niepełne pokrycie, nadmiarowe pokrycie, multiplikowanie jednostek, informacje zespołowe) występujący pomiędzy zbiorem wszystkich elementów rzeczywiście należących do populacji będącej przedmiotem zainteresowania badacza, a wykorzystywanym do losowania próby dostępnym wykazem tych jednostek. Chociaż wyprowadzone wówczas rozróżnienia terminologiczne były wystarczające dla zdefiniowania błędu pokrycia populacji operatem losowania, to jednak wymagają doprecyzowania w toku prowadzonych obecnie rozważań. Istotnie bowiem, jeżeli spojrzy się na dostępny rejestr jednostek jako na operacyjną definicję populacji (por. Henry 1990: 50), a na stopień, w którym ten operat przystaje do populacji, jako na wskaźnik jakości owej operacjonalizacji, to w pierwszej kolejności należy odpowiedzieć na pytanie, do czego odnosi się pojęcie operatu losowania lub, inaczej, co ma się na myśli, mówiąc o populacji w kontekście błędów operatu.

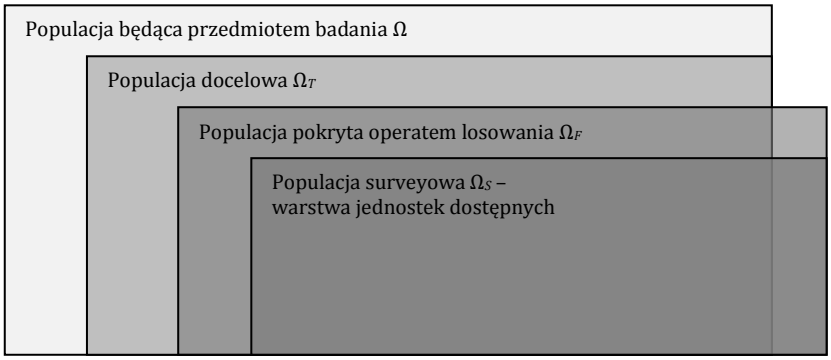
Niezwykle pomocne okazują się dystynkcje wyprowadzone przez L. Kisha w artykule *Populations for Survey Sampling,* zamieszczonym w 1979 roku

w periodyku wydawanym przez International Association of Survey Statisticians (por. Kish 1979: 14–15). Zostały one później wykorzystane przez Grovesa (1989: 82–83) do wyróżnienia czterech typów populacji: (a) będącą przedmiotem badania (*inferential population*), oznaczaną jako Ω, (b) docelową (*target population*), oznaczaną symbolem Ω_T, (c) pokrytą operatem losowania (*frame population*), oznaczaną dalej jako Ω_F oraz (d) surveyową (*survey population*), oznaczaną symbolem Ω_S. Rozróżnienie to nie jest wprawdzie szczególnie systematycznie i konsekwentnie wykorzystywane przez metodologów badań sondażowych[62], pozwala jednak oddzielić problemy pojawiające się w trakcie losowania próby (zależności pomiędzy populacjami Ω_T oraz Ω_F) od tych, które napotykane są w trakcie jej realizacji (wzajemne odniesienie populacji Ω_F i Ω_S).

Krótką charakterystykę każdego z tych pojęć warto poprzedzić uwagą, że w przypadku badań sondażowych ma się w większości wypadków do czynienia z populacjami o skończonej liczbie elementów. Ponadto, w rozumieniu teorii zbiorów będzie tak, że $\Omega_T \subseteq \Omega$ oraz $\Omega_S \subseteq \Omega_F$, ale już nie zawsze $\Omega_F \subseteq \Omega_T$. Jest to oczywistą konsekwencja tego, iż operaty losowania zawierają niejednokrotnie nadmiarowe jednostki, co oznacza, że zbiór elementów należących do Ω_F nie musi być podzbiorem Ω_T. Parafrazując rozważania Stefana Nowaka (2007: 177–179) dotyczące zakresów wskaźników i *indicatów*, można o operatach

---

[62] W drugim tomie *Encyclopedia of Survey Research Methods* James M. Lepkowski (2008) przeprowadza analizę sposobów użycia w badaniach sondażowych terminu *populacja*, dochodząc do konkluzji, iż: „definicje populacji nie są wystandaryzowane w obszarze [badań surveyowych – P.J.]. Niektórzy autorzy używają różnej terminologii na zdefiniowanie [tych samych typów populacji – P.J.]" (Lepkowski 2008: 591). Szczególne kontrowersje budzić może przy tym pojęcie *populacji surveyowej*, która w ujęciu zaprezentowanym na schemacie III.1. oznacza „zbiór osób, które, jeśli byłyby wylosowane do próby badawczej, byłyby jednocześnie respondentami [wzięłyby udział w badaniu – P.J.]" (Groves 1989: 83). Taki sposób definiowania populacji surveyowej zakłada, że każda jednostka jest przypisana do warstwy respondentów lub jednostek niedostępnych. Oczywiście przynależność jednostek populacji do tych rozłącznych kategorii nie jest znana na etapie losowania próby, badacz doświadcza jedynie skutków doboru jednostek z populacji lub spoza populacji surveyowej, co przejawia się występowaniem osób dostępnych oraz niedostępnych. Innymi słowy, jak wskazuje Groves (1989: 83), rozbieżności pomiędzy populacją surveyową oraz populacją pokrytą operatem losowania są niczym innym jak warstwą jednostek niedostępnych (*non–respondents*). Jest to o tyle problematyczne, że skłonność do udziału w badaniu nie jest zmienną dychotomiczną, a raczej pozostaje zależna nie tylko od cech osobowościowych jednostek, ale także od specyfiki techniki badawczej wykorzystanej do zbierania danych. Nie powinno zatem dziwić, że w wielu opracowaniach pojęcie populacji surveyowej używane jest w zupełnie innym znaczeniu. Wystarczy przywołać artykuł autorstwa Geerta Loosveldta oraz Nathalie Sonck z 2008 roku pt. *An Evaluation of the Weighting Procedures for an Online Access Panel Survey*, w którym pojęcie populacji surveyowej używane jest szerzej od pojęcia populacji pokrytej operatem losowania. We fragmencie dotyczącym rozważań nad błędem pokrycia populacji generalnej operatami użytkowników Internetu odnaleźć można następujące stwierdzenie „błąd pokrycia występuje wtedy, gdy nie wszystkie elementy z populacji surveyowej posiadają znaną i niezerową szansę wylosowania do próby" (Loosveldt i in. 2008: 94). W tym kontekście populacja surveyowa jest tym samym, czym u L. Kisha była populacja docelowa. Podobnie pojęcie populacji surveyowej definiują S. Dorofeev oraz P. Grant (2006: 10), stosując je zamiennie z określeniem „populacja docelowa".



Ryc. III.1. Typy populacji – rozróżnienia pojęciowe

Źródło: opracowanie własne

doboru próby powiedzieć, że mają doskonałą moc zawierania jednostek należących do badanej populacji, jeżeli w ich zakres wchodzi zbiór wszystkich elementów przynależnych do populacji, niezależnie jednak od tego, ile jednocześnie zawartych jest elementów do populacji nienależących. Podobnie można powiedzieć, iż operaty mają doskonałą moc odrzucania elementów nienależących do populacji docelowej, jeżeli w ich zakres nie wejdą żadne elementy spoza populacji, niezależnie od tego, ile jednocześnie elementów populacji będzie w operacie pominiętych. Kontynuując analogię zapożyczoną z teorii pomiaru, można wykorzystać również pojęcie idealnej mocy rozdzielczej do określenia sytuacji, w której w zakres operatu wejdą wszystkie elementy z populacji oraz nic poza tym. Analogia ta jest o tyle uzasadniona, że zdefiniowana przez Nowaka (2007: 178) miara mocy zawierania odpowiada znanemu z metodologii badań sondażowych wskaźnikowi pokrycia populacji operatem losowania, natomiast dopełnienie miary mocy odrzucania (por. Nowak 2007: 178) jest równoważne wskaźnikowi nadmiarowego pokrycia. Z kolei operat charakteryzujący się idealną mocą rozdzielczą (por. Nowak 2007: 178) można uznać (z założeniem równoważności poziomu zagregowania jednostek w próbie i w operacie) za operat „idealny".

Pierwszy z wyróżnionych typów populacji związany jest z przedmiotem badania. Ponieważ celem każdego badania o charakterze reprezentatywnym jest wyciąganie wniosków nie tyle o przebadanej próbie, ile o całej zbiorowości, to w sposób zupełnie oczywisty wymaga się określenia tego, kto (lub co) na tę badaną zbiorowość się składa. G. Lissowski i in. (2008: 23), charakteryzując przedmiot badania statystycznego, zwracają uwagę na fakt, że populację można określić na dwa sposoby. Po pierwsze – w sposób jawny (konstruktywny), tj. poprzez podanie wykazu wszystkich jednostek statystycznych, które na

populację się składają[63]. Po drugie natomiast – w sposób niejawny (deskryptywny), tj. poprzez wskazanie zestawu pewnych wspólnych cech jednostek należących do badanej populacji[64]. W notach metodologicznych większości badań sondażowych populacja określana jest w sposób deskryptywny. Dla przykładu, w badaniach Europejskiego Sondażu Społecznego „populację stanowią osoby w wieku 15 lat lub starsze, niezależnie od ich narodowości oraz obywatelstwa, języka lub statusu prawnego" (*ESS2 Sampling Report* 2004: 3). Z kolei w studiach Diagnozy Społecznej przedmiotem badania jest populacja wszystkich – bez wyjątku – gospodarstw domowych w Polsce[65]. Podobnie można powiedzieć, że przedmiotem badań preferencji politycznych jest populacja dorosłych mieszańców Polski posiadających czynne prawo wyborcze. Przy określaniu populacji będącej przedmiotem badania konieczne jest zatem wskazanie cech pozwalających jednoznacznie określić, które elementy należą, a które nie należą do badanej populacji. Poza opisem cech jednostek wymagane jest również ustalenie ram przestrzennych oraz czasowych, do jakich odnoszona będzie populacja będąca przedmiotem badania.

Populacja docelowa pozostaje z kolei zbiorem osób, które w rzeczywistości mają być przedmiotem wnioskowania prowadzonego w oparciu o pomiar próby (por. Cox 2008: 875–876; Groves i in. 2004: 67). W wielu sytuacjach nie istnieje żadna różnica (lub inaczej mówiąc, nie powinna istnieć różnica) pomiędzy populacją docelową a populacją będącą przedmiotem badania. Wnioskowanie powinno bowiem obejmować te kategorie jednostek należących do populacji, których uwzględnienie jest wymagane z uwagi na cel badania oraz związane z tym celem problemy badawcze. Jeśli zatem istnieje jakaś różnica pomiędzy oboma typami populacji, to raczej ze względu na uzasadnione merytorycznie względy praktyczne. Typowym przykładem takiego podejścia jest wyłączenie z zakresu populacji takich osób, z którymi realizacja badań byłaby niezwykle utrudniona lub po prostu niemożliwa. Dla przykładu, w badaniach Europejskiego Sondażu Społecznego w większości krajów wyklucza się z populacji osoby bezdomne oraz przebywające w miejscach zbiorowego zakwaterowania (na przykład w szpitalach, więzieniach, wojsku, klasztorach itp.). W innych natomiast wyłącza się te kategorie populacji, z którymi realizacja badań byłaby pro-

---

[63] Jeżeli populacja składa się ze skończonej liczby elementów, to w sposób jawny określić można ją jako zbiór $\Omega = \{\varpi_1, \varpi_2, \dots, \varpi_N\}$.

[64] Jeżeli populację określamy poprzez wskazanie zestawu k cech (oznaczanych jako α), wśród których będą charakterystyki przestrzenne oraz czasowe, to populację można zapisać jako zbiór tych wszystkich elementów, które posiadają jednocześnie wszystkie określone przez badacza cechy, to znaczy $\Omega = \{\varpi: \alpha_1(\varpi) \wedge \alpha_2(\varpi) \wedge \dots \wedge \alpha_k(\varpi)\}$.

[65] Wprawdzie populacji badawczej nie zdefiniowano wprost, jednak w nocie metodologicznej raportu *Diagnoza Społeczna 2011 Warunki i Jakość Życia Polaków* można odnaleźć informację, że: „badaniu podlegały gospodarstwa jednoosobowe oraz wieloosobowe" (Panek i in. 2011: 35–44).

blematyczna z uwagi na sytuację polityczną danego kraju. Wystarczy przywołać badania ESS realizowane w Izraelu, w których poza zakresem wnioskowania znajduje się mniejszość palestyńska we Wschodniej Jerozolimie oraz na Zachodnim Brzegu Jordanu (por. na przykład *ESS4 – 2008 Documentation Report*: 163). Decyzja o wyłączeniu z badanej populacji pewnej kategorii osób może być też podyktowana względami finansowymi. Typowym tego przykładem jest grecka część projektu ESS, w której poza populacją badawczą znaleźli się mieszkańcy słabo zaludnionych wysp archipelagu Cyklad oraz Dodekanez (por. *ESS1 Sampling Report 2002*: 18), czy też hiszpańska odsłona pierwszej edycji ESS, w której z populacji badawczej wyłączono mieszkańców dwóch afrykańskich miast-enklaw: Ceuty oraz Melilli (por. *ESS1 Sampling report 2002*: 34). Taka w gruncie rzeczy arbitralna decyzja badaczy o nieuwzględnianiu pewnych kategorii osób należących do populacji zawęża w sposób zupełnie oczywisty zakres indukcji statystycznej, ograniczając możliwość uogólniania wniosków z próby badawczej wyłącznie do zakresu populacji docelowej. Warto jednak zauważyć, że w wielu sytuacjach liczebności wykluczonych kategorii są na tyle niewielkie, w porównaniu z liczebnością całej populacji, że skutki takich decyzji nie będą – mimo wszystko – przekładać się w żaden istotny sposób na możliwość wnioskowania o całej populacji, której prowadzone badanie powinno dotyczyć.

Trzeci ze wskazanych typów populacji związany jest z operatami doboru próby. O ile populację będącą przedmiotem badania oraz populację docelową można zdefiniować w sposób deskryptywny, tj. poprzez podanie cech, jakie jednostki zakwalifikowane do populacji mają posiadać, to w przypadku populacji pokrytej operatem losowania wymaga się już jej określenia w sposób konstruktywny, to znaczy poprzez wskazanie konkretnych elementów, które do populacji tej należą. Innymi słowy, populacja pokryta operatem losowania jest wykazem jednostek (na przykład osób) lub rejestrem zespołów jednostek grupujących jednostki indywidualne wchodzące w skład populacji docelowej (na przykład adresów budynków, danych gospodarstw domowych), które wykorzystuje się do losowania prób badawczych. W tym ujęciu populacja objęta operatem losowania utożsamiana jest po prostu z operatem doboru próby badawczej. Warto wskazać, że informacje zawarte w takich rejestrach pozwalają nie tylko na identyfikację jednostek (lub ich zespołów), ale zawierają też dodatkowe dane przydatne przy projektowaniu oraz wyborze określonych schematów doboru prób badawczych. Dla przykładu, rejestr PESEL zawiera informacje umożliwiające warstwowanie próby według miejsca zamieszkania, płci, wieku, a nawet stanu cywilnego. Z kolei baza TERYT, której depozytariuszem jest GUS, zawiera informacje o adresach jednostek mieszkalnych, pozwalające na podział próby względem struktury terytorialnej kraju. Relacja pomiędzy populacją do-

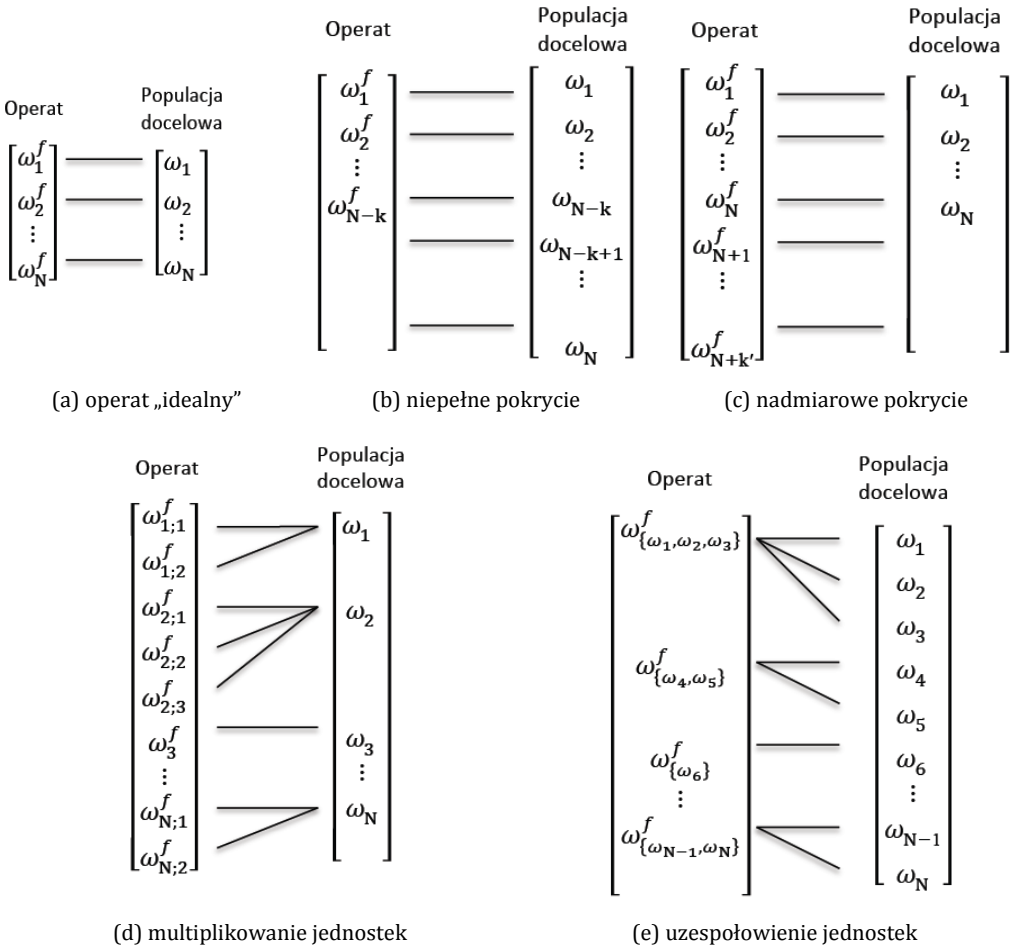
celową a tą pokrytą przez operat losowania jest zatem taka, że ta pierwsza określa, kto do populacji należy oraz jakie cechy wyróżniają włączone do niej jednostki, druga natomiast jest rejestrem jednostek wykorzystywanym w losowaniu konkretnej próby badawczej.

III.2. Błędy operatów doboru prób badawczych – uszczegółowienie problemu

W drugim rozdziale monografii zdefiniowane zostały już błędy badań sondażowych mające swoje źródło w uchybieniach operatów doboru prób badawczych, opisano też cztery rodzaje ułomności charakteryzujące operaty. Wskazywano przy tym na takie problemy jak: (a) niepełne pokrycie populacji docelowej, (b) nadmiarowe pokrycie jednostkami, które do populacji nie należą, (c) multiplikowanie informacji o jednostkach, a także (d) zespołowe grupowanie wielu jednostek indywidualnych. W każdej z tych czterech sytuacji populacja pokryta operatem losowania nie przystaje do populacji docelowej.

Zaprezentowana typologia błędów – podawana przez wielu autorów za L. Kishem (1965: 53–59) – nie wyczerpuje oczywiście wszystkich komplikacji, przed którymi stają badacze przystępujący do losowania prób sondażowych. W wielu przypadkach pojawiają się typy mieszane, kumulujące kilka problemów jednocześnie. Dobrym przykładem takich „kłopotliwych" operatów są rejestry punktów adresowych (tj. budynków, mieszkań lub jednostek terytorialnych), w których ma się do czynienia zarówno z niepełnym pokryciem (na przykład z brakiem informacji o nowych budynkach), pokryciem nadmiarowym (na przykład z wykazem obiektów zbiorowego zakwaterowania, adresów instytucji lub innych adresów niemieszkalnych), jak i z wielokrotnym uzespołowieniem jednostek losowania[66]. Kolejnym przykładem operatu charakteryzującego się wieloma typami błędów jednocześnie są rejestry wykorzystywane w badaniach realizowanych techniką wywiadów telefonicznych z losowo dobieranymi/generowanymi numerami abonentów (por. na przykład Brick 2008: 675–678). W takiej sytuacji mogą wystąpić wszystkie typy błędów operatów. Po pierwsze, wygenerowanie numeru telefonicznego daje zerowe szanse doboru do próby tym jednostkom, które nie posiadają telefonu komórkowego, a ich gospodarstwo domowe nie ma dostępu do telefonii stacjonarnej. Po drugie, operat taki daje szanse wylosowania osobom spoza populacji, np. obcokrajow-

[66] Dla przykładu, jeżeli populacja docelowa składa się z osób, to wylosowanie adresu budynku mieszkalnego grupuje zarówno gospodarstwa domowe, prowadząc do niedoreprezentacji mieszkań z dużych bloków, jak również (w obrębie gospodarstwa domowego) osoby, skutkując niedoreprezentowaniem jednostek z wieloosobowych gospodarstw domowych.



Ryc. III.2. Populacja docelowa *vs*. operat losowania

Źródło: opracowanie własne

com, lub też umożliwia dobór numerów firm, instytucji itd., czyli wylosowanie takich obiektów, które stanowią przedmiot pomiaru wyłącznie w pewnych specyficznych sondażach. Po trzecie, osoby korzystające z więcej niż jednego numeru mobilnego lub posiadające jednocześnie dostęp do telefonu stacjonarnego oraz telefonu komórkowego mają większe prawdopodobieństwa selekcji do próby badawczej od tych osób, z którymi kontakt możliwy jest wyłącznie poprzez jeden telefon stacjonarny lub komórkowy. Po czwarte wreszcie, zastosowanie operatu tego typu rodzi problemy związane z doborem zespołowym. Wystarczy wskazać, że w sytuacji wylosowania numeru stacjonarnego przypisanego do określonego gospodarstwa domowego zachodzi zazwyczaj koniecz-

ność wyboru konkretnego respondenta. W takich przypadkach szansa wylosowania jednostki populacji uwarunkowana jest już od wielkości gospodarstwa domowego.

Przedstawione przykłady mogły wzbudzić pewien niepokój oraz pozostawić wrażenie raczej pesymistyczne. Warto jednak zauważyć, że chociaż „operaty idealne" wydają się konstruktem czysto teoretycznym, to jednak w przypadku badań sondażowych realizowanych na próbach generalnych ludności dysponuje się dość często operatami zbliżonymi parametrami jakościowymi do typu idealnego. W Polsce rolę takiego rejestru – przynajmniej do końca 2016 roku – pełni PESEL[67], zawierający aktualizowane na bieżąco informacje o wszystkich osobach zameldowanych na pobyt stały lub czasowy na terytorium Rzeczypospolitej Polskiej. Wśród krajów biorących udział w badaniach ESS dane jednostkowe dostępne są też między innymi w Belgii, Danii, Estonii, Finlandii, Islandii, Niemczech oraz Norwegii, i pokrywają w każdym kraju jeśli nie całą, to prawie całą populację docelową. Po wykorzystaniu takich operatów doboru prób okazuje się najczęściej, że wpływ uchybień operatów na wielkość całkowitego błędu pomiaru badań sondażowych jest zupełnie marginalny[68]. Innymi słowy, badacz staje przed problemami wynikającymi z niedoskonałości operatów losowania najczęściej wtedy, gdy rejestry indywidualne są niedostępne, niepełne, niewiarygodne lub gdy rejestrów takich nie da się wykorzystać w ramach wybranej techniki badawczej.

Charakteryzując błędy operatów doboru prób badawczych, warto również zaznaczyć, że najpoważniejsze konsekwencje niesie ze sobą niepełne pokrycie populacji docelowej. W pierwszym oraz drugim rozdziale pracy wskazywano już, że wynika to ze specyfiki błędu pokrycia, który oddziałuje na systematyczny komponent całkowitego błędu pomiaru (może zatem skutkować poważnym wypaczeniem wartości estymatorów pewnych parametrów populacyjnych). Podobny efekt może dać też nadmiarowe pokrycie, jednakże jednostki nieprawidłowo zaklasyfikowane do populacji docelowej udaje się w większości przypadków zidentyfikować jeszcze przed losowaniem próby, w czasie jej doboru albo też w trakcie nawiązywania kontaktu z respondentem, a zatem skutki tych uchybień (poza wzrostem kosztów) bywają mało dotkliwe. Z kolei problemy wynikające z wielokrotnego pokrycia oraz z grupowania jednostek indywidualnych oddziałują „jedynie" na losowy komponent błędu pomiarowego, obniżając

---

[67] Więcej o rejestrze PESEL powiedziane zostało w ostatniej części tego rozdziału.

[68] L. Kish (1965: 54–55) postuluje nawet, aby w sytuacji prawie pełnego pokrycia populacji docelowej operatem losowania nie przejmować się błędami pokrycia. Po pierwsze, koszty związane z poprawą dopasowania takich operatów do populacji docelowej są zbyt duże. Po drugie natomiast, błąd operatu stanowi w takich przypadkach rzeczywiście niewielką część błędu całkowitego. Dużo poważniejsze są na przykład błędy wynikające z niedostępności jednostek czy też uchybień pomiarowych.

tym samym efektywność próby badawczej. Innymi słowy, związana jest z nimi konieczność zwiększenia liczebności próby oraz zastosowania ważenia rekompensującego nierówne szanse selekcji, jednak nie wpływają one na błąd systematyczny.

Zresztą tym czynnikiem, który powinien w sposób decydujący wpływać na wykorzystanie określonego operatu, jest przede wszystkim odsetek populacji, jaką operat ten pokrywa. Innymi słowy, zasadniczym kryterium wyboru operatu jest zakres populacji docelowej w nim zawartej lub, inaczej, liczba jednostek, do której operat taki pozwala dotrzeć. Jest to szczególnie widoczne w badaniach ESS-u w odniesieniu do krajów, dla których rejestry indywidualne były niedostępne lub ich jakość była niska. W takich krajach poszukiwało się innych wykazów (najczęściej gospodarstw domowych, punktów adresowych, rzadziej operatów opartych na mapach administracyjnych), których zastosowanie wymagało zwiększenia liczebności próby badawczej, ale jednocześnie umożliwiało całkowite lub prawie całkowite dotarcie do elementów populacji docelowej. Owym najczęściej wykorzystywanym typom operatów doboru prób badawczych poświęcona będzie czwarta część tego rozdziału, w której przeanalizowane zostaną dane metodologiczne z badań ESS-u.

III.3. Procedury ograniczania błędów operatów doboru prób badawczych

W książce *Survey Sampling* L. Kish (1965: 56–59) opisał założenia oraz podstawy metodologiczne kilku działań, które można podjąć w celu ograniczenia błędów operatów. Wprawdzie autor przywoływanej monografii scharakteryzował tylko ogólne pomysły takich przedsięwzięć, a nie przedstawił konkretnych procedur, które należy podjąć, to jednak wiele z opisanych przez niego pomysłów stało się podstawą stosowanych obecnie w badaniach surveyowych metod ograniczających błędy operatów. Przyglądając się ustaleniom literaturowym w tym zakresie, można zauważyć, że większość takich działań podejmowanych na rzecz polepszenia jakości operatów losowania ma na celu zwiększenie pokrycia populacji, tj. uwzględnienie jak największej liczby osób, które do niej należą, oraz – chociaż już nie tak często – zmniejszenie udziału tych elementów, które w skład populacji nie wchodzą. W mniejszym stopniu natomiast procedury te służą wyeliminowaniu uzespołowienia jednostek oraz wielokrotnych powtórzeń w operacie tych samych jednostek populacji. Skutki dwóch ostatnich typów błędów mogą być bowiem z powodzeniem redukowane poprzez zastosowanie procedur ważenia danych oraz zwiększanie liczebności prób badawczych. Innymi słowy, błędy niepełnego pokrycia oraz pokrycia

nadmiarowego mogą doprowadzać do systematycznego zniekształcania wyników pomiaru, natomiast błędy grupowania oraz multiplikowania jednostek – do obniżania precyzji estymacji. A zatem, ponieważ błędów nielosowych nie da się minimalizować poprzez standardowe działania badawcze, lepiej ograniczać szanse ich wystąpienia, nawet kosztem przyrostu wariancji estymatorów.

Przykładem takiej filozofii postępowania jest równoczesne stosowanie wielu operatów. Ma to swoje uzasadnienie wtedy, gdy kilka operatów pokrywa łącznie znacznie większą część populacji docelowej niż każdy z nich osobno. Choć ogranicza to ryzyko wystąpienia błędu systematycznego, to jednak konsekwencją stosowania wielu operatów jest zróżnicowanie szans doboru jednostek, co obniża poziom precyzji estymacji. Ponieważ jednak błąd systematyczny ma konsekwencje dużo poważniejsze niż przyrost błędu statystycznego, to wzrost wariancji jest kosztem, który ponosi się w celu wyeliminowania ryzyka kierunkowego zniekształcenia wyników. Podobnie rzecz ma się ze stosowaniem doboru zespołowego (na przykład operatu adresowego gospodarstw domowych). Losowanie takie ma zastosowanie przede wszystkim wtedy, gdy rejestry indywidualne są niepełne[69]. Co prawda, jeżeli wywiadów nie prowadzi się z każdym członkiem wylosowanego gospodarstwa, a jedynie w obrębie mieszkania dobiera się jakąś jedną osobę, to szanse selekcji respondenta pozostają zależne od liczebności członków gospodarstw domowych. Cała procedura wymaga ważenia danych oraz odpowiedniego zwiększenia wielkości próby, ale ogranicza ryzyko pojawienia się błędu systematycznego.

W kolejnych sekcjach rozdziału trzeciego omówione zostaną najczęściej wykorzystywane w praktyce badawczej metody ograniczania błędy operatów. Scharakteryzowane będą założenia owych procedur, omówione zostaną też wyniki analiz empirycznych oraz metodologicznych poświęconych ocenie efektywności tych metod w redukcji błędów operatów doboru próby.

III.3.1. Sieciowanie jednostek

Sieciowanie jednostek wchodzących w skład populacji docelowej jest metodą łączenia elementów populacji poprzez określenie relacji wiążących jednostki (por. Christman 2009: 115; Lee 2008: 506–507; Potter 2008: 491–492; Groves i in. 2004: 86–87; Groves 1989: 122–124). Procedura ta – będąca pewną modyfikacją losowania określanego mianem *multiplicity sampling* – przedstawiona została po raz pierwszy przez Monroe'a G. Sirkena (1970: 257–266) w artykule *Household Surveys with Multiplicity*. W zamyśle autora metoda ta

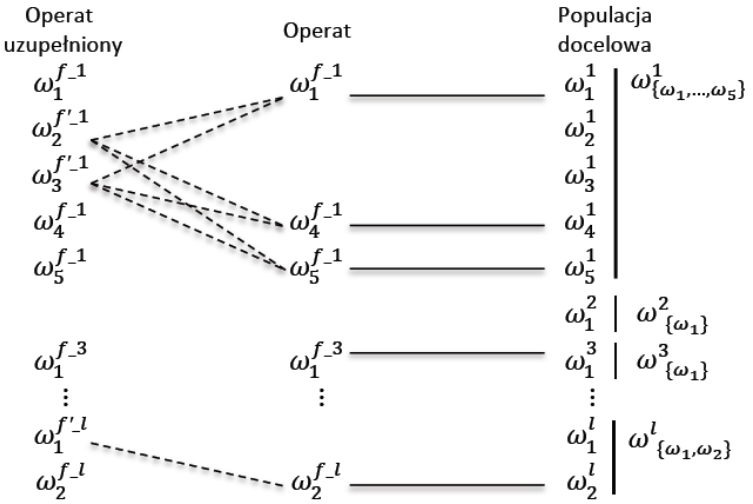
[69] Operat adresowy, o ile tylko jest dobrej jakości, umożliwia dotarcie (w sposób pośredni) do wszystkich jednostek z populacji.

miała być alternatywą dla wielostopniowego schematu doboru próby osób opartego na rejestrach gospodarstw domowych. Innowacyjność pomysłu Sirkena przejawiała się w tym, że o ile w typowym sondażu prowadzonym na adresowych próbach gospodarstw domowych każda jednostka przyporządkowana zostaje tylko i wyłącznie do tego gospodarstwa, którego jest mieszkańcem (stąd, gdy w rejestrze brakuje wykazu jakichś gospodarstw domowych, to też osoby tworzące owo gospodarstwo są wyłączone poza operat), o tyle usieciowienie łączy jednostki z różnych gospodarstw domowych, co pozwala na odnalezienie nawet tych osób, których gospodarstwa są niewykazane w operacie. M.B. Sirken ukazuje również, że sieciowanie pozwala osiągnąć większą precyzję estymacji i to nawet pomimo niebezpieczeństw związanych z pojawieniem się specyficznych kategorii błędów pomiarowych (por. Sirken 1970: 266). W procedurze sieciowania jednostek każda osoba przyporządkowywana zostaje zarówno do tego gospodarstwa domowego, którego jest mieszkańcem, jak i do innych gospodarstw domowych zamieszkiwanych przez osoby tworzące sieć z tą jednostką. Reguły łączenia jednostek oparte są najczęściej na pytaniach o relacje bliskiego pokrewieństwa (z rodzicami, rodzeństwem oraz potomstwem) lub na powiązaniach przestrzennych pomiędzy jednostkami (por. Sirken 1970: 257)[70].

W tekście opublikowanym dwa lata później Sirken przedstawił również inny wariant procedury sieciowania odpowiedni dla stratyfikacyjnych schematów doboru prób badawczych (por. Sirken 1972: 224–227). Metoda ta wymaga, aby wylosowane osoby wskazywały wszystkich swoich domowników oraz tych bliskich członków rodziny, którzy zamieszkują w innych oddzielnych gospodarstwach domowych. Lista takich osób tworzy sieć jednostek podlegającą pomiarowi (por. Groves i in. 2004: 86). Procedura ta umożliwia, przynajmniej teoretycznie, zidentyfikowanie wszystkich osób z populacji, nawet tych, które w operatach się nie znalazły. Jest to możliwe pod jednym wszakże warunkiem, a mianowicie: że każda jednostka pominięta w rejestrze będzie miała szanse utworzenia sieci z przynajmniej jedną osobą wykazaną w operacie. Oznacza to, że sieciowanie jednostek nie spełni swojej funkcji w odniesieniu do tych osób pominiętych w dostępnych rejestrach populacji, które nie mają bliskiej rodziny (osoby takie przez nikogo nie będą miały szansy być wskazane). Całe przedsięwzięcie sieciowania można scharakteryzować w formie graficznej za pomocą ryciny III.3.

---

[70] Pewną alternatywą wobec zadawania dodatkowych pytań sieciujących jednostki w obrębie rodziny jest poszukiwanie osób niepokrytych operatem metodą „kuli śniegowej" (por. Groves i in. 2004: 86). Chociaż jest to nieprobabilistyczny sposób doboru respondentów (por. Sawiński 2005: 83), to jednak bywa wykorzystywany w badaniach realizowanych na (sub)populacjach osób posiadających pewne specyficzne i rzadkie cechy (por. Groves i in. 2004: 86).



Ryc. III.3. Redukcja błędu niepełnego pokrycia poprzez sieciowanie jednostek

Źródło: opracowanie własne

W zaprezentowanym układzie operat doboru próby nie pokrywa w pełnym zakresie populacji docelowej. Jednostki z populacji, które okazały się niedostępne w operacie losowania, oznaczono symbolami ω_2^1 – druga jednostka z pierwszej sieci, ω_3^1 – trzecia jednostka z pierwszej sieci, ω_1^2 – pierwsza jednostka z sieci drugiej (jednoosobowej) oraz ω_2^l – druga jednostka z sieci l–tej. Sieciowanie umożliwi identyfikację jednostek: ω_2^1, ω_3^1 oraz ω_1^l, jednakże pomimo jej zastosowania poza operatem pozostanie nadal jednostka ω_1^2, która nie ma szansy być przez nikogo wskazana.

Jedną z ciekawszych implementacji procedury sieciowania są analizy empiryczne przeprowadzone przez Roberta D. Tortorę i in. (2008: 133–148). Weryfikacja skuteczności sieciowania prowadzona była w tych badaniach na próbie dobieranej z operatu abonentów telefonii stacjonarnej, a jej podstawowym celem było wyszukanie osób niepokrytych tym rejestrem. Respondenci proszeni byli o podanie informacji o wszystkich członkach swojej najbliższej rodziny (rodzicach, rodzeństwie oraz pełnoletnich dzieciach), którzy mieszkają w gospodarstwach domowych bez dostępu do telefonu stacjonarnego, ale jest z nimi możliwy kontakt poprzez telefon komórkowy. Badania terenowe dały niejednoznaczne rezultaty. Z jednej strony procedura sieciowania umożliwiła odszukanie pewnej części jednostek niedostępnych w operatach abonentów telefonii stacjonarnej, z drugiej jednak ujawniła niepokojące prawidłowości. Po pierwsze, wiedza o członkach rodziny zamieszkujących w oddzielnych gospodarstwach

domowych uwarunkowana była stopniem zintegrowania jednostek w obrębie sieci (rodziny). Po drugie, zaobserwowano powszechną niechęć do ujawniania numerów telefonów komórkowych. Po trzecie, wiele z tych osób, które wskazywano jako posiadaczy wyłącznie numerów komórkowych, miało jednak w swoich gospodarstwach domowych dostęp do telefonu stacjonarnego.

Być może z uwagi na wskazane ograniczenia praktyczne sieciowanie jednostek nie stało się procedurą wykorzystywaną na wyjątkowo szeroką skalę. Podstawowym wyzwaniem okazały się bowiem błędy pomiarowe (niewłaściwe lub niepełne raportowanie o sieciach), braki danych wynikające ze znacznej niedostępności wykazanych jednostek oraz przyrost wariancji, będący konsekwencją ważenia danych rekompensującego nierówne szanse selekcji jednostek do próby badawczej[71] (por. Lee 2008: 507; Potter 2008: 491; Groves i in. 2004: 87). Na niektóre z tych komplikacji wskazał Graham Kalton (2009: 135–136) w artykule *Methods for Oversampling Rare Subpopulations in Social Surveys.* W ramach podsumowania części poświęconej próbom sieciowym autor ten zamieszcza następującą konkluzję:

> Korzyści wynikające z próbkowania sieciowego są częściowo niwelowane przez przyrost błędu losowego będącego konsekwencją ważenia zmiennych, którego procedura ta wymaga, oraz poprzez koszty zlokalizowania jednostek wskazanych w sieci. (Kalton 2009: 136)

Otwarte pozostaje zatem pytanie o to, czy i ewentualnie w jaki sposób, a także w odniesieniu do jakich przypadków procedura ta mogłaby przyczynić się do ograniczenia błędu niepełnego pokrycia populacji operatami doboru prób badawczych. Kilka przykładów pozwoli zobrazować możliwości jej empirycznej implementacji.

W pierwszej kolejności warto rozpatrzyć schemat realizacji badań Europejskiego Sondażu Społecznego w dwóch krajach: w Irlandii (wszystkie rundy ESS) oraz we Włoszech (rundy 1 oraz 2). Przypadki tych dwóch krajów są o tyle interesujące, że w każdym z nich dysponuje się operatami (rejestrami wyborców) umożliwiającymi losowanie proste w obrębie populacji mieszkańców w wieku 18 lat i więcej posiadających prawa wyborcze (por. Lynn i in. 2007: 110). Pomimo to, na podstawie tych rejestrów dobiera się jedynie adresowe próby osób. Problem polega na tym, że w badaniach ESS populację docelową stanowią

---

[71] Osoby wchodzące w skład sieci mają zwielokrotnione szanse znalezienia się w próbie, gdyż liczebność sieci wyznacza prawdopodobieństwo selekcji. Aby zrekompensować niejednakowe prawdopodobieństwa wyboru, konieczne jest zatem ważenie danych oraz zwiększanie liczebności próby badawczej. Odpowiednie procedury ważenia danych odnaleźć można w artykule Sirkena (1970: 258). Opierają się one na prostej zasadzie, podobnej do tej, którą wykorzystuje się w ważeniu danych prób zespołowych (por. Weisberg 2005: 221), to znaczy każda jednostka otrzymuje wagę równą odwrotności liczby wystąpień w operacie (por. Groves i in. 2004: 86).

wszyscy mieszkańcy w wieku powyżej 15 lat. Operaty dostępne w Irlandii i we Włoszech nie zawierają osób 15-, 16- oraz 17-letnich, co w konsekwencji oznacza, że charakteryzuje je błąd niepełnego pokrycia. Ponieważ kohorta wiekowa z przedziału 15–17 lat stanowi we Włoszech mniej więcej 3% całej populacji osób w wieku 15 lat i więcej, a w Irlandii niecałe 5%, to pokrycie populacji rejestrami wyborców wynosi odpowiednio 97% oraz 95%[72]. Operaty pokrywają zatem populacje w stopniu na tyle znacznym, że w wielu sytuacjach pokrycie to uznano by za wystarczające. Problem wynika jednak z tego, że jednostki niepokryte wykazami wyborców stanowią zbiór homogeniczny względem kryterium wieku, to znaczy ich niewystępowanie w operacie nie jest losowe. W konsekwencji, w obu krajach konieczne jest przyjęcie takich schematów doboru prób, które polegają na wykorzystaniu rejestrów indywidualnych do losowania punktów adresowych (budynków lub gospodarstw domowych); dopiero w ich obrębie losuje się konkretnego respondenta[73]. Taki sposób realizacji badań skutkuje nierównymi szansami doboru jednostek w obrębie gospodarstw domowych o różnej liczbie członków, czyli obniża poziom precyzji wnioskowania (por. Lynn i in. 2007: 112–113).

Wydaje się, że sieciowanie jednostek mogłoby być alternatywą dla ustalonego w obu krajach schematu doboru respondentów. Otóż, ponieważ operaty jednostkowe pokrywają w tych państwach mniej więcej 95% oraz 97% całej populacji osób w wieku 15 lat i więcej, a populację 18+ pokrywają całkowicie, to dla tej w pełni pokrytej części populacji można by zastosować losowanie proste (ewentualnie warstwowanie lub wiązkowanie części próby), natomiast błąd pokrycia osób w wieku 15–17 lat ograniczany byłby w trakcie terenowej fazy badań. Warto bowiem zauważyć, że ankieterzy prowadzący wywiady kwestionariuszowe w ramach ESS zadają bardzo szczegółowe pytania dotyczące składu oraz charakterystyk osób w obrębie gospodarstw domowych; nic nie stałoby na przeszkodzie, aby takie informacje wykorzystać do sieciowania osób z kohorty wiekowej 15–17 lat. Wywiady kwestionariuszowe realizowane mogłyby być wtedy ze wszystkimi wyszukanymi w ten sposób osobami lub też losowano by jakąś ich próbę. Co oczywiste, szansa udziału w badaniu osób z kohorty wiekowej 15–17 lat uzależniona byłaby od prawdopodobieństw selekcji jednostek powiązanych z nimi relacją sieci (wymagałoby to zastosowania jakiejś procedury ważenia), jednak ów schemat realizacji próby mógłby być interesującą alternatywą dla stosowanych wieloetapowych schematów losowania budynków lub gospodarstw domowych.

---

[72] Obliczenia własne na podstawie repozytorium Eurostatu. http://epp.eurostat.ec.europa.eu/portal/page/portal/statistics/search_database [data pobrania: 3.12.2012].

[73] W Irlandii do wywiadu dobierano osoby, których data urodzin przypadała w terminie najbliższym wizyty ankieterskiej, natomiast we Włoszech losowano respondentów, wykorzystując siatkę Kisha.

Kolejnym obszarem zastosowania procedury sieciowania jednostek mogłyby być sondaże prowadzone techniką wywiadów telefonicznych (z generowaniem numerów stacjonarnych oraz komórkowych) uzupełniane wywiadami bezpośrednimi z osobami bez dostępu do telefonu. Wprawdzie operaty RDD dają możliwość dotarcia do 95% całej populacji dorosłych obywateli Polski, lecz jednocześnie poza operatem znajdują się najczęściej osoby starsze, zamieszkujące obszary wiejskie, gorzej sytuowane finansowo oraz te o niższym poziomie wykształcenia (por. Batorski 2011: 309)[74]. Zresztą jest to prawidłowość zaobserwowana nie tylko w Polsce, ale też w innych krajach, na przykład w Stanach Zjednoczonych (por. Duncan i in. 2001: 121). Wyłączając w tym momencie poza nawias rozważań inne konsekwencje wynikające z losowego generowania numerów telefonicznych, można zauważyć, że nieobecność w operacie znacznej części osób z takich właśnie specyficznych kategorii jednostek populacji mogłaby stanowić poważny problem w badaniach ukierunkowanych na kwestie związane z poziomem życia czy też sytuacją materialną. Każdą wylosowaną do próby osobę można by wpierw zapytać, czy ma ona jakichś bliskich członków rodziny (rodziców, rodzeństwo, dzieci) zamieszkujących wspólnie lub w osobnych gospodarstwach domowych, a następnie, czy te gospodarstwa domowe mają dostęp do telefonu stacjonarnego. Jeżeli nie, to należałoby dodatkowo poprosić o wskazanie, która z takich osób posiada telefon komórkowy, a która z takiego urządzenia nie korzysta. Być może na podstawie serii pytań filtrujących udałoby się zidentyfikować osoby, z którymi realizacja wywiadów możliwa byłaby jedynie poprzez wywiad osobisty. Rodzi to wprawdzie wiele komplikacji wynikających z zastosowania technik mieszanych oraz

---

[74] Dane uzyskane w ramach projektu badawczego Diagnozy Społecznej 2011 ukazują, że w miastach pow. 500 tys. mieszkańców odsetek gospodarstw domowych mających dostęp do telefonu stacjonarnego wynosi 61 pp., podczas gdy na wsiach kształtuje się na poziomie niecałych 53 pp. Podobnie wyglądają zróżnicowania w odsetku osób korzystających z przynajmniej jednego telefonu komórkowego. W największych miastach poziom ten wynosi prawie 94 pp., z kolei na wsiach jest o 15 pp. niższy. Jeszcze większe zróżnicowanie daje się zauważyć, gdy weźmie się pod uwagę strukturę wiekową osób korzystających z telefonii komórkowej. Dla przykładu, w kohortach wiekowych 16–24, 25–34 oraz 35–44 lat posiadanie telefonu komórkowego deklaruje mniej więcej 97 procent badanych, natomiast wśród osób w wieku 65 lat i więcej użytkownicy telefonów mobilnych stanowią już niecałe 49 procent. Równie znaczne dysproporcje widoczne są po uwzględnieniu wykształcenia jako kategorii różnicującej korzystanie z telefonu komórkowego. Wśród osób z wykształceniem wyższym do korzystania z telefonii komórkowej przyznaje się prawie 96 procent badanych, podczas gdy wśród osób z wykształceniem podstawowym odsetek ten wynosi już nieco powyżej 51 pp. Warto wreszcie zwrócić uwagę na zróżnicowanie w obrębie dochodów przypadających na jedną osobę w badanych gospodarstwach. Kontrastując ze sobą respondentów o dochodach najniższych (do pierwszego kwartyla) oraz najwyższych (powyżej trzeciego kwartyla) widać, że wśród tej pierwszej kategorii osoby korzystające z telefonu komórkowego stanowią nieco ponad 76 procent, w drugiej natomiast już ponad 94 procent (por. *Diagnoza społeczna: zintegrowana baza danych*, www.diagnoza.com [data pobrania: 10.11.2012] oraz tabela 7.2.1. w: Batorski 2011: 309).

skutkować może pojawieniem się błędów pomiarowych, jednak sieciowanie byłoby alternatywą dla procedury losowania wykorzystującej wiele różnych operatów jednocześnie.

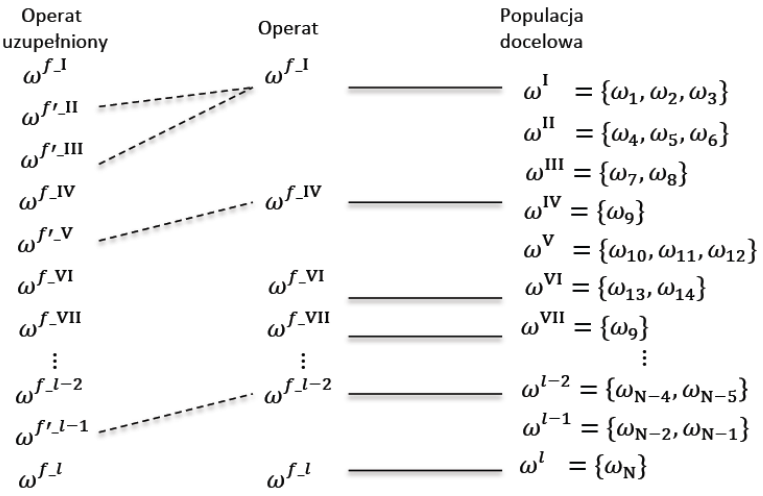
III.3.2. Procedura przedziałów półotwartych

Kolejną metodą umożliwiającą redukcję błędu niepełnego pokrycia operatu doboru prób badawczych jest procedura przedziałów półotwartych (por. Kim 2008: 310; Groves i in. 2004: 84–86; Groves 1989: 127–128; Kish 1965: 56). Podstawowym wymogiem jej zastosowania jest konieczność uporządkowania jednostek objętych rejestrem według jakiegoś kryterium, na przykład uszeregowanie adresów gospodarstw domowych w porządku rosnącym. Cała istota tej techniki opiera się bowiem na tym, że łączy się elementy udokumentowane w rejestrze z takimi wszystkimi pominiętymi obiektami populacji docelowej, które poprzedzają kolejny (w ustalonym porządku) element wykazany w operacie. Innymi słowy, procedura ta polega na wyszukiwaniu takich pominiętych jednostek statystycznych lub ich zespołów, które znajdują się pomiędzy dwoma sąsiadującymi obiektami wykazanymi w operacie doboru próby. Istotę procedury przedziałów półotwartych świetnie oddają rozważania R. Grovesa (1989), który stwierdza, że:

> wynikiem [jej zastosowania – P.J] jest to, że każdy element z operatu losowania koresponduje z wiązką elementów populacji docelowej, to znaczy jednym elementem z operatu oraz z innymi elementami nieznajdującymi się w operacie, które poprzedzają następny w kolejności obiekt wykazany w operacie. Dlatego proces ten, w razie powodzenia, pozwala na przypisanie każdego elementu z populacji docelowej do jednego i tylko jednego elementu wykazanego w operacie, ale jednocześnie też do każdego elementu z operatu losowania może być przypisany więcej niż jeden element populacji docelowej. (Groves 1989: 127)

Założenia te można zaprezentować również w formie graficznej (ryc. III.4).

Ponieważ procedurę przedziałów półotwartych wykorzystuje się najczęściej do redukcji błędu niepełnego pokrycia operatami adresowymi, to na zaprezentowanym na ryc. III.4 schemacie elementy populacji docelowej grupują jednostki indywidualne. Niedostępne w operacie losowania zespoły jednostek oznaczono symbolami ω^{II}, ω^{III}, ω^{V} oraz ω^{l-1}. Zauważyć można, że ponieważ pierwszym elementem wykazanym w operacie jest ω^{I}, a kolejnym ω^{IV}, to w zakres przedziału półotwartego, którego początkowym elementem jest ω^{I}, wchodzi zbiór jednostek $\{\omega^{\mathrm{II}}, \omega^{\mathrm{III}}\}$, ale już nie ω^{IV}. Analogicznie można zauważyć, że ponieważ po obiekcie oznaczonym ω^{IV} występuje obiekt ω^{VI}, to niewykazany w rejestrze zbiór jednoelementowy $\{\omega^{\mathrm{V}}\}$ jest powiązany z ω^{IV}. Z kolei, jako że ω^{VII} nastę-



Ryc. III.4. Redukcja błędu niepełnego pokrycia poprzez przedziały półotwarte

Źródło: opracowanie własne

puje po ω^{VI}, a pomiędzy nimi nie ma żadnych innych niewykazanych obiektów, to zbiór jednostek powiązanych z ω^{VI} jest pusty. W sytuacji tej uznano by, że żaden element populacji nie został przez operat pominięty.

Procedurę przedziałów półotwartych można wykorzystać w losowaniu jednostek przestrzennych (por. Groves 1989: 127), budynków mieszkalnych (por. Mulry 2008: 165) lub też gospodarstw domowych (por. Hall 2008: 35), a także – przynajmniej teoretycznie – do redukcji błędów pokrycia rejestrów abonentów telefonii stacjonarnej. W tym ostatnim przypadku celem byłoby wyszukanie gospodarstw bez dostępu do linii telefonicznej (por. Groves 1989: 128) [75].

[75] R. Groves (1989), odnosząc się w książce *Survey Errors and Survey Costs* do możliwości redukcji błędu pokrycia populacji gospodarstw domowych operatami abonentów telefonicznych, podaje, że w przypadku zastosowania przedziałów półotwartych „[ankieterzy – P.J.] mogliby prosić osobę należącą do wylosowanego gospodarstwa domowego o określenie tego, czy następne gospodarstwo domowe ma dostęp do telefonu stacjonarnego. Jeżeli nie, to gospodarstwo takie byłoby sieciowane z tym, przez które zostało wskazane. Następnie byłaby podejmowana próba bezpośredniego kontaktu z członkami takich gospodarstw domowych. Dużym wyzwaniem związanym z zastosowaniem przedziałów półotwartych w operatach abonentów telefonicznych jest wieloznaczność pojęcia 'następnego gospodarstwa domowego' oraz brak wiedzy członków jednych gospodarstw domowych o wyposażeniu innych gospodarstw w telefony stacjonarne" (Groves 1989: 128). R. Groves zamieścił to stwierdzenie jeszcze przed rozkwitem telefonii komórkowej. Przy zastosowaniu losowego generowania numerów telefonicznych (stacjonarnych oraz komórkowych) w losowaniu prób badawczych procedura przedziałów półotwartych traci sens, głównie z uwagi na fakt, że poszukiwać powinno się osób niepokrytych operatami *RDD*, a nie tylko gospodarstw domowych bez dostępu do telefonii stacjonarnej. W celu redukcji błędu niepełnego pokrycia operatami *RDD* wykorzystać można jednak procedurę sieciowania jednostek lub też procedurę wielokrotnych operatów (rejestry abonentów oraz wykazy adresowe).

Zauważyć należy, że zastosowanie metody przedziałów półotwartych pociąga za sobą konieczność przeprowadzenia takiego doboru jednostek, który każdemu wylosowanemu obiektowi przyporządkowuje również ten następujący po nim w kolejności. Tylko wówczas, gdy spełni się ten warunek, będzie można ustalić granice przedziałów delimitujących zakres poszukiwania obiektów pominiętych w rejestrach, znajdujących się pomiędzy obiektem wylosowanym a następującym po nim. Empiryczne implementacje procedury przedziałów półotwartych odnaleźć można między innymi w monografii *Survey Methodology* (por. Groves i in. 2004: 84–86). Z kwestii zupełnie fundamentalnych należy wskazać, że w przypadku odszukania jakiegoś nowego adresu (lub wielu adresów pominiętych w rejestrze) należy je wszystkie dołączyć do próby. Badanie powinno zatem objąć zarówno te jednostki, które wylosowano z operatu, jak również te, które do nich przypisano. Dzięki temu wszystkie dodatkowe obiekty będą miały jednakowe szanse selekcji do próby, równe prawdopodobieństwom wylosowania tego obiektu, do którego zostały przypisane. W praktyce, gdy tych adresów jest wiele, stosuje się dodatkowe losowanie w obrębie zidentyfikowanego zbioru obiektów[76]. Oznacza to jednak zróżnicowanie szans ich selekcji do próby, zupełnie tak samo jak w schemacie losowania jednego reprezentanta z wieloosobowych gospodarstw domowych. Konieczne jest zatem ważenie danych lub zwiększenie liczebności próby o wielkość ekwiwalentną przewidywanej utracie precyzji estymacji (por. Groves i in. 2004: 86).

Z procedurą przedziałów półotwartych związane są jednak pewne istotne problemy natury praktycznej. Po pierwsze, pozwala ona na identyfikację jednostek pominiętych w operatach, ale jednocześnie okazuje się nieskuteczna w eliminowaniu wielokrotnych wskazań tych samych jednostek oraz wyłączaniu poza operat jednostek nieprawidłowo zakwalifikowanych do populacji docelowej (por. Mulry 2008: 165). Po drugie, jak ukazują analizy empiryczne, ankieterzy realizujący wywiady skupiają się bardziej na skutecznym dotarciu do respondenta oraz przeprowadzeniu z nim wywiadu, niż na odszukaniu dodatkowych adresów. Co więcej, nawet jeśli takie dodatkowe adresy identyfikują, to często popełniają błędy (por. Groves 1989: 128). Typowym przykładem komplikacji, w jakie uwikłana jest próba empirycznej implementacji przedstawionej metody, są wyniki eksperymentu przeprowadzonego przez Stephanie Eckman oraz Colma O'Muircheartaigha, opublikowane w 2011 roku w artykule *Performance of the Half-Open Interval Missed Housing Unit Procedure*. Eksperyment polegał na przekazaniu ankieterom wykazu punktów adresowych, z któ-

[76] L. Kish (1965: 350) sugeruje przeprowadzenie losowania w obrębie zidentyfikowanych adresów, jeżeli tylko ich liczba przekracza pięć obiektów. Jest to jednak wyłącznie kwestia umowna, bowiem w opracowaniu J. Lepkowskiego i in. (2010: 12) odnaleźć można sugestię, że losowanie powinno być przeprowadzone już w sytuacji odnalezienia dwóch dodatkowych adresów.

rego celowo usunięto informację o pewnych istniejących adresach. Każdy z takich ukrytych adresów powinien być przez ankieterów odnaleziony jako brakujący obiekt, o ile oczywiście procedura przedziałów półotwartych zostałaby przez nich prawidłowo przeprowadzona. Wyniki eksperymentu ukazały, że jedynie w 15 na 140 przypadków ankieterzy odnaleźli brakujące adresy, zatem skuteczność wyszukiwania adresów była bardzo niska. Ponadto, procedurze tej towarzyszył efekt ankieterski stawiający pod znakiem zapytania uczciwość pracy ankieterów. Okazało się bowiem, że mieli oni skłonność do wykazywania adresów, które w rzeczywistości nie powinny być przez nich uwzględniane. Przyczyna tego stanu rzeczy okazała się trywialna, bowiem ankieterzy nieprawidłowo wykazywali adresy z przyczyn czysto instrumentalnych: w celu otrzymania dodatkowej gratyfikacji finansowej za realizację nadprogramowych wywiadów. Innymi słowy, przydatność procedury przedziałów półotwartych okazała się niesatysfakcjonująca. Czynnikiem dodatkowo ograniczającym sens jej wykorzystania była konieczność przeprowadzenia szczegółowej kontroli pracy ankieterów (por. Eckman i in. 2011: 127–130).

III.3.3. Operaty wielokrotne

W wielu sytuacjach problem niepełnego pokrycia udaje się ograniczyć (lub nawet wyeliminować) poprzez jednoczesne zastosowanie wielu operatów. Wydaje się, że procedura ta jest najczęściej wykorzystywaną techniką redukcji błędu niepełnego pokrycia populacji operatem losowania. Całość opiera się na bardzo prostej idei, a mianowicie, że jeśli nawet pojedynczy wykaz jednostek obejmuje wyłącznie część populacji docelowej, to jednak razem zawierają one więcej, niż każdy z nich osobno. Przykładem wykorzystania procedury wielu operatów jest uzupełnianie wykazu numerów telefonicznych (stacjonarnych i/lub mobilnych) o adresową listę gospodarstw domowych. Takie połączenie pozwala na wylosowanie do próby tych osób, które nie mają dostępu do telefonii stacjonarnej oraz/lub nie korzystają w ogóle z telefonii komórkowej (por. Link i in. 2011: 613–635; Groves i in. 2004: 87–88). Być może najważniejszym udogodnieniem stojącym za użyciem różnych operatów pozostaje jednak możliwość redukcji kosztów związanych z realizacją badań terenowych. Dzieje się tak, gdy główna część badań prowadzona jest w oparciu o mniej kosztochłonną technikę wywiadów telefonicznych, uzupełnianą – wymagającą większych nakładów – techniką wywiadów osobistych prowadzonych już na próbach adresowych (por. Opsomer 2011: 227; Lohr 2009: 72–73).

Jednak najbardziej charakterystycznym przykładem zastosowania procedury operatów wielokrotnych pozostają badania prowadzone techniką wywiadów telefonicznych z losowym generowaniem numerów stacjonarnych oraz komór-

kowych[77]. Nie tyle zasadność, ile konieczność wykorzystania w wywiadach telefonicznych takich podwójnych operatów potwierdzają wyniki *Diagnozy Społecznej* z 2011 roku. W raporcie sprawozdawczym z tych badań, w części poświęconej korzystaniu przez Polaków z technologii teleinformatycznych, odnaleźć można niezwykle interesującą konstatację odnośnie zmian w dostępie do telefonii stacjonarnej oraz sieci GSM:

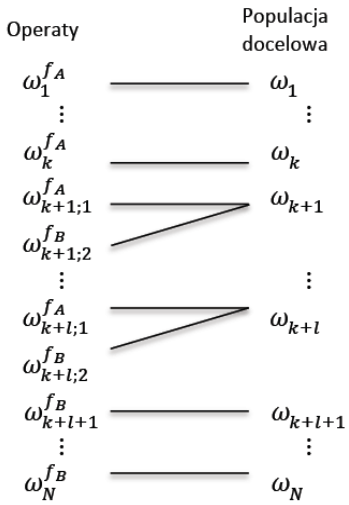
> Nadal maleje liczba gospodarstw domowych posiadających telefon stacjonarny. Obecnie jest ich nieco ponad połowa. Oczywiście wynika to przede wszystkim z upowszechnienia telefonii komórkowej – obecnie telefon komórkowy posiadają osoby z 87,9 proc. gospodarstw domowych, a więc komórki są w znacznie większej liczbie gospodarstw domowych niż telefony stacjonarne. W aż 57,7 proc. gospodarstw telefon komórkowy ma każda osoba. Własny telefon komórkowy ma 85,1 proc. osób w wieku 16 i więcej lat. Wśród osób, które nie mają telefonu komórkowego, 10,1 proc. ma w domu telefon stacjonarny. Zaledwie 4,8 proc. to osoby, które nie mają ani własnej komórki, ani telefonu stacjonarnego. (Batorski 2011: 301)

Innymi słowy, chociaż połączone rejestry nie zawierają wszystkich jednostek z populacji, to jednak pokrywają one dużo większą jej część, niż brane pojedynczo spisy abonentów stacjonarnych oraz wykazy posiadaczy telefonów mobilnych. Oczywiście, choć zwiększenie pokrycia populacji ogranicza ryzyko błędu systematycznego, to jednak może prowadzić do innych błędów losowych oraz nielosowych[78].

---

[77] Można zauważyć, że ma się tu rzeczywiście do czynienia z dwoma różnymi operatami. Choć w obu sytuacjach losowaniu podlega numer telefoniczny, to jednak wygenerowanie numeru stacjonarnego oznacza losowanie gospodarstwa domowego (tj. zespołu osób, z których należy jeszcze dobrać konkretnego respondenta), natomiast wygenerowanie numeru komórkowego oznacza wylosowanie jednostki, tak więc żadnego dodatkowego losowania nie trzeba już przeprowadzać).

[78] Ciekawym studium ukazującym ograniczenia możliwości wykorzystywania losowego generowania numerów stacjonarnych oraz komórkowych w badaniach prowadzonych techniką *CATI* pozostają studia Britty Busse i in. (2012: 1209–1225), jak również analizy M.J. Bricka i in. (2011: 1–12), będące kontynuacją wcześniejszych badań Bricka i in. (1995: 218–235). Autorzy ci ukazują, że potencjalne korzyści wynikające z redukcji błędu pokrycia poprzez zastosowanie wielu operatów niwelowane są przez przyrost błędów braku odpowiedzi oraz błędów pomiarowych. Innymi słowy, w przywoływanych studiach ukazano, że włączenie do próby badawczej użytkowników telefonów komórkowych może skutkować znacznym wypaczeniem danych wynikającym z większej liczby odmów udziału w badaniu oraz z komplikacji pomiarowych. Zatem korzyści wynikające z zastosowania podwójnych operatów okazują się w najlepszym przypadku zupełnie marginalne. W podobnym duchu wypowiadają się też inni badacze, wymieniając choćby P. Vicente i in (2009: 105–111), Tephena Blumberga i in. (2007: 734–749) czy też Eleanor Singer (2006: 637–645). Nie jest to szczególnie zaskakujące, bowiem działania podejmowane w celu ograniczenia jednych typów błędów mogą prowadzić do przyrostu wielkości błędów z innych źródeł. Proces badawczy jest przecież pewną całością i jako taką należy ją ostatecznie rozpatrywać. O tym, czy zastosować procedurę wielokrotnych operatów, decydować będzie zatem nie tylko to, czy działanie takie pozwoli zredukować błędy pokrycia, ale również, czy ogólne korzyści będą przewyższać negatywne konsekwencje.



Ryc. III.5. Redukcja błędu niepełnego pokrycia poprzez wykorzystanie wielu operatów

Źródło: opracowanie własne

Mechanizm ograniczenia błędu pokrycia poprzez wielokrotne operaty można przedstawić w formie graficznej zaprezentowanej na rycinie III.5.

Schemat ten ukazuje, że pokrycie populacji wieloma operatami pociąga za sobą nietrywialne komplikacje natury metodologicznej. Bodaj najpoważniejszą konsekwencją jest to, iż losując próbę z wielu operatów, daje się większe szanse selekcji tym wszystkim jednostkom, które znajdują się w wykazie więcej niż jednego rejestru (por. Groves 1989: 126–127). Dla przykładu, gdy operaty abonentów telefonii stacjonarnej zostają uzupełnione operatami adresowymi gospodarstw domowych, to te gospodarstwa domowe, które posiadają dostęp do linii telefonicznej, mają większe prawdopodobieństwo wylosowania od tych, z którymi można nawiązać wyłącznie bezpośredni kontakt. W konsekwencji te pierwsze będą nadreprezentowane, a drugie niedoreprezentowane w próbie badawczej. Podobnie dzieje się, gdy losowanie opiera się na operatach abonentów telefonii stacjonarnej oraz mobilnej. Po pierwsze, większe szanse selekcji mają wtedy użytkownicy telefonów komórkowych posiadający również w swoich gospodarstwach domowych telefon stacjonarny. Po drugie, abonenci wielu numerów komórkowych lub stacjonarnych mają zwiększone szanse wyboru w porównaniu do osób korzystających wyłącznie z jednego numeru.

W literaturze metodologicznej odnaleźć można opis wielu metod umożliwiających rozwiązanie tego problemu. Do najważniejszych z nich należy zaliczyć: (a) działania zmierzające do wyeliminowania wielokrotnego pokrycia jednostek (por. Groves i in. 2004: 88), (b) szacowanie wartości parametrów

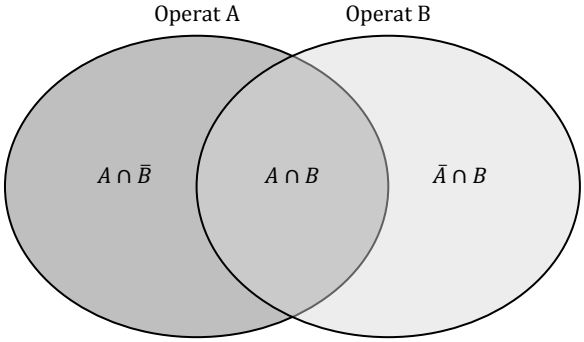
populacyjnych w oparciu o estymatory Horvitza–Thomsona przypisujące każdej jednostce wagi odpowiadające odwrotnościom szans ich selekcji do próby badawczej (por. Kalton 2009: 134; Groves i in. 2004: 88) wraz z różnymi modyfikacjami procedur przypisywania wag (por. Chu i in. 1999: 103–104; Lavallée 1995: 25–32) oraz (c) szacowanie wartości parametrów populacyjnych w oparciu o estymatory Hartleya, minimalizujące wariancję w zbiorze jednostek wylosowanych ze wspólnej części obu operatów doboru próby (por. Kalton 2009: 135; Groves i in. 2004: 88–89; Groves 1989: 125–126) wraz z różnymi modyfikacjami tych estymatorów (por. Lohr 2011: 197–213; Lohr 2009: 76–84; Buskirk 2008: 212–215).

Pierwsza ze wspomnianych metod opiera się na tym, że w trakcie nawiązywania kontaktu z osobą wylosowaną do próby z operatu uzupełniającego niepełne pokrycie operatu głównego ustala się również, czy osoba taka ma szanse doboru z operatu zasadniczego. Jeżeli tak, to nie jest ona włączana do próby badawczej. Dla przykładu, gdyby główna część badań realizowana była w oparciu o próbę dobieraną z rejestru abonentów telefonii stacjonarnej, uzupełnianą adresową próbą gospodarstw domowych, to na etapie aranżacji wywiadu z przedstawicielem gospodarstwa domowego wylosowanego z próby adresowej ankieter ustalałby, czy gospodarstwo to ma również dostęp do telefonii stacjonarnej. Jeżeli tak, to wywiad nie byłby kontynuowany. Pozwala to na uzupełnienie niepełnego pokrycia populacji przez operat główny oraz ogranicza wielokrotne pokrycie tych samych jednostek losowania (por. Brick i in. 2011: 1–12; Groves i in. 2004: 88). Gdyby jednak pomiar przeprowadzony został na wszystkich jednostkach (niezależnie od operatu, z którego zostały one dobrane), to w konsekwencji nadreprezentowane byłyby te jednostki, które obecne są w wykazach wielu operatów jednocześnie.

Jeśli ograniczy się rozważania wyłącznie do dwóch rejestrów, można zauważyć, iż w takiej sytuacji jednostki wchodzące w skład populacji objętej operatem losowania mogłyby zostać przydzielone do trzech różnych warstw[79]: (a) zbioru jednostek należących do operatu *A,* ale już nie do operatu *B*, tj. $A \cap \bar{B}$, (b) zbioru jednostek należących do operatu *B*, ale nie do operatu *A*, tj. $\bar{A} \cap B$, a także (c) zbioru jednostek należących do części wspólnej obu operatów, tj. $A \cap B$.

Z kolei próbę badawczą dobraną z takich rejestrów można by podzielić na cztery zbiory: (a) jednostki wylosowane z operatu *A*, ale nienależące do operatu *B*, (b) jednostki wylosowane z operatu *B* i tylko w tym operacie wykazane,

[79] Jeżeli oba operaty nie będą pokrywać w pełni populacji docelowej, to poza trzema opisanymi zbiorami jednostek będzie jeszcze zbiór czwarty, obejmujący te wszystkie elementy populacji docelowej, które znajdują się poza operatem *A* oraz *B*. Stosując zapis formalny, zbiór jednostek niepokrytych przez te operaty oznaczyć można jako $\bar{A} \cap \bar{B}$.



Ryc. III.6. Procedura wielokrotnych operatów – podział populacji na podzbiory

Źródło: opracowanie własne

(c) jednostki wylosowane z operatu *A*, lecz znajdujące się również w operacie *B* oraz (d) jednostki wylosowane z operatu *B* i znajdujące się też w wykazie *A* (por. Kalton 2009: 135). Problematyczne byłyby zatem dwie ostatnie warstwy, ponieważ ich występowanie oznacza, iż szansa wylosowania konkretnej jednostki zależy od liczby operatów, w których jednostka taka się znajduje. Jednym z częściej stosowanych sposobów rekompensacji nierównych szans losowania jest oczywiście ważenie danych[80]. Szczegółowy opis procedur ustalania wielko-

[80] Wagi przypisane poszczególnym jednostkom definiowane są przy tym jako odwrotności prawdopodobieństw wylosowania jednostek do próby badawczej. Warto wskazać jednak, że prawdopodobieństwo wylosowania obiektu pokrytego przez dwa operaty równe jest prawdopodobieństwu sumy zdarzeń polegających na jego doborze z pierwszego lub drugiego operatu. Ponieważ losowania te odbywają się niezależnie od siebie, to prawdopodobieństwo sumy tych zdarzeń wynosi $P(\mathrm{A} \cup \mathrm{B}) = P(\mathrm{A}) + P(\mathrm{B}) - P(\mathrm{A}) \cdot P(\mathrm{B})$. R. Groves i in. (2004: 88) podają przykładowy sposób wyznaczania wag dla takiego schematu losowania, który wykorzystuje operaty abonentów telefonii stacjonarnej – jako operat główny, a także rejestr adresowy – jako operat uzupełniający. Autorzy ci wychodzą od wyrażonej *explicite* obserwacji, że ponieważ gospodarstwa domowe bez dostępu do telefonu stacjonarnego mają szanse wylosowania wyłącznie z operatu adresowego, natomiast te posiadające telefon stacjonarny zarówno z rejestru abonentów, jak i operatu adresowego, to w próbie badawczej nadreprezentowane byłyby gospodarstwa z dostępem do linii telefonicznej oraz niedoreprezentowane gospodarstwa bez takiego dostępu. Można tam również odnaleźć wzory pozwalające na wyznaczenie wielkości odpowiednich wag (por. Groves i in. 2004: 88). Warto jednak wskazać na pewne specyficzne założenia poczynione w ramach opisywanego przez nich przykładu. Groves i in. (2004: 87) założyli bowiem, iż operat abonentów telefonicznych jest węższy od operatu adresowego, to znaczy że zbiór wszystkich gospodarstw domowych z dostępem do telefonii stacjonarnej zawiera się w operatach adresowych. W szczególnym przypadku może się jednak zdarzyć, że gospodarstwo domowe będzie obecne w wykazie abonentów, ale nie w rejestrze adresowym. Sytuacja taka może mieć miejsce w odniesieniu do nowo zamieszkanych budynków, których operaty adresowe jeszcze nie obejmują. W praktyce nie udałoby się jednak ustalić, czy dane gospodarstwo domowe wylosowane z rejestru abonentów stacjonarnych jest też obecne w wykazie adresowym, czyli niemożliwe byłoby ustalenie prawdopodobieństwa doboru takich gospodarstw domowych. Zapewne z tych właśnie względów bierze się – poczynione przez Grovesa i in. (2004: 87) – założenie o inkluzji operatu telefonicznego przez rejestr adresowy.

ści wag odnaleźć można w licznych pozycjach literaturowych, w tym między innymi w: Kalton (2009: 134), Groves i in. (2004: 88), Bankier (1986: 1074–1075) oraz Kalton i in. (1986: 76–77). Istotne jest to, że estymatory punktowe parametrów populacyjnych przyjmują w takich sytuacjach postać szacunków Horwitza–Thomsona (1952), co skutkuje często przyrostem wariancji wyników w porównaniu do wariancji w zbiorze danych nieważonych[81].

Z koniecznością ustalania wielkości wag będących odwrotnością szans wylosowania jednostek do prób badawczych związane są jednak pewne ograniczenia praktyczne. Na najpoważniejsze z nich zwrócił uwagę G. Kalton (2009: 135) w artykule *Methods for Oversampling Rare Subpopulations in Social Surveys*, podkreślając, że procedura ta wymaga wiedzy o prawdopodobieństwach selekcji każdej jednostki, w ramach każdego operatu, tj. niezależnie od tego, z którego rejestru jednostka taka została pobrana. Mówiąc precyzyjniej, problematyczne okazuje się wyznaczenie prawdopodobieństwa wylosowania jednostek z części wspólnej obu operatów, bowiem informacje o szansach selekcji znane są często wyłącznie w odniesieniu do tych operatów, z których daną jednostkę wylosowano, nieznane pozostają natomiast w obrębie tych rejestrów, w których jednostka figuruje, ale nie została z nich wybrana. Naprzeciw tym ograniczeniom wychodzi metoda opracowana przez Pierre'a Lavallée (1995: 25–32), w której uwzględnia się prawdopodobieństwa selekcji jednostek tylko z tych operatów, z których zostały one pobrane[82]. Propozycję P. Lavallée można zapisać w postaci formuły:

(III.1.) $$w_i \approx \sum_{j=1}^{f} \lambda_{ij} w'_{ij},$$

gdzie:

- $j = \{1,2, \dots, f\}$ jest symbolem oznaczającym *j*-ty operat;
- λ_{ij} jest parametrem zmiennym[83], takim jednak, że dla dowolnej *i*–tej jednostki $\sum_{j=1}^{f} \lambda_{ij} = 1$;
- dla ustalonego *j*–tego operatu: (a) $w'_{ij} = \left(\pi_{ij}\right)^{-1}$, jeżeli *i*–ta jednostka została z niego wylosowana z prawdopodobieństwem równym π_{ij}, natomiast w przeciwnym przypadku: (b) $w'_{ij} = 0$.

Modyfikację tej procedury odpowiednią dla operatów wykorzystujących rejestry numerów telefonicznych odnaleźć można w artykule *Mobile Phone*

---

[81] Skalę tego przyrostu można łatwo wyznaczyć w oparciu o zdefiniowany w II rozdziale monografii miernik *VIF*.

[82] Waga nadana *i*–tej jednostce przeszacowywałaby rzeczywiste szanse selekcji do próby.

[83] W artykule Adama Chu, M. Bricka oraz G. Kaltona (1999: 103–104) pt. *Weights for Combining Surveys Across Time or Space* odnaleźć można sugestię, że parametr λ_{ij} należy wyznaczyć jako proporcję efektywnej wielkości próby dobranej z *j*-tego operatu do wielkości łącznej próby badawczej.

Surveys: Empirical Findings from a Research Project (Häder i in. 2010). Autorzy tego artykuły odwołują się z kolei do pracy S. Gablera oraz Östasa Ayhana (2007: 39–46), w której podano formułę umożliwiającą oszacowanie prawdopodobieństw selekcji jednostek dobieranych z rejestrów abonentów stacjonarnych oraz użytkowników telefonów komórkowych. Estymatory prawdopodobieństw doboru wyrażone są w postaci wzoru:

(III.2.) $$\pi_i \approx k_i^s \frac{n^s}{N^s} \cdot \frac{1}{z_i} + k_i^m \frac{n^m}{N^m}$$ (por. Häder i in. 2010: 15),

gdzie:

- k_i^s oraz k_i^m jest liczbą numerów telefonicznych, które umożliwiają dotarcie do *i*–tej osoby odpowiednio poprzez telefonię stacjonarną oraz mobilną;
- n^s oraz n^m odpowiada liczebności wylosowanych numerów stacjonarnych praz komórkowych;
- N^s oraz N^m jest liczbą wszystkich numerów stacjonarnych oraz mobilnych w populacji objętej badaniem;
- z_i jest wielkością gospodarstwa domowego, do którego należy *i*–ta wylosowana osoba.

Opisane dotąd procedury estymacji wielkości parametrów populacyjnych w oparciu o pomiar prób badawczych losowanych z wielu różnych operatów łączył wspólny mianownik, jakim był estymator Horwitza–Thomsona. W literaturze metodologicznej odnaleźć można jednak również liczne odniesienia do innego typu estymatora zdefiniowanego przez Hermana O. Hartleya (1962, 1974). Podstawową jego właściwością jest zminimalizowanie wariancji w obrębie podzbioru jednostek statystycznych wylosowanych z części wspólnej dwóch (lub więcej) operatów, czyli z tego fragmentu rejestru jednostek populacyjnych, który różnicuje szanse doboru do próby. Szacunki Hartleya wymagają jednak wyznaczenia statystyk punktowych dla każdej części operatu, tj. $\hat{\theta}^A_{A\cap\bar{B}}$, $\hat{\theta}^B_{\bar{A}\cap B}$, $\hat{\theta}^A_{A\cap B}$ oraz $\hat{\theta}^B_{A\cap B}$. Zgodnie z propozycją H.O. Hartleya wartość estymatora można wyznaczać z formuły o ogólnej postaci:

(III.3.) $$\hat{\theta} = \hat{\theta}^A_{A\cap\bar{B}} + \lambda\hat{\theta}^A_{A\cap B} + (1-\lambda)\hat{\theta}^B_{A\cap B} + \hat{\theta}^B_{\bar{A}\cap B},$$

gdzie statystyki częściowe $\hat{\theta}^A_{A\cap\bar{B}}$, $\hat{\theta}^B_{\bar{A}\cap B}$, $\hat{\theta}^A_{A\cap B}$ oraz $\hat{\theta}^B_{A\cap B}$ uwzględniają proporcje poszczególnych warstw jednostek losowanych z podzbiorów operatu *A* oraz *B*, natomiast parametr $\lambda \in \langle 0; 1\rangle$ jest ustalany w taki sposób, aby wariancja estymatorów z części wspólnej operatów była jak najmniejsza. Opis różnych metod służących optymalnemu doborowi parametru λ odnaleźć można w artykule Sharon L. Lohr (2009: 78–84), wspomina o nich także Graham Kalton (2009: 135) oraz Trent D. Buskirk (2008: 212–215). Z kolei w monografii *Survey*

Methodology (por. Groves i in. 2004: 88) podany został przykład zastosowania estymatora Hartleya w sytuacji pokrycia populacji operatami abonentów telefonii stacjonarnej uzupełnionymi o adresowy spis gospodarstw domowych[84].

III.3.4. Procedury ograniczania błędów operatów doboru prób badawczych – podsumowanie

Działania podejmowane przez badaczy w celu ograniczenia negatywnych konsekwencji wynikających z uchybień operatów doboru prób badawczych przyjmują najczęściej postać czterech strategii: (a) ignorowania problemu (tj. jego przemilczania), (b) redefiniowania populacji docelowych (tj. przeformułowywania założeń badawczych w celu ich dostosowania do jakości dostępnych operatów), (c) postbadawczego ważenia danych (tj. dostosowania rozkładów wylosowanej próby badawczej do znanych rozkładów wybranych zmiennych w całej populacji) oraz (d) dążenia do poprawy jakości operatów, (tj. dostosowania rejestru doboru próby do potrzeb wynikających z celów badania, a nie odwrotnie) włącznie z ujednoliceniem nierównych szans selekcji jednostek populacji do prób badawczych. Co oczywiste, najbardziej pożądane wydaje się przyjęcie czwartej strategii. Jednoznacznym tego potwierdzeniem wydaje się dążenie metodologów badań sondażowych do wypracowania procedur ograniczających błędy operatów. Doskonale wyraził to Biemer wraz z Lybergiem, którzy w monografii *Introduction to Survey Quality*, w części poświęconej zagadnieniom operatów, stwierdzają, iż:

> być może najbardziej efektywnym podejściem do problemu ograniczania błędu systematycznego [na skutek błędów operatów – P.J.] jest próba naprawienia operatu poprzez usunięcie duplikatów [wielokrotnych informacji o tych samych jednostkach – P.J.] lub jednostek nieprawidłowo włączonych do operatu, zwiększenie pokrycia poprzez działania terenowe mające na celu identyfikację jednostek pominiętych w operacie. Procedury te są oczywiście kosztowne, ale instytucje badawcze […] mogą te koszty amortyzować w wielu badaniach w dłuższym okresie. (Biemer i in. 2003: 77)

Niemniej jednak należy również zauważyć, że w pewnych sytuacjach ignorowanie błędu niepełnego pokrycia nie jest niczym nadzwyczajnym. Wystarczy

---

[84] Przypomnieć można, że Groves i in. (2004: 87) zakładali, iż gospodarstwa domowe znajdujące się w operacie abonentów telefonicznych stanowią podzbiór gospodarstw domowych wykazanych w rejestrach adresowych. Abstrahując od tego, czy tak rzeczywiście musi być, warto jedynie zauważyć, że przyjęcie takiego założenia oznacza, iż zbór jednostek $\bar{A} \cap B$ będzie pusty. Stąd podany przez Grovesa i in. (2004: 88) przykład estymatora Hartleya dla prób losowanych z operatów telefonicznych oraz adresowych składa się z trzech, a nie jak we wzorze (III.3.) – z czterech estymatorów częściowych.

podać przykłady państw uczestniczących w projekcie ESS, w których badacze dysponują prawie „idealnymi" operatami doboru prób badawczych, na przykład urzędowymi rejestrami ludności. Istotnie bowiem, w takich sytuacjach bardziej rozsądne wydaje się zignorowanie błędu pokrycia, niż podejmowanie jakichś kosztownych działań mających na celu wyszukanie oraz zidentyfikowanie w sumie niewielkiej części całej populacji. Podobnie można zauważyć, że redefiniowanie populacji nie jest też czymś wyjątkowym. Przypomina ono działania, które badacz podejmuje, wykluczając pewne kategorie jednostek z populacji objętej badaniem, definiując tym samym populację docelową. Ponieważ osoby z takich wyłączonych kategorii jednostek stanowią niewielki odsetek całej populacji, to redefiniowanie populacji poprzez ich wykluczenie nie jest w stanie, w sposób statystycznie istotny, zmienić uzyskiwanych wyników pomiaru.

III.4. Możliwości wykorzystania rejestrów administracji publicznej w polskiej socjologii sondażowej

Zanim omówione zostaną szczegółowe konsekwencje wykorzystania pewnych typów operatów doboru prób badawczych (imiennych lub zespołowych), warto najpierw przyjrzeć się możliwościom wykorzystania dostępnych w Polsce rejestrów urzędowych – PESEL oraz TERYT – jako operatów doboru sondażowych prób badawczych.

Przynajmniej do końca 2016 roku operatem umożliwiającym losowanie prób imiennych pozostanie Powszechny Elektroniczny System Ewidencji Ludności (tzw. rejestr PESEL)[85], którego gestorem jest Ministerstwo Spraw Wewnętrznych. Z kwestii podstawowych ważne jest przede wszystkim to, że baza PESEL uaktualniana jest na bieżąco[86] i zawiera dane wszystkich osób przebywających na stałe na terytorium Rzeczypospolitej Polskiej, tj. zameldowanych na pobyt stały lub czasowy powyżej 3 miesięcy. Źródłem informacji pozyskiwa-

[85] Tekst ten jest swego rodzaju epitafium dla rejestru PESEL w jego dotychczasowej formie umożliwiającej wylosowanie imiennych prób ludności. Zgodnie z ustawą z dnia 24 września 2010 r. (z późn. zm.) o ewidencji ludności, począwszy od 1 stycznia 2017 roku zniesiony zostaje obowiązek meldunkowy (art. 74). Co więcej, zlikwidowane zostają rejestry mieszkańców i rejestry zamieszkania cudzoziemców (art. 75); nie będzie się również gromadzić danych znajdujących się obecnie w rejestrze PESEL, poza serią, numerem oraz datami ważności dowodów osobistych i paszportów (art. 76).

[86] Odpowiedni organ gminy prowadzący ewidencję ludności przekazuje (w terminie nie dłuższym niż 5 dni od daty zgłoszenia) informację o zmianach meldunkowych w formie elektronicznej do wojewody. Wojewoda przekazuje informację zwrotną potwierdzającą wpisanie zmiany oraz niezwłocznie (w terminie do 3 dni) przekazuje informacje do MSW, gdzie przeprowadzana jest aktualizacja cyfrowej wersji bazy PESEL (por. Nowak i in. 2007: 27).

nych na potrzeby operatu PESEL są odpowiednie organy gminy, które na mocy ustawy z dnia 10 kwietnia 1974 roku o ewidencji ludności i dowodach osobistych prowadzą gminne zbiory meldunkowe (Dz. U. z 1974 roku, Nr 139, poz. 993, z późn. zm.). W zależności od zasięgu terytorialnego prowadzonych badań instytucjami formalnie upoważnionymi do udostępniania danych z rejestru PESEL (dla ośrodków badawczych odpłatnie) są urzędy gmin, wojewodowie lub Minister Spraw Wewnętrznych. W ocenie przydatności repozytorium PESEL do losowania prób badawczych istotne jest przede wszystkim to, że poza numerem porządkowym, imieniem i nazwiskiem oraz adresem zameldowania (stałego lub czasowego) istnieje również możliwość pozyskania innych danych, niezwykle ważnych na etapie projektowania, losowania oraz realizacji prób badawczych. Są to między innymi: (a) daty i miejsca urodzenia, (b) płeć, (c) obywatelstwo oraz (d) stan cywilny.

Niezwykle interesującej egzemplifikacji ograniczeń związanych z możliwością zastosowania rejestru PESEL w badaniach sondażowych dostarcza opracowanie Sawińskiej i in. (2009: 6–11). Chociaż przeprowadzona analiza poświęcona była wyłącznie różnicom pomiędzy zakresem populacji pokrytej PESEL oraz zakresem populacji będącej przedmiotem badania w Europejskim Sondażu Społecznym, to jednak, wyłączając poza nawias dyskusji pewne problemy specyficzne dla ESS, można w przywoływanym materiale odnaleźć również pouczające informacje oraz wnioski o charakterze ogólnym. Dla przykładu, autorzy tego opracowania ukazują, że dane zawarte w rejestrze PESEL pozwalają na wylosowanie osób bezdomnych, ale tylko tych, które są gdzieś zameldowane[87]. Problematyczną kategorię stanowią jednak przede wszystkim osoby przebywające w kraju, lecz niedostępne w miejscu zameldowania – rejestr PESEL zawiera wyłącznie informacje o aktualnym adresie stałego lub czasowego zameldowania, nie podaje informacji o miejscu faktycznego pobytu. Doświadczenia ESS_PL ukazują, że problem ten dotyczy dość znacznego odsetka respondentów (mniej więcej 8 pp. wszystkich jednostek próby w ESS_PL–2008) przy czym, jedynie w połowie przypadków ankieterom udaje się (w trakcie terenowej realizacji próby badawczej) ustalić faktyczne miejsce zamieszkania takich osób (por. Sawińska i in. 2009: 9). Zatem, pomimo iż, formalnie rzecz biorąc, owe osoby są w operacie wykazane (nie ma w stosunku do nich błędu niepełnego pokrycia), to jednak dotarcie do nich jest niemożliwe. Stąd w praktyce sondażowej rejestr PESEL będzie charakteryzował się – mimo wszystko – niepełnym pokryciem badanej populacji. Ostatnią z problematycznych kategorii są osoby, które wyjechały poza granice kraju z zamiarem pozostania tam na

[87] Problemy wynikające z badania metodą sondażową osób bezdomnych stanowią osobny przedmiot rozważań metodologicznych (por. na przykład Martin i in. 1997: 59–73, Laska i in. 1993: 209–220).

stałe, ale jednocześnie faktu tego nie zgłosiły odpowiednim organom administracji publicznej. Szacunki przeprowadzone na podstawie ESS–2008 ukazują, iż odsetek takich osób wynosi prawie 6 pp. (por. Sawińska i in. 2009: 11). Przyjmując terminologię błędów, można wszystkie takie osoby zaklasyfikować do warstwy jednostek nadmiarowo pokrytych przez operat doboru próby. Co ważne jednak, pomimo pewnych niedoskonałości rejestr PESEL pozostaje najczęściej wykorzystywanym operatem doboru prób imiennych w sondażach o charakterze naukowym[88].

Podczas gdy repozytorium PESEL umożliwia wybór prób imiennych, to operatem pozwalającym na losowanie prób adresowych (ulic, budynków mieszkalnych lub mieszkań) jest Krajowy Rejestr Urzędowy Podziału Terytorialnego Kraju (tzw. TERYT) prowadzony przez GUS. Ważne jest to, że znaczenie operatu TERYT wzrośnie wraz ze zniesieniem obowiązku meldunkowego oraz – *de facto* – z formalną likwidacją PESEL. Podstawą prawną funkcjonowania bazy TERYT jest ustawa z dnia 29 czerwca 1995 roku o statystyce publicznej (Dz. U. z 1985 roku, Nr 88, poz. 439 z późn. zm.). W obecnym kształcie repozytorium to zostało wprowadzone rozporządzeniem Rady Ministrów z dnia 15 grudnia 1998 roku w sprawie szczegółowych zasad prowadzenia, stosowania i udostępniania krajowego rejestru urzędowego podziału terytorialnego kraju oraz związanych z tym obowiązków organów administracji rządowej i jednostek samorządu terytorialnego (Dz. U. z 1998 roku, Nr. 157, poz. 1031 z późn. zm.). Rejestr TERYT składa się z pięciu podstawowych komponentów o hierarchicznej strukturze, w tym z: (1) systemu identyfikatorów i nazw jednostek podziału terytorialnego kraju na województwa, powiaty oraz gminy (tzw. TERC)[89], który pozostaje zbieżny z nomenklaturą Klasyfikacji Jednostek Terytorialnych do Celów Statystycznych[90]. Drugim komponentem rejestru

---

[88] Wystarczy przyjrzeć się informacjom zamieszczonym w repozytorium Archiwum Danych Społecznych prowadzonym przez Instytut Studiów Społecznych UW oraz Instytut Filozofii i Socjologii PAN, aby zauważyć, że w projektach badawczych opartych na próbach imiennych rejestr PESEL wykorzystywany był w zdecydowanej większości badań (ze zbioru wszystkich 51 projektów badawczych zamieszczonych w ADS [stan na dzień 27.11.2013 r.], baza PESEL użyta została w 46). Specyfika pięciu pozostałych przedsięwzięć badawczych (stanowią one znaną serię lokalnych badań mieszkańców Łodzi z lat 60., 70. oraz 80. XX wieku, prowadzonych pod kierunkiem Włodzimierza Wesołowskiego, Kazimierza M. Słomczyńskiego oraz Krystyny Janickiej) wymagała jednak od badaczy doboru próby z populacji ograniczonej do mieszkańców jednego miasta, zatem operatem uczyniono wykazy wyborców oraz spisy mieszkańców Łodzi (por. Wesołowski i in. 1965, 2004; Słomczyński i in. 1994).

[89] Baza TERC umożliwia ponadto wyróżnienie: (a) miast na prawach powiatu, (b) gmin miejskich, miejsko–wiejskich oraz wiejskich, (c) miast i obszarów miejskich w gminach miejsko-wiejskich oraz (d) dzielnic i delegatur w gminach miejskich.

[90] Tzw. podział NUTS (skrót od *Nomenclature of Teritorial Units for Statistics*) jest wystandaryzowaną formą analizy jednostek terytorialnych na różnych poziomach agregacji danych (regionalnym – NUTS1, wojewódzkim – NUTS2, podregionalnym – NUTS3, powiatowym – NUTS4 oraz gminnym – NUTS5) stosowaną we wszystkich krajach członkowskich UE. Podstawę prawną sta-

TERYT jest (2) system identyfikatorów i nazw miejscowości (tzw. SIMC) zawierający urzędowe nazwy miejscowości oraz przynależność miejscowości do gmin, powiatów oraz województw. Trzecim natomiast jest (3) system rejonów statystycznych i obwodów spisowych (tzw. BREC), obejmujący podział terytorialny kraju na rejony statystyczne i obwody spisowe utworzone dla potrzeb realizacji spisów powszechnych oraz prowadzenia badań reprezentatywnych (losowaniu podlegają wówczas odpowiednio: rejon statystyczny, obwód spisowy oraz numer porządkowy adresu mieszkalnego w ramach wylosowanego obwodu spisowego). Numery porządkowe adresów mieszkalnych z bazy BREC mają swoje odpowiedniki w (4) systemie identyfikacji adresowej ulic, nieruchomości, budynków i mieszkań (tzw. NOBC). Elementem składowym systemu NOBC jest z kolei ostatni komponent rejestru TERYT, to znaczy (5) centralny katalog ulic (tzw. ULIC) zawierający identyfikatory nazw ulic zgodne z brzmieniem uchwał odpowiednich organów administracyjnych o ich nadaniu. Operat doboru prób adresowych do badań sondażowych konstytuują zatem dwa główne komponenty rejestru TERYT: BREC oraz NOBC.

Rozważając możliwość wykorzystania repozytorium TERYT w losowaniu sondażowych prób badawczych, należy przede wszystkim zwrócić uwagę na wysoką wiarygodność baz danych adresowych prowadzonych przez GUS. Nieco więcej problemów stwarza natomiast aktualność danych. Zgodnie z informacjami zawartymi na stronach internetowych biuletynu informacji publicznej GUS-u, każdy z komponentów rejestru TERYT podlega aktualizacji przynajmniej raz w roku, przy czym systemy TERC, SIMC, ULIC oraz NOBC (ale tylko w części dotyczącej nazw miejscowości, ulic oraz numerów porządkowych nieruchomości i budynków) uaktualniane są na bieżąco (w cyklach tygodniowych), system NOBC (w kluczowej dla losowania prób badawczych części dotyczącej przyrostów i ubytków budynków i mieszkań) uaktualniany jest kwartalnie, z kolei system BREC podlega aktualizacji co najmniej raz w roku. Co za tym idzie, repozytorium NOBC nie cechuje się pełnym pokryciem budynków oraz mieszkań (rejestr zawiera z całą pewnością adresy już niezamieszkane oraz pomija adresy niezamieszkane w momencie uaktualnienia operatu, ale już zamieszkane w okresie losowania oraz realizacji próby), jednakże wydaje się, iż skala tych uchybień nie jest znacząca, a jakość operatu pozostaje – mimo wszystko – wysoka. Ważne jest również to, że ponieważ architektura rejestru TERYT zawiera uporządkowane dane adresowe, to umożliwia zastosowanie

---

nowią**:** rozporządzenie WE nr 1059/2003 z dnia 26 maja 2003 r.; załącznik nr 1 do rozporządzenia Wspólnoty Europejskiej (WE) o nr 1059/2003**;** rozporządzenie Komisji WE nr 11/2008 z dnia 8 stycznia 2009 r**.** Od 1 stycznia 2008 roku, zgodnie z rozporządzeniem Rady Ministrów z dn. 14 listopada 2007 r. (Dz. U. z 2007 roku, Nr 214, poz. 1573), wprowadzono zmianę w Nomenklaturze Jednostek Terytorialnych do Celów Statystycznych, polegającą na zwiększeniu do 66 liczby podregionów w Polsce.

metody przedziałów półotwartych do redukcji błędu niepełnego pokrycia populacji operatem doboru próby[91].

Niezwykle ważnym kryterium oceny przydatności rejestru TERYT jako operatu doboru sondażowych prób badawczych jest też dostępność danych z tego repozytorium. Podobnie jak w przypadku operatu PESEL, wykorzystanie rejestru TERYT pozostaje całkowicie zgodne z prawem, w tym przede wszystkim z ustawą o ochronie danych osobowych z dnia 29 sierpnia 1997 roku (Dz. U. z 1997 roku, Nr 133, poz. 883 z późn. zm.). Należy przy tym wskazać, że informacje z systemów TERC, SIMC oraz ULIC udostępniane są bezpłatnie za pośrednictwem strony internetowej GUS, natomiast dane zawarte w systemie BREC i NOBC podlegają każdorazowej subskrypcji za pośrednictwem centrali lub wojewódzkich oddziałów GUS w zależności od zasięgu terytorialnego badanej populacji. Innymi słowy, badacze sondażowi mają nieograniczony oraz bezpłatny dostęp tylko i wyłącznie do operatu umożliwiającego wylosowanie próby losowej ulic (co sprowadza się *de facto* do możliwości losowania próby zgodnej ze schematem doboru prób przestrzennych); natomiast wtedy, gdy losowaniu podlegają adresy budynków lub mieszkań, udostępnienie pożądanego przez badaczy wykazu adresowego odbywa się już na zasadach komercyjnych.

Doskonałym przykładem wykorzystania rejestru TERYT (oraz jego wcześniejszych wersji) w reprezentatywnych badaniach o charakterze społecznym okazuje się siedem pierwszych edycji Polskiego Generalnego Sondażu Społecznego (PGSS 1992 – PGSS 2002) [92]. W dokumentacji metodologicznej tych badań – zamieszczonych na stronach internetowych Archiwum Danych Społecznych oraz bezpośrednio w witrynie poświęconej badaniu PGSS – odnaleźć można szczegółowe informacje odnośnie operatów doboru prób badawczych oraz schematów losowania respondentów w obrębie wybranych mieszkań (por. Cichomski i in. 2009: 3–4). Trzeba przy tym zaznaczyć, że procedura doboru próby badawczej miała w tych badaniach charakter wielostopniowy. Zastosowano przy tym odmienne schematy doboru próby dla każdego z trzech segmentów populacji wyodrębnionych z uwagi na wielkomiejski, miejski lub wiejski charakter jednostek terytorialnych. W każdej warstwie segmentu wielkomiejskiego (którą tworzyły całe miejscowości lub też dzielnice w pięciu największych miastach Polski: w Warszawie, Krakowie, Łodzi, Poznaniu oraz we Wrocławiu) próbę gospodarstw domowych losowano bez zwracania, z jednakowym prawdopodobieństwem, proporcjonalnie do wielkości każdej warstwy, otrzymując tym samym prostą próbę losową gospodarstw domowych. W war-

[91] Jak wiadomo, procedura ta wymaga, aby każdemu wylosowanemu adresowi przyporządkować adres następujący po nim. Autorowi nie są znane jednak żadne polskie badania, w których zastosowano by procedurę przedziałów półotwartych.

[92] W badaniach PGSS zrealizowanych w 2005 roku operatem losowania był już rejestr PESEL (por. Cichomski i in. 2009: 3–4).

stwach miejskich (cztery warstwy miast wyróżnione z uwagi na liczebność mieszkańców) losowano w pierwszej kolejności (w sposób zwrotny oraz proporcjonalnie do liczby gospodarstw domowych w danej warstwie miejskiej) pewną liczbę miast, a następnie w każdym wylosowanym mieście wybierano wymaganą liczbę mieszkań. Z kolei w każdej warstwie segmentu wiejskiego przeprowadzono losowanie trójstopniowe: najpierw dobierano gminy wiejskie proporcjonalnie do liczby mieszkań w tych gminach, następnie w gminie losowano jeden rejon spisowy, z którego w ostatnim etapie wybierano próbę mieszkań ze zbioru wszystkich gospodarstw domowych znajdujących się w granicach rejonu spisowego. Warto zauważyć jednak, że chociaż wykorzystany schemat doboru próby w PGSS sprawia, że „każde z gospodarstw domowych znajdujące się w operacie losowania ma takie samo prawdopodobieństwo bycia wylosowanym, […] [to jednak – P.J.] zastosowana metoda doboru próby, uwzględnia na wszystkich etapach liczebność gospodarstw domowych, nie zaś liczbę ludności" (Cichomski i in. 2009: 3). Oznacza to, że jedyną poważną konsekwencją zastosowania operatu adresowego w zamian imiennego, pozostało w badaniu PGSS zróżnicowanie szans selekcji jednostek indywidualnych w obrębie jednostek zespołowych.

III.5. Operaty doboru prób badawczych – komplikacje metodologiczne oraz konsekwencje praktyczne na przykładzie Europejskiego Sondażu Społecznego

Prowadzona do tej pory dyskusja skoncentrowana była w głównej mierze na czterech źródłach błędów wynikających z ułomności operatów doboru prób badawczych, a także na wpływie tych uchybień na losowy oraz systematyczny komponent całkowitego błędu pomiaru. Scharakteryzowano również procedury służące poprawie jakości operatów, w tym klasę metod wykorzystywanych w celu redukcji niepełnego pokrycia populacji operatami losowania, a także przedstawiono możliwe do zastosowania w polskiej socjologii sondażowej administracyjne rejestry ludności oraz adresów budynków mieszkalnych. Rozważania metodologiczne warto teraz odnieść do świata empirii, nadając tym samym prowadzonym dotąd studiom zdecydowanie bardziej praktyczny charakter. Trudność polega jednak na tym, że w zdecydowanej większości projektów badawczych kwestie dotyczące jakości operatów doboru próby nie są w ogóle przedmiotem jakiegokolwiek namysłu[93]. Mówiąc dobitniej, niewiele

[93] Zwracano już zresztą na to uwagę w rozdziale II. Przywoływana była wówczas konstatacja F. Sztabińskiego, który w odniesieniu do błędów operatów mówił, iż są one „najrzadziej uwzględ-

jest takich przedsięwzięć, w których – niezależnie od użytych procedur oraz technik zbierania danych – podejmowano by wysiłki na rzecz określenia relacji zachodzących pomiędzy jakością operatu a jakością badania.

Jednym z nielicznych projektów, w którym podejmowane są takie zagadnienia, jest Europejski Sondaż Społeczny. Poza celem substantywnym w programie tym realizuje się także cele metodologiczne i praktyczne, z których bodaj najważniejszym jest wypracowanie standardów przygotowania i realizacji naukowych badań sondażowych (por. Domański 2006: 30–31; Sztabiński P.B. 2004: 27). W świetle prowadzonych obecnie rozważań ważne jest to, że po każdej rundzie projektu ESS publikowane są raporty dokumentujące metodologiczną warstwę badań. Zawierają one bogaty zestaw informacji o schematach losowania prób badawczych, efektywnych wielkościach tych prób, jednostkach niedostępnych, jakości prowadzonego pomiaru, a także o innych ważnych aspektach związanych z realizacją reprezentatywnych badań społecznych. Choć dokumenty te zawierają niezwykle wartościowe informacje o przebiegu każdej rundy badań ESS, to akurat w odniesieniu do problematyki operatów doboru prób badawczych okazują się niepełne. Wprawdzie na stronach internetowych projektu znaleźć można dwa dokumenty poświęcone procesowi próbkowania (por. *ESS1 Sampling report 2002*, *ESS2 Sampling report 2004*), jednak dotyczą one wyłącznie pierwszej (ESS1–2002) oraz drugiej (ESS2–2004) rundy tych badań. Mimo wszystko, nawet ten niepełny zakres danych okazuje się lekturą niezwykle pouczającą i pozwalającą na ukazanie relacji pomiędzy rodzajem operatu doboru jednostek populacji oraz typem próby badawczej.

Przechodząc do szczegółowej analizy metodologii projektu ESS, warto rozpocząć od przypomnienia tego, o czym już mówiono wcześniej, że we wszystkich krajach obowiązywała ta sama definicja populacji będącej przedmiotem badania, co było oczywistym wymogiem porównawczej natury projektu. Deskryptywny opis populacji odnaleźć można na stronach internetowych ESS, a także w wielu publikacjach naukowych poświęconych tym badaniom. Odnosząc się do jednego z takich opracowań należy przypomnieć, iż na badaną populację składały się „wszystkie osoby w wieku 15 lat i więcej zamieszkujące w prywatnych gospodarstwach domowych, w granicach kraju, niezależnie od narodowości, obywatelstwa, języka lub statusu prawnego" (Lynn i in. 2007: 110). Chociaż definicja ta odnosiła się do wszystkich krajów, to jednocześnie dało się zaobserwować istotne różnice w sposobie określania populacji docelowych (czyli takich, których badanie w rzeczywistości, a nie tylko w zamierzeniach, dotyczyło). Ponieważ kwestie te analizowane były już wcześniej w ramach opi-

---

nianym rodzajem błędów zawartych w wynikach badania" (Sztabiński F. 2011: 50), jak również wyniki studiów Kasprzyka i in. (2003: 343–363) prowadzące do podobnych wniosków.

su rozróżnień terminologicznych związanych z różnymi typami populacji, to zostaną one pominięte w toku obecnej dyskusji. Uwaga skupiać będzie się więc wyłącznie na właściwościach operatów doboru prób badawczych oraz wynikających z tego konsekwencjach praktycznych.

W tabeli A1. zamieszczonej w aneksie książki zestawiono informacje o populacjach docelowych wraz z charakterystyką operatów służących losowaniu prób badawczych. Materiał ten pozwala na przypisanie wszystkich państw biorących udział w ESS do pięciu kategorii, wyróżnionych z uwagi na dwa kryteria, to znaczy typ zastosowanego operatu (imienny *vs*. zespołowy) oraz rodzaj losowanej próby badawczej (imienna, adresowa, przestrzenna)[94]. Klasyfikacja ta pozwala podzielić wszystkie kraje uczestniczące w ESS na takie, w których: (a) istniały oraz dostępne były operaty jednostkowe służące losowaniu prób imiennych, (b) istniały oraz dostępne były operaty imienne, choć z uwagi na niepełne pokrycie populacji docelowej wykorzystywano je do losowania adresowych prób osób, (c) istniały oraz dostępne były adresowe wykazy gospodarstw domowych wykorzystywane do losowania prób adresowych, (d) istniały oraz dostępne były adresowe rejestry budynków mieszkalnych wykorzystane do losowania prób adresowych, a także takie, (e) dla których nie istniały lub niedostępne były jakiekolwiek wiarygodne wykazy imienne lub rejestry adresowe. To z kolei oznaczało konieczność realizacji tzw. prób przestrzennych.

Zestawienie rodzajów operatów wykorzystywanych przez krajowe ośrodki realizujące narodowe komponenty ESS znakomicie obrazuje relacje zachodzące między populacją będącą przedmiotem badania, populacją docelową oraz populacją pokrytą przez operat losowania. Uwidacznia także istotę oraz skalę problemów, na jakie narażone jest prowadzenie badań sondażowych w sytuacji niedostępności imiennych rejestrów obywateli. Chociaż podstawowym wyzwaniem w badaniach międzykrajowych staje się zastosowanie operatów zapewniających porównywalność otrzymywanych wyników (por. Słomczyński 1999: 242–243), na przykład poprzez wykorzystanie rejestrów dających podobne błędy systematyczne, to jednak, abstrahując od tych komplikacji fundamentalnych dla badań porównawczych, ale nieistotnych w sondażach prowadzonych w jednym kraju (chyba że realizuje się badania diachroniczne losując próby

[94] W niezwykle interesującym artykule pt. *Methods for Achieving Equivalence of Samples in Cross–National Surveys: The European Social Survey Experience* opublikowanym w 2007 roku w czasopiśmie „Journal of Official Statistics" odnaleźć można podobną klasyfikację operatów wykorzystanych w państwach uczestniczących w badaniach *ESS* (por. Lynn i in. 2007: 110). Autorzy artykułu stosują jednak nieco inne kryteria klasyfikacji uwzględniające wyłącznie rodzaj zastosowanego operatu doboru próby. Wyróżniają cztery grupy państw, to znaczy takie, w których istnieją i są możliwe do zastosowania (a) imienne operaty jednostkowe, (b) adresowe wykazy gospodarstw domowych oraz (c) adresowe wykazy budynków mieszkalnych. Natomiast czwartą kategorię stanowią te kraje, w których operaty (a), (b) oraz (c) nie istnieją lub nie ma możliwości ich wykorzystania.

z różnych typów operatów), należy przede wszystkim wskazać, że w znacznej części państw uczestniczących w projekcie ESS badacze dysponowali – mimo wszystko – rzetelnymi oraz wiarygodnymi rejestrami urzędowymi dającymi możliwość losowania prób imiennych. We wszystkich takich krajach wskaźniki pokrycia populacji okazywały się pełne lub prawie pełne. Ponadto, nawet jeśli nie mówiono o tym wprost lub *explicite* nie napisano tego w raporcie, rejestry administracyjne, jeśli tylko istniały, traktowano jako najbardziej wiarygodne źródło danych o członkach populacji docelowych. Co więcej, szczegółowa analiza dokumentacji metodologicznej ukazuje, że we wszystkich krajach, w których badacze mieli możliwość dostępu do urzędowych jednostkowych rejestrów ludności, nie podejmowano już żadnych dodatkowych działań zwiększających pokrycie populacji operatami imiennymi. Jedynym wyjątkiem od tej reguły są badania realizowane w Irlandii oraz we Włoszech. Jak już jednak wiadomo, ma się tutaj do czynienia z przypadkiem szczególnym, bowiem rejestry imienne zawierały dane o osobach należących do kohorty wiekowej 18+ i nie mogły być stosowane do losowania próby z populacji 15+. Wykorzystano je zatem wyłącznie do doboru gospodarstw domowych (we Włoszech) lub budynków mieszkalnych (w Irlandii)[95]. Innymi słowy, w celu zwiększenia pokrycia populacji docelowych dostępnymi w tych krajach rejestrami wyborców, badacze decydowali się na wykorzystanie prób adresowych w zastępstwie prób imiennych.

W badaniach ESS operaty adresowe stanowiły zresztą oddzielną i ważną klasę rejestrów służących doborowi prób badawczych. Wprawdzie losowaniu podlegały wówczas nie jednostki indywidualne, ale ich zespoły, na przykład gospodarstwa domowe lub budynki mieszkalne, to jednak wobec braku dostępu do operatów imiennych repozytoria adresowe umożliwiały pośrednie pokrycie jednostek populacji docelowej poprzez losowanie wielostopniowe – konkretne osoby uczestniczące w wywiadach podlegały wtórnemu doborowi dopiero w obrębie wylosowanego gospodarstwa domowego lub budynku mieszkalnego. Co prawda, przy zastosowaniu doboru wielostopniowego konieczne okazuje się zazwyczaj ważenie rekompensujące nierówne prawdopodobieństwa selekcji jednostek indywidualnych w obrębie jednostek zespołowych, lecz jest to koszt ponoszony niejako z konieczności (ryzyko systematycznego błędu pokrycia minimalizuje się, wybierając operat wymagający zwiększenia liczebności próby). Rodzi to natural-

[95] Chociaż w Irlandii rejestr wyborców daje możliwość wylosowania danych adresowych budynków mieszkalnych, to jednak ma on właściwości zbliżone do operatu gospodarstw domowych. Zgodnie z danymi zamieszonymi w raporcie charakteryzującym krajowe próby badawcze z I rundy ESS okazuje się, że w Irlandii istnieje bardzo mały odsetek budynków wielomieszkalnych (por. *ESS1 Sampling report 2002*: 20). Jest to korzystne o tyle, że wybór adresu budynku wielomieszkalnego oznacza konieczność dodatkowego losowania gospodarstwa domowego w obrębie bloku, co zwiększa wariancję wyników. Ponieważ jednak bloków takich jest bardzo mało, to również konsekwencje zastosowania operatu adresowego budynków są niewielkie w porównaniu do konsekwencji losowania próby gospodarstw domowych.

nie oddzielne komplikacje, mające swoje źródło w konieczności sporządzania spisu członków wylosowanych gospodarstw domowych oraz zastosowania zrandomizowanego doboru przedstawiciela wybranego gospodarstwa.

W literaturze metodologicznej odnaleźć można kilka klas metod służących wewnątrzzespołowej selekcji jednostek. Przynajmniej teoretycznie, wszystkie takie procedury powinny zapewnić możliwość przeprowadzenia doboru w sposób zrandomizowany i chociaż w niewielu przypadkach warunek ten zostaje spełniony – większość z proponowanych metod jest quasi-losowa lub nawet nielosowa – to wymaga się bezwzględnie, aby dobór przebiegał w sposób zobiektywizowany, tj. niezależny od osoby ankietera dokonującego wyboru respondenta. Świetnie wyraził to Zbigniew Sawiński (2005), który w siódmym rozdziale podręcznika *Fieldwork jest sztuką*, w części poświęconej schematom doboru losowego w adresowych próbach osób, podkreśla, iż:

> Wybór konkretnej osoby w ramach gospodarstwa dokonywany jest zawsze za pomocą procedury, która pozwala na zachowanie pełnego obiektywizmu. Dokonany wybór musi być niezależny od tego, który z ankieterów będzie realizował wywiad pod wskazanym adresem. Nie może zatem dopuszczać żadnej swobody ani stwarzać żadnych preferencji co do wyboru pewnych osób, a pominięcia innych. Co więcej, wybór konkretnej osoby nie może również zależeć od momentu, w którym ankieter trafił pod wskazany adres. A w szczególności od tego, którzy z domowników byli w tym czasie obecni, którzy zaś nie. (Sawiński 2005: 88–89)

Warto zatem rozpocząć od tego, że bodaj najbardziej znaną metodą wewnątrzzespołowej selekcji jednostek pozostaje tzw. *procedura obiektywnego doboru respondentów w gospodarstwie domowym*, która w metodologii badań sondażowych funkcjonuje pod nazwą siatki Kisha, od nazwiska jej autora. Leslie Kish opublikował założenia owej metody w 1949 roku w czasopiśmie „Journal of the American Statistical Association". Ponieważ jej podstawy są doskonale znane, wystarczy jedynie przypomnieć, że pierwotnie procedura ta polegała na uszeregowaniu dorosłych członków wylosowanych gospodarstw domowych według wieku (w podziale z uwagi na płeć) oraz zastosowaniu – z ustaloną częstotliwością – jednej z sześciu (lub jednej z ośmiu) tablic, w których, dla określonej liczebności danego gospodarstwa domowego, wskazane były numery porządkowe osób, które należało wybrać do próby (por. Kish 1949: 383–385). Co do istoty, obecnie stosowane modyfikacje tej procedury niewiele różnią się od swojego pierwowzoru.

Chociaż empiryczne implementacje siatki Kisha dały obiecujące rezultaty (por. Smith i in. 1995: 33–38)[96] – i to nawet w porównaniu z procedurami we-

[96] W artykule z 1995 roku pt. *Increasing Response Rates in Telephone Surveys: A Randomized Trial* opublikowanym w czasopiśmie „Journal of Public Health" odnaleźć można empiryczną wery-

wnątrzzespołowej selekcji, które uważa się powszechnie za znacznie mniej ingerujące w prywatność respondentów (por. Yan 2009: 6143–6144; Gaziano 2005: 124–157; Czaja i in. 1982: 381–385), to jednak wielu autorów zwraca przede wszystkim uwagę na potencjalne negatywne konsekwencje jej zastosowania. Krytyka (szczególnie w odniesieniu do możliwości zastosowania procedury Kisha w wywiadach telefonicznych) koncentrowała się głównie na tym, iż konieczność sporządzania składu osobowego wylosowanych gospodarstw domowych jest niezwykle czasochłonna, wzbudza poczucie „nachalności" ze strony ankietera, obniża poziom poczucia bezpieczeństwa osoby udzielającej informacji o członkach wylosowanej jednostki mieszkalnej, prowadzi do wzrostu odmów udziału w badaniu lub przerwania fazy aranżacji wywiadu, co w konsekwencji przekłada się na obniżenie wskaźników realizowalności próby (por. Binson i in. 2000: 53–59; Oldendick i in. 1988: 307–318; O'Rourke i in. 1983: 428–432). Innymi słowy, jak ukazuje Cecilie Gaziano w niezwykle interesującym artykule poświęconym procedurom doboru jednostek z wylosowanych do próby zespołów:

> Pomimo że metody probabilistyczne są preferowane, prowadzą one zazwyczaj do wzrostu liczby jednostek niedostępnych. Quasi-probabilistyczne oraz nieprobabilistyczne metody zostały opracowane po to, aby zwiększyć wskaźniki kooperacji z ankieterem oraz obniżyć koszty, pomimo utraty zalet wynikających z losowości selekcji. (Gaziano 2005: 124)

Wprawdzie owe alternatywne (wobec procedury Kisha) metody selekcji respondentów z wylosowanych do próby zespołów różnią się między sobą pod wieloma względami (także takimi o charakterze substantywnym)[97], to jednak łączy je wspólny mianownik, a mianowicie, dążenie do takiego sformułowania procedury, aby wyeliminować konieczność sporządzania listy osób zamieszkujących wylosowane gospodarstwa domowe. Komplikacje, na jakie narażone jest prowadzenie takiego wewnątrzzespołowego spisu jednostek rozpatrzone zostaną nieco później, teraz można wskazać, iż do najbardziej znanych – choć nie jedynych – procedur wewnątrzzespołowej selekcji zaliczyć można (poza wspomniana już siatką Kisha) nieprobabilistyczne metody, takie jak: (1) proce-

---

fikację poprawności losowania respondentów zgodnie z siatką Kisha. Wnioski z tych analiz są niezwykle interesujące, bowiem ukazują, iż wykorzystanie procedury Kisha w losowaniu gospodarstw domowych (w porównaniu z próbami imiennymi) powoduje zmniejszenie wskaźników realizowalności próby o niecałe 10 pp. Eksperyment przeprowadzony przez W. Smitha i in. (1995) ukazał również, że metoda Kisha została poprawnie przeprowadzona w 93% wybranych gospodarstw domowych, natomiast w 99% ankieterzy w sposób właściwy stosowali reguły inkluzji respondentów należących do wylosowanych gospodarstw domowych.

[97] Interesujące zestawienie kilkunastu procedur wewnątrzzespołowej selekcji jednostek odnaleźć można w artykule C. Gaziano (2005: 124–157) pt. *Comparative Analysis of Within-Household Respondent Selection Techniques*.

durę Troldahla-Cartera (1964), wraz z jej modyfikacją zaproponowaną przez Barbarę Bryant (1975), (2) Hagana-Collier (1983), czy też (3) Rizzo–Bricka-Parka (2004), a także procedury quasi-losowe, w tym (4) klasę metod doboru wykorzystujących daty urodzin domowników, polegających na wyborze takiej osoby, która (a) obchodziła urodziny jako ostatnia, (b) będzie następna obchodzić urodziny lub też, (c) której data urodzin przypada najbliżej dnia wizyty ankietera, niezależnie od tego, czy dzień urodzin już minął, czy też dopiero nastąpi.

Wydaje się zresztą, że właśnie klasa procedur wykorzystujących daty urodzin cieszy się szczególną popularnością wśród badaczy i wykorzystywana jest dość często w zastępstwie siatki Kisha. Wszystko opiera się jednak na założeniu, iż rozkład dat urodzin pozostaje losowy wśród domowników, dając tym samym losową próbę populacji. Chociaż wczesne studia nad tą procedurą (por. Salmon i in. 1983: 270–276; O'Rourke i in. 1983: 428–432) ukazały, że pozwala ona na uzyskanie próby reprezentatywnej, a także posiada wiele zalet w porównaniu z innymi – bardziej „inwazyjnymi" – procedurami wewnątrzzespołowej selekcji, jednak wyniki późniejszych analiz każą już powątpiewać w losowość rozkładu dat urodzin w obrębie gospodarstw domowych (por. Groves i in. 1988: 191–212). Do niezwykle pouczających wniosków w tym zakresie doprowadziły także dwie edycje badań przeprowadzonych pod kierunkiem P. Lavrakasa (por. Lavrakas i in. 1993, Lavrakas i in. 2000). Studia te ukazują, iż procedura daty urodzin ma tę przewagę nad metodą Kisha, że prowadzi do większych wskaźników kooperacji ankietera z respondentem oraz umożliwia obniżenie kosztów badań. Co ważne jednak, zidentyfikowano też cały szereg problemów wynikających z błędów w niewłaściwej implementacji procedury daty urodzin. W sumie okazało się bowiem, iż w około 20%–25% wszystkich przypadków proces wewnątrzzespołowej selekcji przeprowadzany był w sposób nieprawidłowy, co stanowiło znacznie większy odsetek uchybień niż w procedurze Kisha. Badacze ci zauważyli, że podstawowym błędem popełnianym w procesie selekcji było wskazywanie na siebie, tj. przez osobę, z którą nawiązano kontakt, jako na tę, która spełnia warunki selekcji, niezależnie od tego, czy akurat ta osoba obchodziła urodziny w terminie przypadającym najbliżej daty nawiązania kontaktu z danym gospodarstwem. Innymi słowy, zyski wynikające z większej kooperacji eliminowane były utratą poziomu reprezentatywności próby.

Przyglądając się ponownie dokumentacji metodologicznej projektu ESS, można zauważyć, że w badaniach tych wykorzystuje się dwie – najbardziej rozpowszechnione w sondażowej praktyce badawczej – metody losowania reprezentantów z wylosowanych do próby obiektów (mieszkań lub adresów/budynków): siatkę Kisha w celu doboru mieszkania w obrębie budynku,

Tabela III.2. Strategie doboru respondentów w badaniach ESS – próby adresowe *vs.* imienne

Nazwa kraju	Próba adresowa budynków mieszkalnych		Próba adresowa gospodarstw domowych		Próba imienna
	siatka Kisha & siatka Kisha	siatka Kisha & data urodzin	siatka Kisha	data urodzin	
Austria				ESS:1[i]–2–3–4	
Belgia					ESS:1–2–3–4–5
Bułgaria				ESS:3–4–5	
Chorwacja		ESS:4		ESS:5	
Cypr				ESS:3–4–5	
Czechy			ESS:1–2, 4–5		
Dania					ESS:1–2–3–4–5
Estonia				ESS:1	ESS:2–3–4–5
Finlandia					ESS:1–2–3–4–5
Francja				ESS:1–2–3–4–5	
Grecja			ESS:1–2,4–5		
Hiszpania		ESS:1			ESS:2–3–4–5
Holandia	ESS:5	ESS:1–2–3–4			
Irlandia				ESS:1–2–3–4–5	
Islandia					ESS:2
Izrael				ESS:1,4–5	
Litwa				ESS:2,4	
Luksemburg				ESS:1–2	
Łotwa				ESS:3–4	
Niemcy					ESS:1–2–3–4–5
Norwegia					ESS:1–2–3–4–5
Polska					ESS:1–2–3–4–5
Portugalia				ESS:1–2–3–4–5	
Rosja	ESS:3–4–5				
Rumunia				ESS:3–4	
Słowacja	ESS:4–5				ESS:2–3
Słowenia					ESS:1–2–3–4–5
Szwajcaria			ESS:1–2–3–4		ESS:5
Szwecja					ESS:1–2–3–4–5
Turcja		ESS:1–2,4			
Ukraina		ESS:2–3–4–5			
Węgry					ESS:1–2–3–4–5
Wielka Brytania	ESS:1–2–3–4–5				
Włochy			ESS:1		

Źródło: opracowanie własne na podstawie danych z repozytoriów ESS
[i] Oznaczenie edycji badań ESS

a przy wyborze konkretnego respondenta spośród członków gospodarstw – siatkę Kisha lub procedurę daty najbliższych urodzin[98].

Pamiętać należy również o tym, że chociaż tworzenie wykazu gospodarstw domowych w adresowych próbach budynków mieszkalnych (lub osób w adresowych próbach gospodarstw domowych) narażone jest na pewne ograniczenia wynikające bardziej z natury pomiaru, niż z właściwości rejestrów adresowych jako takich, to ułomności z tym związane mają jednak bezpośrednie przełożenie na jakość takich operatów i należy je uwzględnić na etapie wyboru konkretnego rejestru, z którego przeprowadzane będzie losowanie próby. Potwierdzają to rozważania R. Grovesa (1989: 108–115), który w trzecim rozdziale monografii *Survey Errors and Survey Costs*, w części poświęconej błędom pokrycia badanej populacji przez operaty gospodarstw domowych zamieszcza następujące stwierdzenie ukazujące, iż źródłem błędu niepełnego pokrycia jest nie tylko sam operat, ale również proces pomiaru:

> Tworzenie [zgodnej z rzeczywistością – P.J.] listy osób [zamieszkujących wylosowane gospodarstwo domowe – P.J.] nie zależy jedynie od obserwacji przeprowadzonej przez ankietera, ale wymaga zazwyczaj zadania pytań członkom gospodarstw domowych. Stopień pokrycia jednostek w obrębie gospodarstw domowych jest zatem zależny od zachowań ankieterskich, jak też od zachowań osób udzielających informacji […]. (Groves 1989: 109–110)

Komplikacje związane z poprawnym spisywaniem jednostek tworzących wylosowane zespoły stanowiły zresztą przedmiot wielu analiz metodologicznych poświęconych jakości operatów adresowych. Wystarczy przywołać ustalenia niektórych z takich studiów, by ukazać skalę oraz złożoność problemu. Dla przykładu, w niezwykle pouczającym opracowaniu *Who Lives Here? Survey Undercoverage and Household Roster Question* autorstwa Rogera Tourangeau i in. (1997: 1–18) zaprezentowane zostały wyniki eksperymentu poświęconego analizie tego, czy sposób sformułowania pytania o osoby zamieszkujące w wylosowanych gospodarstwach domowych wpływa na prawdziwość podawanych informacji. Analizy ukazała między innymi, że prośba o wskazanie inicjałów

---

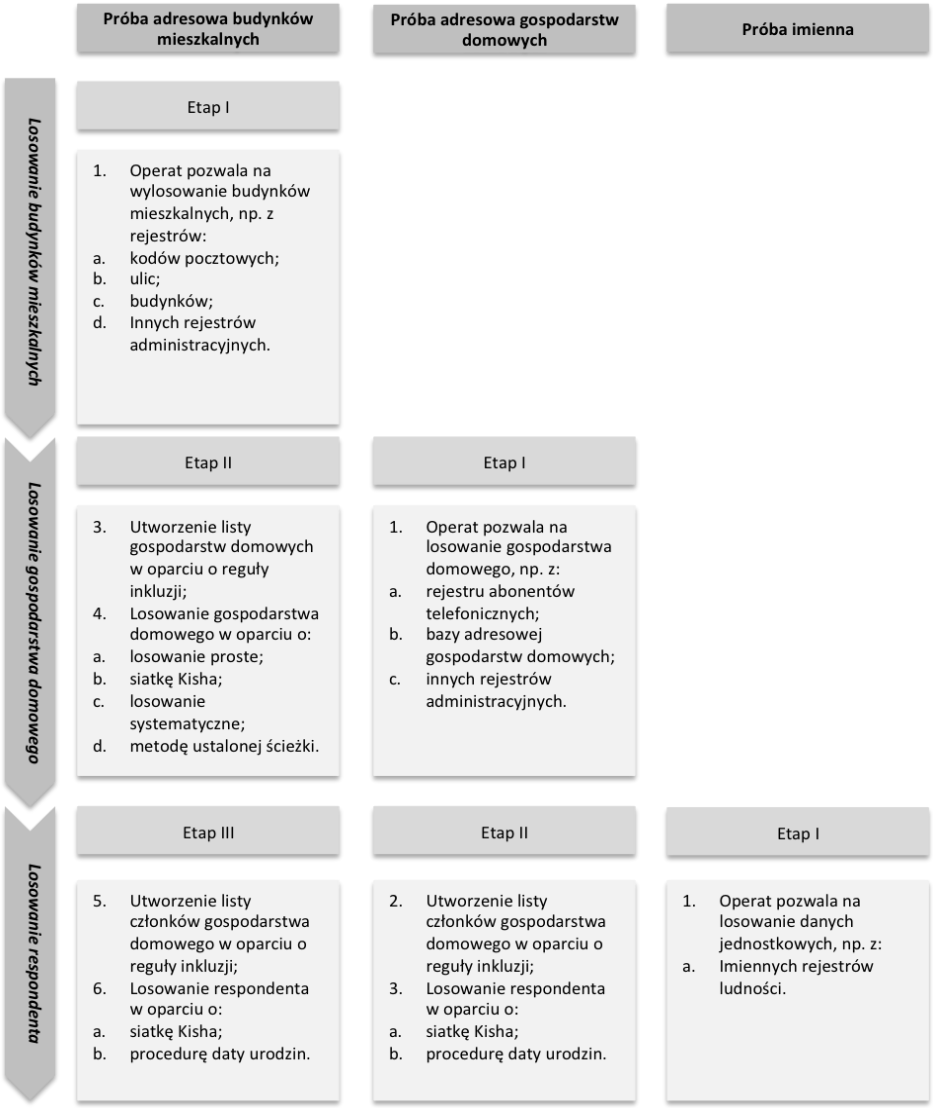
[98] W przywoływanym już siódmym rozdziale podręcznika *Fieldwork jest sztuką*, Z. Sawiński (2005) zamieszcza następujące stwierdzenie odnośnie procedury najbliższych urodzin: „Mimo swojej prostoty, metoda osoby obchodzącej urodziny jest rzadko stosowana przez instytuty badawcze. Główną przyczynę stanowią niekorzystne reakcje domowników na zadane przez ankietera pytanie o to, kto ostatnio obchodził urodziny. Pytania tego rodzaju nie sprzyjają tworzeniu klimatu, jaki ankieter stara się uzyskać w fazie aranżacji wywiadu" (Sawiński 2005: 92). Jakkolwiek niewątpliwie słuszna wydaje się uwaga o potencjalnie negatywnym wpływie „wypytywania" przez ankietera o daty urodzin domowników wylosowanych do próby gospodarstw, to jednak trudno zgodzić się z tym, że metoda ta jest rzadko wykorzystywana. Wystarczy przyjrzeć się danym z badań ESS (por. tab. III.2), by zauważyć, że procedura daty urodzin użyta była jednak w zdecydowanej większości państw, w których losowano adresowe próby osób.

członków gospodarstw domowych umożliwia „uchwycenie" większej liczby osób, podczas gdy żądanie podania dokładnych imion i nazwisk prowadzi do znacznego błędu pokrycia. W artykule opublikowanym dwa lata później w czasopiśmie *Public Opinion Quarterly* Elizabeth Martin (1999: 220–236) kontynuuje studia nad tymi zagadnieniami, skupiając uwagę na komplikacjach związanych z poprawnym klasyfikowaniem osób jako mieszkańców wylosowanych gospodarstw domowych[99]. Studia prowadzone w USA jeszcze w latach 70. XX wieku ukazały także, iż problem niepełnego wewnątrzzespołowego pokrycia[100] dotyczy zdecydowanie bardziej mężczyzn (por. Valentine i in. 1971: 40), a także osób mniej zamożnych oraz gorzej wykształconych (por. Korns 1977: 60–69). Ponadto, jak ukazuje R. Groves (1989: 110), powołując się na opracowanie autorstwa Camilli Brooks oraz Barbary Bailar (1978), w operatach adresowych podstawowym źródłem błędu niepełnego pokrycia populacji nie jest wcale pominięcie gospodarstw domowych, lecz jednostek tworzących owe gospodarstwa. Nie jest to żadnym zaskoczeniem, wszak: „[…] w próbie adresowej […] mamy […] do czynienia z próbą osób, mimo że instytut badawczy przekazuje ankieterom jedynie adresy gospodarstw domowych" (Sawiński 2005: 93). Z przywołanych opracowań można zatem wysunąć niezwykle interesujące wnioski o charakterze fundamentalnym dla oceny przydatności operatów adresowych w doborze jednostek do sondażowych prób badawczych. Najważniejsze jest bowiem to, że stopień pokrycia populacji gospodarstw domowych operatami adresowymi nie ma większego przełożenia na pokrycie jednostek indywidualnych. Zatem, nawet jeśli operat adresowy obejmie – przynajmniej teore-

---

[99] C. Gaziano (2008: 962), kontynuując tę problematykę, wskazuje na kilka powodów takich komplikacji, z czego dwa wydają się najważniejsze. Po pierwsze, sytuacja taka wynika z rozbieżności w operacyjnej definicji gospodarstwa domowego przyjętej przez badacza oraz społecznym rozumieniu tego pojęcia. Dodać do tego można również problemy w zrozumieniu – przez osobę udzielającą informacji o innych członkach gospodarstwa – reguł inkluzji, a więc tego, kogo należy uznać, a kogo nie, za osobę rzeczywiście zamieszkującą w danym gospodarstwie. Wiąże się z tym drugi problem, na który zwraca uwagę Gaziano (2008) a mianowicie, że osoby udzielające informacji o członkach gospodarstwa domowego mają tendencję do pomijania tych osób, które choć zamieszkują w wybranych gospodarstwach przez większą część roku, to jednak w pewnych okresach, na przykład tych obejmujących terenową fazę badań, bywają nieobecne.

[100] W metodologii badań sondażowych spotkać można specjalne określenie na ten specyficzny rodzaj uchybień w pokryciu populacji. Nosi on nazwę błędu niepełnego wewnątrzzespołowego pokrycia badanej populacji (oryg. *within–unit coverage error*). Błąd ten wynika oczywiście z tego, że respondenci wybrani dzięki zastosowaniu którejś z technik pozwalającej na wewnątrzzespołową selekcję jednostek, różnią się w sposób znaczący od tych, którzy zostali pominięci, a teoretycznie (przy poprawnym zastosowaniu techniki doboru) powinni zostać wybrani. Omawiana kategoria błędu wpływać będzie zarówno na losowy, jak i systematyczny komponent całkowitego błędu pomiaru badań sondażowych. Z tym pierwszym przypadkiem będzie się miało do czynienia wtedy, gdy procedura wykorzystana do wewnątrzzespołowej selekcji zróżnicuje szanse doboru jednostek, z drugim natomiast, gdy pewne specyficzne kategorie osób zostaną pominięte w procesie doboru (por. Zinniel 2008: 962–963).



Ryc. III.6. Typologia prób badawczych (adresowych oraz imiennych) z uwagi na rodzaje operatów

Źródło: opracowanie własne

tycznie – całą populację gospodarstw domowych, to i tak stopień pokrycia populacji docelowej (to znaczy jednostek tworzących ową populację) będzie niepełny, co ma swoje podstawowe źródło w błędach pomiarowych.

Można także zauważyć, że choć w sytuacji losowania gospodarstw domowych oraz budynków mieszkalnych ma się rzeczywiście w obu przypadkach do

czynienia z tzw. adresowymi próbami osób (por. Sawiński 2005: 88–93), to jednak w rzeczywistości występują tu dwa różne typy prób badawczych. Po pierwsze, w losowaniu budynków mieszkalnych pojawia się podwójne „uzespołowienie" jednostek indywidualnych (najpierw gospodarstw domowych w obrębie budynków, a następnie osób w obrębie gospodarstw domowych). Po drugie zaś, wylosowanie adresu jakiegoś budynku oznacza konieczność przeprowadzenia spisu jednostek mieszkalnych, wylosowania konkretnego mieszkania, sporządzenia wykazu osób w ramach gospodarstw domowych oraz wylosowania respondenta. Realizacja adresowej próby budynków mieszkalnych jest więc o wiele bardziej złożona niż realizacja adresowej próby gospodarstw domowych.

Zresztą statystyczne konsekwencje losowania adresów budynków mieszkalnych okazują się również dużo poważniejsze od konsekwencji związanych z losowaniem gospodarstw domowych. Wynika to z tego, że losowanie wielostopniowe oznacza prawie zawsze konieczność wyboru jednego przedstawiciela całego zespołu[101], a zatem, w konsekwencji, szanse selekcji w obrębie zespołu pozostają zależne od liczebności tworzących je jednostek. Ponieważ skutki nierównych prawdopodobieństw doboru ogranicza się poprzez ważenie danych, to losowanie adresów budynków mieszkalnych oznacza konieczność zrealizowania próby o większej liczebności, bowiem zrekompensowane muszą być zarazem nierówne szanse selekcji gospodarstw domowych w obrębie budynków, jak też wyrównane szanse selekcji osób w obrębie gospodarstw domowych. Problemy wynikające ze zróżnicowania szans selekcji jednostek populacji będą przedmiotem rozważań w kolejnym rozdziale tej pracy.

Powracając do analizy dokumentacji metodologicznej projektu ESS, można zauważyć, że bazą źródłową dla repozytoriów gospodarstw domowych lub budynków mieszkalnych były dane o dość zróżnicowanym charakterze. Dla przykładu, w większości krajów biorących udział w pierwszej oraz drugiej odsłonie ESS-u źródłem informacji o gospodarstwach domowych były (a) bazy abonentów telefonii stacjonarnej (Austria, Izrael, Szwajcaria), w innych natomiast (b) repozytoria klientów usług komunalnych (Czechy ESS1) lub też (c) urzędowe spisy gospodarstw domowych (na przykład w Hiszpanii adresowy rejestr gospodarstw domowych, natomiast w Luksemburgu – baza gospodarstw zamieszkiwanych przez osoby ubezpieczone w ramach systemu opieki społecznej). Z kolei źródłem danych wykorzystanych do konstruowania opera-

[101] Dla przykładu, w badaniach ESS prowadzonych w Izraelu losowaniu (zresztą zgodnie z zasadą prostej próby losowej) podlegają jednostki mieszkalne. Jeżeli wylosowany obiekt zawiera więcej niż jedno gospodarstwo domowe, to każde z nich wchodzi do próby badawczej z równym prawdopodobieństwem (por. *ESS1 Sampling report 2002*: 21). Znacznie częstszą praktyką jest jednak przeprowadzenie dodatkowego losowania w obrębie zespołu jednostek.

tów adresowych budynków mieszkalnych były (a) dane kodów pocztowych (Wielka Brytania, Holandia) oraz (b) inne rejestry administracji publicznej (na przykład urzędowe wykazy budynków mieszkalnych, wykorzystane w badaniach ESS2 w Czechach). Z informacji metodologicznych ESS można też odczytać, że wskaźniki pokrycia populacji takimi operatami okazywały się dość znaczne, osiągając wartość ponad 95 pp.[102]. Pamiętać należy jednak, iż w operatach adresowych nie chodzi praktycznie w ogóle o stopień pokrycia takim operatem gospodarstw domowych czy też budynków mieszkalnych, ale o to, do jakiego odsetka jednostek indywidualnych operat taki pozwala dotrzeć. Chociaż w badaniach ESS informacje o pokryciu badanych populacji rejestrami adresowymi dostępne były jedynie w odniesieniu do kilku krajów (dla przykładu w Hiszpanii rejestr adresowy pozwalał na dotarcie do 97,7% populacji kraju, natomiast w Wielkiej Brytanii wykaz kodów pocztowych pokrywał 97% populacji docelowej mieszkańców), to jednak nawet z tych cząstkowych danych można odczytać, że choć wskaźniki pokrycia były znaczne, to jednak już niższe niż w przypadku operatów imiennych.

Ostatnią z wyróżnionych pod względem charakteru użytych operatów doboru prób badawczych grupą państw uczestniczących w projekcie ESS były te wszystkie kraje, w których niedostępne były zarówno operaty imienne, jak i adresowe. Warto odwołać się przy tej okazji do przywoływanego już wcześniej artykułu P. Lynna i in. (2007: 107–124), poświęconego problemom ekwiwalentności prób badawczych w ESS, w którym odnaleźć można spory fragment rozważań metodologicznych poświęcony procedurom losowania prób badawczych w sytuacji braku dostępu do jakichkolwiek całościowych rejestrów (indywidualnych lub zespołowych) jednostek wchodzących w skład populacji docelowych. Autorzy artykułu wychodzą przy tym od oczywistej konstatacji, że „losowanie próby jest najbardziej skomplikowane wtedy, gdy żaden operat nie jest dostępny" (Lynn i in. 2007: 110). Dalej odnaleźć można informację o tym,

---

[102] Jedynym odstępstwem od tej reguły są badania realizowane w Austrii. Zastosowany w tym kraju operat abonentów telefonii stacjonarnej obejmował około 90% wszystkich gospodarstw domowych. Wskaźnik pokrycia populacji docelowej uznano za niewystarczający. Badacze zastosowali jednak prostą procedurę ograniczającą błąd pokrycia. Otóż, ponieważ podstawową jednostką losowania były miejscowości, do których proporcjonalnie do liczebności populacji przyporządkowano wiązki gospodarstw domowych (12 jednostek), to w obrębie każdej takiej wiązki 6 obiektów losowano z operatu abonentów telefonii stacjonarnej, a następnie, wykorzystując metodę ustalonej ścieżki (*random route*), dobierano do tego 6 pozostałych gospodarstw domowych. Punktem startowym był adres gospodarstwa wylosowanego z operatu abonentów telefonicznych, z kolei interwał wynosił 5 (w ESS1) lub 10 adresów (w ESS2). Zastosowanie takiego schematu doboru dawało szanse wylosowania gospodarstw domowych, które nie miały dostępu do telefonii stacjonarnej. Innymi słowy, wśród adresów losowanych z operatu abonentów telefonicznych były jedynie te, które posiadały dostęp do telefonii stacjonarnej, natomiast wśród adresów dobieranych zgodnie z metodą ustalonej ścieżki pewną część stanowiły gospodarstwa stelefonizowane, pozostałą natomiast niestelefonizowane.

że w takich szczególnych przypadkach w badaniach ESS wykorzystano – doskonale znaną w światowej literaturze metodologicznej pod nazwą *area sampling* (por. Marker i in. 2009: 490, Hall 2008: 33–36, Kennel 2008: 31–32, Särndal i in. 1992) – procedurę losowania prób przestrzennych[103]. Typowym przykładem implementacji takiego schematu doboru próby są badania zrealizowane w Grecji, gdzie podstawową jednostką losowania (*primary sampling unit*) były statystyczne obwody spisowe. W obrębie dobranych w ten sposób jednostek terytorialnych ankieterzy tworzyli listy adresowe gospodarstw domowych i przekazywali je do ośrodka koordynującego badanie, w którym przeprowadzano systematyczne losowanie wymaganej liczby obiektów, te zaś zwracano ankieterom w celu terenowej realizacji próby badawczej. Warto zauważyć, że tworzenie spisu gospodarstw domowych oddzielone zostało od etapu doboru adresów do próby, co miało wyeliminować potencjalne niebezpieczeństwo pojawienia się wpływu i efektu ankieterskiego już w fazie losowania mieszkań (por. *ESS1 Sampling report 2002*: 18) [104]. Chociaż praktyka taka jest godna uwagi oraz niezwykle pożądana z uwagi na konieczność zapewnienia odpowiedniej jakości próby sondażowej, to jednak w ESS nie była stosowana powszechnie. Istotnie bowiem, we Francji, Portugalii oraz na Ukrainie, czyli we wszystkich pozostałych państwach, w których – w ESS1 oraz ESS2 – dobierano próby przestrzenne, losowanie gospodarstw domowych przeprowadzane było już przez ankieterów. Stosowano przy tym różne strategie takiego doboru. We Francji wybierano najpierw gminy, którym przyporządkowywano odpowiednią do wielkości populacji liczbę gospodarstw domowych, a następnie procedurą ustalonej ścieżki z adresami „startowymi" losowanymi z książki telefonicznej wybierano systematycznie po 5 gospodarstw domowych (por. *ESS1 Sampling report 2002*: 14). Podobną strategię losowania zastosowano w Portugalii, gdzie w pierwszym etapie dobierano 100 miejscowości, każdej z nich przypisywano – proporcjonalnie do liczebności populacji – odpowiednią liczbę gospodarstw domowych, których losowanie prowadzone było systematycznie zgodnie z pro-

---

[103] W dużym skrócie można powiedzieć, że dobór próby jest wówczas wielostopniowy, przy czym pierwszym etapem losowania jest wybór niewielkich (pod względem liczby ludności) jednostek terytorialnych o dobrze określonych granicach (na przykład gmin, miejscowości, obwodów spisowych lub też ulic). Dopiero w ich obrębie dokonuje się spisu wszystkich budynków mieszkalnych oraz gospodarstw domowych i z nich losuje do próby badawczej odpowiednią liczbę gospodarstw domowych. Alternatywnym podejściem jest metoda ustalonej ścieżki. Nie wymaga ona sporządzenia pełnego wykazu gospodarstw domowych, lecz jedynie doboru (z jakiegoś operatu) adresów startowych (por. Häder i in. 2003).

[104] P. Lynn i in. autorzy artykułu *Methods for Achieving Equivalence of Samples in Cross-National Surveys: The European Social Survey Experience* podkreślali, że „[panel ekspertów ds. losowania prób badawczych – P.J.] nalegał, aby etap losowania próby został oddzielony od spisu [gospodarstw domowych w obrębie wylosowanych obszarów – P.J.] oraz by był przeprowadzany przez personel ośrodka koordynującego badanie lub kierownika badania, czyli przez osoby, które nie dokonywały spisu [jednostek mieszkalnych – P.J.]" (Lynn i in. 2007: 110).

cedurą ustalonej ścieżki. Punktem startowym był adres wybrany z książkowego wykazu abonentów telefonii stacjonarnej lub też punkt wylosowany z mapy zawierającej adresowe wykazy budynków (por. *ESS1 Sampling report 2002*: 31). Jeszcze inny schemat losowania wykorzystano na Ukrainie, gdzie w pierwszej kolejności dobierano miejscowości, a następnie „dolosowano" do nich (proporcjonalnie do liczebności populacji) ulice. W kolejnym etapie ankieterzy zliczali wszystkie jednostki mieszkalne znajdujące się przy wylosowanej ulicy, przekazywali zdobyte informacje do zespołu koordynującego badania, który ustalał (odpowiednio do liczby wszystkich zliczonych mieszkań), wymaganą liczbę gospodarstw domowych. Te z kolei dobierane były przez ankieterów w sposób systematyczny zgodnie z metodą ustalonej ścieżki (por. *ESS2 Sampling report 2004*: 51–52).

Studia przypadków państw biorących udział w projekcie ESS potwierdzają zatem, iż w sytuacji niedostępności operatów imiennych lub adresowych wylosowanie reprezentatywnej próby okazuje się przedsięwzięciem niezwykle ekwilibrystycznym. W świetle problematyki poruszanej w tym rozdziale najważniejsze jest jednak to, że w procedurach losowania przestrzennego dosyć trudno jest ocenić jakość takich strategii doboru prób badawczych. Niezwykle wymowny jest tutaj fakt, że w żadnym z krajów wykorzystujących losowanie jednostek terytorialnych nie pojawia się w dokumentacji metodologicznej projektu ESS informacja, czy też przynajmniej wzmianka o tym, jaki był stopień pokrycia krajowych populacji schematem losowania prób przestrzennych. Co za tym idzie, trudno jest ustalić choćby przybliżoną wielkość błędów systematycznych, pojawiających się tu z całą pewnością na skutek braku dostępu do wiarygodnych operatów doboru prób badawczych.

ROZDZIAŁ IV

Losowanie jednostek – schematy doboru próby

Przedstawione w poprzednim rozdziale usystematyzowanie prób badawczych zostało oparte w znacznej mierze na właściwościach rejestrów jednostek populacji wykorzystywanych w procesie losowania próby do badań reprezentatywnych. Chociaż związane z tym działania wpływały już – w pewnym sensie – na sposób doboru prób badawczych, to jednak nie wynikały jeszcze – *sensu stricto* – z przyjęcia określonych schematów losowania respondentów. Można bowiem zauważyć, iż decyzja o wykorzystaniu jakiegoś rejestru motywowana była przede wszystkim koniecznością wyeliminowania ryzyka pojawienia się systematycznego błędu niepełnego lub nadmiarowego pokrycia i w zasadzie nie wynikała z niczego innego. Co więcej, większość z podejmowanych przez badaczy działań skoncentrowana była na utworzeniu lub też uzyskaniu dostępu do takich rejestrów jednostek populacji, które, jeśli nie bezpośrednio, to przynajmniej pośrednio dawały każdej (lub prawie każdej) jednostce szanse wylosowania do próby. Innymi słowy, dążyło się do wyeliminowania błędu systematycznego, nawet jeśli konsekwencją tego miałby być przyrost wariancji estymatorów oraz utrata precyzji prowadzonego pomiaru.

W schematach losowania sondażowych prób badawczych nie chodzi już o zabiegi podejmowane w celu poprawy jakości operatów ani też o konsekwencje wykorzystania pewnych rejestrów, ale o określenie, w jaki sposób jednostki populacji miałyby być z takich operatów dobierane do próby. Ujmując to inaczej, problemem nie jest już to, z jakiego wykazu jednostek należałoby skorzystać, ale to, w jaki sposób z wybranego już operatu losowanie powinno przebiegać. Punktem odniesienia pozostaje – rzecz jasna – schemat losowania próby w sposób prosty, wraz ze swoimi charakterystykami definicyjnymi: (a) losowaniem indywidualnym (czyli doborem jednostek, a nie ich wiązek lub zespołów), (b) losowaniem jednostopniowym (czyli doborem przeprowadzanym wprost z rejestru populacyjnego, bez konieczności pośredniej selekcji jednostek), (c) losowaniem nieograniczonym (czyli doborem elementów z całej populacji,

a nie oddzielnie z jej poszczególnych warstw), oraz (d) losowaniem jednostek populacji z jednakowymi prawdopodobieństwami selekcji (por. Lissowski i in. 2008: 513).

Zauważyć można zatem, że wyprowadzone w poprzednim rozdziale rozróżnienie na próby imienne oraz adresowe dotyczyło takich działań badawczych, w wyniku których uchylane było kryterium jednostopniowości losowania. Rzecz jasna, operaty imienne pozwalały na spełnienie tego wymogu, natomiast adresowe wykazy gospodarstw domowych lub też budynków mieszkalnych wiązały się już z koniecznością losowania jednostek w sposób złożony – poprzez wstępny dobór punktu adresowego grupującego jednostki (losowanie zespołowe nieograniczone na pierwszym etapie), a następnie poprzez wtórne losowanie konkretnej jednostki spośród wszystkich osób zamieszkujących w dobranych gospodarstwach domowych (ograniczone losowanie indywidualne na drugim etapie)[105]. Innymi słowy, decyzja o zastosowaniu próby imiennej lub adresowej uwarunkowana była możliwością wykorzystania operatów zawierających indywidualne dane o jednostkach populacji.

Przyjęcie pewnego określonego schematu doboru próby wiąże się natomiast z uchyleniem któregoś z trzech pozostałych warunków nakładanych na próbę prostą, tj. losowania indywidualnego, doboru nieograniczonego lub losowania z zachowaniem równości szans selekcji jednostek z populacji do próby badawczej. Chociaż w praktyce nie sposób określić, wymienić oraz nazwać wszystkich schematów, według których losowanie mogłoby przebiegać, to jednak daje się wyodrębnić główne klasy takich procedur doboru. Pierwszy z owych schematów typologicznych różni się od losowania prostego tym, że jednostki populacji mają zróżnicowane szanse doboru do próby; mowa wówczas o schemacie losowania z nierównymi prawdopodobieństwami selekcji (por. Lissowski i in. 2008: 528). Drugi schemat opiera się na idei losowania zespołów, czyli grup jednostek, a następnie na przeprowadzeniu pomiaru wśród wszystkich osób tworzących owe zespoły (lub też na pomiarze wiązki, tj. części osób z zespołu); w takich przypadkach mówi się o doborze zespołowym[106] (por. Lissowski i in. 2008: 538). Z kolei trzeci schemat doboru próby różni się od losowania prostego tym, że elementy populacji dobierane są rozłącznie z poszczególnych jej części, nie zaś z całej populacji. Te części nazywa się warstwami, a schemat doboru próby – losowaniem warstwowym lub stratyfikacyjnym (por. Lissowski i in. 2008: 532).

---

[105] Może ono zresztą przebiegać na dwa sposoby. Po pierwsze, w każdym wylosowanym gospodarstwie dobierać można jedną osobę do wywiadu. Po drugie, można sporządzić listę osób ze wszystkich gospodarstw domowych i dopiero wówczas wylosować wymaganą liczbę respondentów. Konsekwencje wynikające z zastosowania obu tych procedur przeanalizowane będą w części rozdziału poświęconej zróżnicowanym prawdopodobieństwom doboru jednostek z populacji do próby.

[106] Różni się to tym od omówionego w poprzednim rozdziale losowania dwustopniowego, że badaniu podlega wiele osób tworzących zespół, a nie jedna wylosowana z każdego zespołu.

Istotę działań podejmowanych w ramach doboru sondażowych prób badawczych w sposób niezwykle trafny wyraził Robert Groves, który w monografii *Survey Errors and Survey Costs* zwraca uwagę na to, iż:

> chociaż precyzja pomiaru [tj. błąd wynikający z próbkowania – P.J.] częściowo zależy [od pozostającego poza wpływem badacza – P.J.] zróżnicowania populacji, to […] pozostaje [w znacznej mierze – P.J.] pod kontrolą osoby określającej schemat doboru próby. […] Działania, które mają na to największy wpływ, to: (1) stratyfikacja, czyli podział populacji na rozłączne warstwy jeszcze przed fazą losowania jednostek, (2) [zróżnicowanie – P.J.] szans selekcji poszczególnych elementów populacji do próby, (3) uzespołowienie, czyli dobór wiązek respondentów zamiast losowania pojedynczych osób. Czwartym działaniem mającym na to wpływ jest ustalenie wielkości próby badawczej. (Groves 1989: 252)

W praktyce sondażowej ma się częściej do czynienia ze schematami doboru próby innymi niż losowanie proste głównie z uwagi na to, iż „umożliwiają [one – P.J.] wykorzystanie posiadanej wiedzy o populacji do uzyskania dokładniejszych i bardziej wiarygodnych wyników, bez zwiększania liczebności próby, oraz ułatwiają realizację badań" (Lissowski i in. 2008: 528). W rozdziale tym wykazane zostanie w sposób jednoznaczny, że zarówno precyzja estymacji, jak i optymalizacja działań podejmowanych w trakcie terenowej fazy badań powinny być głównymi kryteriami branymi pod uwagę przy wyborze określonych schematów doboru prób do badań sondażowych.

Warto jednak rozpocząć od kwestii ogólniejszej, a mianowicie od ustaleń terminologicznych pozwalających na identyfikację pewnych najbardziej charakterystycznych typów prób badawczych, zróżnicowanych pod względem schematu losowania. Zauważyć można, że poprzez skrzyżowanie kryterium (a) doboru z równymi lub zróżnicowanymi prawdopodobieństwami selekcji, (b) doboru jednostek lub ich zespołów, a także (c) losowania nieograniczonego z całej populacji lub doboru niezależnego prowadzonego z poszczególnych jej warstw, otrzymuje się osiem typów prób badawczych.

Dla każdego z wyróżnionych typów podane zostały zakresy wartości mierników przyrostu wariancji, odpowiednio dla efektu stratyfikacji ($DEFF_s$), uzespołowienia ($DEFF_c$) oraz zróżnicowania prawdopodobieństw doboru jednostek ($DEFF_p$). Z analiz zaprezentowanych w rozdziale II wiadomo już bowiem, że konsekwencją przyjęcia jakiegoś schematu doboru próby odmiennego od losowania prostego[107] (o takiej samej liczebności) jest przyrost (lub ograniczeniu

[107] Prosta próba losowa pozostaje przykładem nieograniczonego doboru indywidualnego przeprowadzanego z jednakowymi prawdopodobieństwami selekcji każdej jednostki. Ponieważ efektywność każdego schematu losowania ujmowana jest w sposób relatywny do doboru prostego, to wszystkie komponenty schematu losowania prostego pozostają równe jedności.

nie) wariancji estymatorów. Poszczególnym schematom losowania, jak też czynnikom warunkującym ich efektywność, poświęcone będą kolejne sekcje tego rozdziału. W ostatniej części rozdziału rozważania teoretyczne zobrazowane zostaną analizami empirycznymi.

Tabela IV.1. Typologia prób badawczych z uwagi na schemat doboru respondentów

	Losowanie indywidualne		**Losowanie zespołowe**	
	Losowanie nieograniczone	**Losowanie warstwowe**	**Losowanie nieograniczone**	**Losowanie warstwowe**
Jednakowe szanse selekcji	1) imienna prosta próba losowa $\mathrm{DEFF}_s = 1$ $\mathrm{DEFF}_p = 1$ $\mathrm{DEFF}_c = 1$	3) imienna próba warstwowa z lokalizacją proporcjonalną $\mathrm{DEFF}_s \leq 1$ $\mathrm{DEFF}_p = 1$ $\mathrm{DEFF}_c = 1$	5) jednostopniowa próba zespołowa (badanie ze wszystkimi jednostkami w obrębie zespołu) lub dobór prosty przeprowadzany ze zbioru wszystkich członków wylosowanych zespołów $\mathrm{DEFF}_s = 1$ $\mathrm{DEFF}_p = 1$ $\mathrm{DEFF}_c \neq 1$	7) wielostopniowa, wiązkowana, adresowa próba warstwowa zespołu jednostek z lokalizacją proporcjonalną – badanie ze wszystkimi jednostkami w obrębie zespołu lub dobór prosty przeprowadzany ze zbioru wszystkich członków wylosowanych zespołów $\mathrm{DEFF}_s \leq 1$ $\mathrm{DEFF}_p = 1$ $\mathrm{DEFF}_c \neq 1$
Zróżnicowane szanse selekcji	2) imienna próba losowa nierównymi szansami doboru jednostek $\mathrm{DEFF}_s = 1$ $\mathrm{DEFF}_p > 1$ $\mathrm{DEFF}_c = 1$	4) imienna próba warstwowa lokalizacją nieproporcjonalną $\mathrm{DEFF}_s \neq 1$ $\mathrm{DEFF}_p > 1$ $\mathrm{DEFF}_c = 1$	6) adresowa próba osób – losowanie dwustopniowe jednostki w obrębie zespołu $\mathrm{DEFF}_s = 1$ $\mathrm{DEFF}_p \geq 1$ $\mathrm{DEFF}_c = 1$	8) wielostopniowa, warstwowana oraz wiązkowana próba adresowa osób lub próba imienna $\mathrm{DEFF}_s \neq 1$ $\mathrm{DEFF}_p \geq 1$ $\mathrm{DEFF}_c \neq 1$

Źródło: opracowanie własne: (2), (5), (6), (7). Za Groves (1989: 253) podano: (1), (3), (4), (8)

IV.1. Schematy doboru prób badawczych – analizy teoretyczne

IV.1.1. Losowanie warstwowe/stratyfikacyjne

Jednym z najczęściej wykorzystywanych schematów doboru prób badawczych pozostaje rozwarstwienie/stratyfikacja, czyli taka procedura losowania, której idea polega na niezależnym doborze obiektów z rozłącznych warstw populacji docelowej[108] (por. Kalsbeek 2008: 849–850). Co prawda,

[108] G. Lissowski i in. (2008: 532), autorzy podręcznika *Podstawy statystyki dla socjologów*, dodają, że w losowaniu warstwowym doborowi podlegać muszą jednostki (tzn. nie ich zespoły), a także, że próbę powinno się losować w jednym etapie oraz z jednakowymi szansami wylosowa-

problematyka losowania warstwowego została szczegółowo opisana w wielu opracowaniach naukowych o znaczeniu fundamentalnym dla metodologii badań sondażowych – należy tu wspomnieć choćby prace Jerzego Neymana (1934: 558–625), Tore Daleniusa (1950: 203–213), T. Daleniusa i in. (1959: 88–101), L. Kisha (1965: 75–91), Williama G. Cochrana (1977: 89–146), czy też R. Grovesa (1989: 253–256) – jednakże również obecnie zagadnienia te pozostają przedmiotem rozważań wielu metodologów, zwłaszcza w kontekście pomiaru zmiennych o rozkładach skrajnie asymetrycznych (por. np. Baillargeon i in. 2011: 56–65, Kozak i in. 2006: 157–163, Gunning i in. 2004a: 159–166; Gunning i in. 2004b; Kozak 2004: 797–806; Lednicki i in. 2003: 287–305; Hedlin 2000: 15–28; Lavallée i in. 1998: 33–43). Mnogość literatury stawia autora tej pracy przed poważnym dylematem, a mianowicie: czy warto szczegółowo rozważać kwestie, które dla większości czytelników będą zupełnie oczywiste? Nie ma sensu drobiazgowe referowanie znanych przecież podstaw losowania warstwowego, warto jednak skoncentrować uwagę na tych wątkach literaturowych, które są kluczowe dla podejmowanych w tej pracy zagadnień. Będą to zatem rozważania dotyczące: (a) efektu losowania warstwowego (tzn. oddziaływania stratyfikacyjnego schematu doboru próby na losowy komponent całkowitego błędu pomiaru), (b) procedur umożliwiających poprawę efektywności losowania (to znaczy działań zmierzających zarówno do optymalnego określenia warstw, jak i liczności podprób dobieranych z tych warstw) oraz (3) innych ważnych kwestii związanych z poruszaną w pracy problematyką.

Z uwagi na konieczność wprowadzenia formalnych oznaczeń, jak i dla zachowania porządku w strukturze narracji pracy, należy rozpocząć jednak – mimo wszystko – od kwestii podstawowej, a mianowicie od wskazania tego, że stratyfikacja lub, inaczej mówiąc, rozwarstwienie populacji jest schematem doboru próby polegającym na jej podziale na rozłączne zbiory w ten sposób, że każdy element populacji przynależy tylko i wyłącznie do jednej warstwy[109] ($h = 1, 2, \dots, H$ o licznościach $N_1, N_2, \dots, N_H$), tym samym wszystkie warstwy obejmują wszystkie elementy badanej populacji ($N = N_1 + N_2 + \cdots + N_H$). Możliwość zastosowania schematu losowania warstwowego wymaga jednak wstępnej wiedzy o strukturze populacji (tj. o licznościach warstw i/lub ich wewnętrznym zróżnicowaniu), jak też dostępu do takich operatów doboru prób badawczych, które pozwalają na niezależne losowanie jednostek z każdej war-

---

nia elementów w poszczególnych warstwach Taki rodzaj losowania nazywa się czasami prostą próbą warstwową (por. Barnett 1982: 119). W wielu projektach badawczych losowanie warstwowe obejmuje jednak również takie schematy doboru respondentów, które są wielostopniowe lub zespołowe (por. np.: *ESS1 Sampling report 2002, ESS2 Sampling report* 2004).

[109] W ramach każdej takiej warstwy dobór próby prowadzony może być zgodnie z zasadami losowania prostego albo też w oparciu o inny schemat selekcji respondentów, w zależności od przyjętych przez badacza kryteriów merytorycznych lub względów praktycznych.

stwy. Trudności z tym związane doskonale obrazują rozważania S. Dorofeeva oraz P. Granta, którzy w pierwszym rozdziale monografii *Statistics for Real-Life Sample Surveys. Non-Simple-Random Samples and Weighted Data*, w części dotyczącej procedur próbkowania warstwowego zamieszczają następującą konkluzję:

> W ściśle rozumianej stratyfikacji warstwy są predefiniowane, a próba losowana niezależnie: istnieje możliwość przydzielenia każdej osoby do odpowiedniej warstwy, jeszcze przed przeprowadzeniem wywiadu. Tym niemniej, choć w wielu przypadkach informacje o strukturze populacji są dostępne, [...] nie istnieją jednocześnie rejestry pozwalające [przyporządkować respondentów do warstw – P.J.]. (Dorofeev i in. 2006: 22)

Innymi słowy, jeśli wartości zmiennych warstwowych udaje się ustalić dopiero w oparciu o wyniki przeprowadzonego pomiaru, to w rzeczywistości nie ma się do czynienia z rozwarstwieniem populacji, lecz z rozwarstwieniem *post hoc* próby badawczej, czyli z poststratyfikacją, która w badaniach sondażowych nie jest schematem doboru próby, lecz ma określone zastosowanie w fazie obróbki danych wynikowych. W tym momencie ważne jest to, że liczebności dobieranych podprób badawczych $(n_1, n_2, \dots, n_H)$ mogą być proporcjonalne lub też nieproporcjonalne do liczności warstw w całej populacji[110]. W tym pierwszym przypadku będzie tak, iż dla każdej dowolnej *j*-tej warstwy zachodzić będzie równość $\frac{n_j}{n} = \frac{N_j}{N}$ (gdzie $n = n_1 + n_2 + \cdots + n_H$), co oznaczać będzie też, iż względne liczebności wszystkich podprób (to znaczy $f_j = \frac{n_j}{N_j}$) pozostaną równe wielkości $f = \frac{n}{N}$. Właściwości te wykorzystane zostaną nieco później, w ramach oceny efektywności schematów doboru prób z populacji rozwarstwionej.

Przyglądając się literaturze przedmiotu, odnaleźć można przynajmniej kilka argumentów przemawiających na rzecz stosowania schematów losowania prób z populacji rozwarstwionych. Część z nich ma charakter merytoryczny, inne

---

[110] Decyzję o lokalizacji proporcjonalnej lub nieproporcjonalnej podejmuje się, uwzględniając szereg kryteriów. Zostaną one omówione w dalszej części tej sekcji rozdziału. Pamiętać należy jednak o trzech niezwykle istotnych konsekwencjach wynikających z takich lokalizacji. Po pierwsze, proporcjonalne ulokowanie liczebności prób zachowuje równość szans selekcji jednostek w obrębie całej populacji, natomiast nieproporcjonalne wiąże się zawsze z koniecznością ważenia danych wyrównującego owe nierówne szanse losowania. Po drugie, proporcjonalne rozlokowanie jednostek populacji docelowej w obrębie poszczególnych warstw próby badawczej (o ile tylko takie wewnątrzwarstwowe losowanie prowadzone jest zgodnie z doborem prostym) umożliwi zawsze uzyskanie próby o efektywności losowania lepszej lub tak samo dobrej, jaką posiada efektywność doboru prostej próby losowej. Po trzecie, o większej lub mniejszej efektywności nieproporcjonalnego rozlokowania jednostek populacji w warstwach próby decyduje w znacznej mierze poziom wewnątrzwarstwowej oraz międzywarstwowej wariancji estymatorów parametrów populacyjnych (por. Groves, 1989: 256).

natomiast wynikają z kwestii organizacyjnych. Podstawowym motywem skłaniającym badaczy ku warstwowemu doborowi próby jest to, iż ów schemat pozwala zwiększyć poziom precyzji estymacji parametrów populacyjnych bez konieczności zwiększania liczebności próby. Jednakże zysk wynikający z losowania warstwowego nie jest wcale pewny. To, czy uda się wylosować próbę efektywniejszą od tej dobieranej zgodnie z zasadą losowania prostego, uwarunkowane jest wieloma czynnikami. Komplikacje wynikające z różnych przesłanek losowania warstwowego świetnie obrazują ustalenia Vica Barnetta (1982), zamieszczone w czwartym rozdziale podręcznika *Elementy teorii pobierania prób*[111]:

> […] rozwarstwienie populacji, przy zachowaniu odpowiednich warunków może zwiększyć efektywność estymacji, […], może być też pożądane ze względów organizacyjnych. […] [ale jednocześnie – P.J.] nie ma powodów, dla których rozwarstwienie wynikające z potrzeb organizacyjnych, miałoby koniecznie prowadzić do [poprawy efektywności losowania – P.J.]. (Barnett 1982: 115)

Z tego względu zasadne wydaje się postawienie kilku pytań, które pozwolą ocenić proces losowania warstwowego. Zacząć warto od spraw o charakterze ogólnym, przechodząc następnie do kwestii bardziej szczegółowych. W pierwszej kolejności należy zapytać: jakie warunki muszą być spełnione, aby efektywność losowania warstwowego była większa od efektywności doboru prostego o tej samej liczebności próby? Dalej: jakie strategie doboru próby należy zastosować, aby losowanie warstwowe miało największą efektywność? Ponadto: czy na efektywność losowania warstwowego wpływa sposób definiowania warstw, ich liczba oraz liczebność jednostek dobranych do próby badawczej w obrębie tych warstw populacji? Wreszcie: jakie względy praktyczne oraz organizacyjne przemawiają za losowaniem próby z populacji rozwarstwionych?

Odpowiedź na te pytania wymaga oczywiście wcześniejszego zdefiniowania efektywności doboru warstwowego. W drugim rozdziale monografii wskazane zostało już, że miernik efektu losowania z populacji rozwarstwionej (oznaczany jako $DEFF_s$) jest jednym z trzech komponentów miary całkowitego efektu doboru próby[112] ($DEFF_{TOTAL}$). Z przyjętej wówczas definicji wynikało (por. wzór II.6.), że miarę globalną, jak też poszczególne jej składniki, wyznacza się jako stosunek wariancji estymatorów, których wielkości dałoby się ustalić, przeprowadzając pomiar na wszystkich możliwych *n*-elementowych próbach losowanych zgodnie z ustalonym schematem doboru próby, oraz wariancji estyma-

---

[111] Przywoływane jest tutaj polskie tłumaczenie publikacji V. Barnetta *Elements of Sampling Theory*, wydanej w 1974 roku.

[112] Inne komponenty miary $DEFF_{TOTAL}$ to miernik przyrostu wariancji na skutek nierównych prawdopodobieństw selekcji ($DEFF_p$) oraz miara efektu losowania zespołowego ($DEFF_c$).

torów tych samych parametrów, których wartość można by ustalić, gdyby pomiar prowadzono na próbach prostych. Podobnie jak poprzednio, tak i tutaj, efektywność doboru prób badawczych zobrazowana zostanie na przykładzie dwóch klas estymatorów punktowych: wskaźników struktury oraz średnich arytmetycznych. Zacząć należy od przypomnienia tego, że dla *n*-elementowej prostej próby losowej dobieranej z populacji o skończonej liczbie *N* jednostek, wariancję estymatorów parametrów wskaźników struktury $\hat{p}$ oraz estymatorów parametrów średnich arytmetycznych $\bar{X}$ określa się jako:

$$\mathrm{Var}_{SRS}(\hat{p}) = (1-f)\frac{p(1-p)}{n-1}$$ (por. wzór II.8.)

oraz

$$\mathrm{Var}_{SRS}(\bar{X}) = \left(1-f\right)\frac{\sigma^2}{n}$$ (por. wzór II.9.),

gdzie: $f = \frac{n}{N}$ jest frakcją próby badawczej, p – jest wartością parametru wskaźnika struktury w całej populacji, natomiast σ^2 jest wielkością wariancji populacyjnej.

W próbach losowanych zgodnie ze schematem warstwowym wariancję estymatorów definiuje się w sposób bardzo podobny. Uwzględnić należy jednak nie tylko poziom wewnątrzwarstwowego zróżnicowania jednostek, ale też proporcje poszczególnych warstw w całej populacji. Jeżeli zatem z każdej *j*-tej warstwy (o liczbie elementów równych $N_1, N_2, \ldots, N_H$) losuje się w sposób niezależny próby badawcze o liczebnościach $n_1, n_2, \ldots, n_H$, to wariancję estymatorów wskaźników struktury:

(IV.1.) $$\hat{p}_s = \sum_{j=1}^{H} h_j\,\hat{p}_j,$$

gdzie $h_j = \frac{N_j}{N}$ jest frakcją *j*-tej warstwy w całej populacji, natomiast $\hat{p}_j$ jest estymatorem parametru wskaźnika struktury p_j w obrębie *j*-tej warstwy, można zapisać w postaci wzoru:

(IV.2.) $$\mathrm{Var}(\hat{p}_s) = \sum_{j=1}^{H} h_j^2\left(1-f_j\right)\frac{p_j(1-p_j)}{n_j-1}$$ (por. Groves i in. 2004: 114),

natomiast wariancję estymatorów średnich arytmetycznych:

(IV.3.) $$\bar{X}_s = \sum_{j=1}^{H} h_j\,\bar{X}_j,$$

gdzie $\bar{X}$ jest oszacowaniem parametru średniej arytmetycznej w *j*-tej warstwie, daje się zapisać w postaci równania:

(IV.4.) $$\mathrm{Var}(\bar{X}_s) = \sum_{j=1}^{H} h_j^2\left(1-f_j\right)\frac{\sigma_j^2}{n_j}$$ (por. Groves i in. 2004: 112),

w którym σ_j^2 oznacza wewnątrzwarstwową wariancję w *j*-tej warstwie.

Ponieważ określone zostały już wszystkie wielkości potrzebne do zdefiniowania miernika efektu losowania warstwowego, można podać, iż dla parametrów wskaźników struktury przyjmuje on postać wyrażenia:

$$\text{DEFF}_s = \frac{\sum_{j=1}^{H} h_j^2 (1-f_j) \frac{p_j(1-p_j)}{n_j - 1}}{(1-f)\frac{p(1-p)}{n-1}}, \tag{IV.5.}$$

które – jeśli tylko uwzględnić trzeba efekt (DEFF_j) jakiegoś odbiegającego od doboru prostego schematu losowania próby w obrębie każdej *j*-tej warstwy – można doprecyzować jako:

$$\text{DEFF}_s = \frac{\sum_{j=1}^{H} h_j^2 (1-f_j) \frac{p_j(1-p_j)}{n_j - 1} \text{DEFF}_j}{(1-f)\frac{p(1-p)}{n-1}}. \tag{IV.5'.}$$

Podobnie też zapisać można miernik efektu doboru warstwowego dla parametru średniej arytmetycznej:

$$\text{DEFF}_s = \frac{\sum_{j=1}^{H} h_j^2 (1-f_j) \frac{\sigma_j^2}{\text{n}_\text{j}}}{(1-f)\frac{\sigma^2}{n}} \tag{IV.6.}$$

lub dokładniej:

$$\text{DEFF}_s = \frac{\sum_{j=1}^{H} h_j^2 (1-f_j) \frac{\sigma_j^2}{n_j} \text{DEFF}_j}{(1-f)\frac{\sigma^2}{n}}. \tag{IV.6'.}$$

Po spełnieniu pewnych założeń wyrażenia (IV.5.), (IV.5'.), (IV.6.) oraz (IV.6'.) udaje się zapisać w nieco prostszej postaci. Po pierwsze, jeśli tylko liczebności poszczególnych warstw w populacji są dostatecznie duże, to dla dowolnej *j*-tej warstwy można przyjąć, że $\frac{N_j}{N} \sim \frac{N_j - 1}{N-1} \sim \frac{N_j}{N-1}$, a także, iż $f \sim 0$ oraz $f_j \sim 0$. Podobnie zresztą, jeżeli liczebności prób są wystarczająco duże, to $h_j^{'} = \frac{n_j}{n} \sim \frac{n_j - 1}{n-1}$. Dla estymatora parametru wskaźnika struktury miernik DEFF_s można wówczas zapisać w uproszczonej postaci jako:

$$\text{DEFF}_s \approx \sum_{j=1}^{H} \frac{h_j^2}{h_j'} \frac{p_j(1-p_j)}{p(1-p)} \text{DEFF}_j, \tag{IV.5''.}$$

co jest równoważne formule podanej przez S. Dorofeeva i in. (2006: 94), natomiast dla estymatora średniej arytmetycznej – w postaci przybliżenia:

$$\text{DEFF}_s \approx \sum_{j=1}^{H} \frac{h_j^2}{h_j'} \frac{\sigma_j^2}{\sigma^2} \text{DEFF}_j. \tag{IV.6''.}$$

Ważne jest to, że w wyrażeniach (IV.5'') oraz (IV.6'') znak przybliżenia można zastąpić równością, jeśli tylko jednostki w próbie rozlokowane są proporcjonalnie do liczebności poszczególnych warstw w całej populacji.

Obie formuły, (IV.5'') oraz (IV.6''), wykorzystane zostaną teraz do określenia czynników warunkujących efektywność losowania warstwowego. Można przy tym rozpatrzyć różnicę wariancji estymatorów otrzymanych w losowaniu prostym oraz stratyfikacyjnym[113], tak jak czynią to R. Groves (1989: 255) oraz V. Barnett (1982: 125), lub też rozważyć iloraz obu wariancji. Chociaż nie ma znaczenia, którą metodę się wybierze, to jednak warto przeanalizować stosunek obu wariancji, co wydaje się bardziej naturalne w kontekście działań zmierzających do ukazania skali przyrostu owej wariancji, pozostaje też zbieżne ze sposobem definiowania miary $DEFF_s$. Bez utraty ogólności dla formułowanych wniosków analizy ograniczone będą do przypadku estymacji parametru średniej populacyjnej. Założone będzie też, iż w ramach każdej warstwy losowanie elementów przeprowadzane jest w sposób odpowiadający doborowi prostemu, czyli że dla dowolnej *j*-tej warstwy populacji docelowej $DEFF_j = 1$. Choć w praktyce badawczej nie zawsze tak jest[114], to dla zobrazowania właściwości doboru warstwowego przyjęcie takiego założenia będzie bardziej właściwe. Wykorzystana będzie również niezwykle istotna charakterystyka wariancji, która pozwala wyrazić wielkość σ^2 poprzez sumę zróżnicowania wewnątrzwarstwowego oraz międzywarstwowego, co w sposób formalny wyraża równanie (por. Barnett 1982: 118):

$$\sigma^2 = \frac{1}{N-1}\sum_{j=1}^{H}\left(N_j - 1\right)\sigma_j^2 + \frac{1}{N-1}\sum_{j=1}^{H} N_j(\mu_j - \mu)^2, \qquad \text{(IV.7.)}$$

gdzie μ jest średnią populacyjną, natomiast dla każdej *j*-tej subpopulacji μ_j jest średnią warstwową. Ponieważ założono już wcześniej, że dla dowolnej *j*-tej warstwy $\frac{N_j}{N} \sim \frac{N_j - 1}{N-1} \sim \frac{N_j}{N-1}$, to formułę (IV.7.) można wyrazić prościej jako:

$$\sigma^2 \approx \sum_{j=1}^{H} h_j \sigma_j^2 + \sum_{j=1}^{H} h_j(\mu_j - \mu)^2. \qquad \text{(IV.7'.)}$$

Analizując efektywność losowania warstwowego warto w pierwszej kolejności rozważyć przypadek lokalizacji proporcjonalnej, takiej, w której zachodzi rów-

---

[113] Dodatnia wartość różnicy świadczy o większej efektywności losowania warstwowego, ujemna zaś o efektywności mniejszej.

[114] Dla przykładu w polskiej części badań ESS stosuje się losowanie warstwowe w subpopulacjach wyróżnionych z uwagi na typ i wielkość miejscowości. We wszystkich miastach o wielkości powyżej 50 tys. mieszkańców przeprowadza się losowanie proste, natomiast w warstwie wiejskiej oraz miast do 49,9 tys. mieszkańców wykorzystuje się losowanie zespołowe wymagające uwzględnia efektu wiązkowania respondentów (por. Sawińska i in. 2009: 12). Zagadnienia te omówione będą dokładniej w dalszej części rozdziału.

ność $h_j = h_j^{'}$. Po przekształceniach (IV.6".) oraz (IV.7'.) otrzymuje się wówczas, że:

(IV.8.) $$\mathrm{DEFF}_s = \frac{\sum_{j=1}^{H} h_j \sigma_j^2}{\sum_{j=1}^{H} h_j \sigma_j^2 + \sum_{j=1}^{H} h_j (\mu_j - \mu)^2}.$$

Ponieważ wartość licznika oraz mianownika wyrażenia (IV.8.) jest zawsze liczbą dodatnią a mianownik jest większy od licznika o wielkość równą $\sum_{j=1}^{H} h_j (\mu_j - \mu)^2$, to w lokalizacji proporcjonalnej $\mathrm{DEFF}_s \leq 1$. Ukazuje to niezwykle interesującą konsekwencję losowania warstwowego z lokalizacją proporcjonalną, bowiem ów schemat doboru próby posiada efektywność co najmniej tak dobrą jak losowanie proste z całej populacji (por. Dorofeev i in. 2006: 18). Im mniejsze będą zróżnicowania wyników w obrębie warstw, a większe pomiędzy nimi, tym efektywność schematu warstwowego będzie większa. Wyrażenie (IV.8.) będzie równe jedności (losowanie warstwowe proporcjonalne będzie miało efektywność losowania prostego), jeśli tylko komponent $\sum_{j=1}^{H} h_j (\mu_j - \mu)^2$ osiągnie wartość zero, czyli wówczas, gdy wszystkie średnie warstwowe będą sobie równe. W takiej sytuacji całość wariancji ulokowana będzie wewnątrz warstw, a nie między nimi. Doskonale ujął to V. Barnett, który w przywoływanym już podręczniku poświęconym doborowi sondażowych prób badawczych stwierdza, iż:

> charakterystyki populacji mogą być z większą precyzją estymowane na podstawie próby warstwowej, niż na podstawie prostej próby losowej, jeśli tylko średnie warstwowe się między sobą różnią, zaś wariancje wewnątrzwarstwowe są małe. Im większy jest właśnie efekt rozwarstwienia, tym większa jest efektywność odpowiednich estymatorów. […] Jeśli […] rozwarstwienia dokonuje się ze względu na wygodę zorganizowania badań, to wybór warstw jest ograniczony, a zatem wzrost efektywności nie jest pewny (chociaż z praktyki wynika, że wyznacza się takie podpodziały populacji na warstwy, które sprzyjają wzrostowi efektywności. (Barnett 1982: 159)

Problem polega właśnie na tym, że w praktyce badawczej ma się znacznie częściej do czynienia z nieproporcjonalnym rozlokowaniem liczebności jednostek próby w poszczególnych warstwach populacji. Chociaż ma to swoje merytoryczne uzasadnienie[115], to jednocześnie oznacza też, iż efektywność losowania warstwowego może być mniejsza od efektywności doboru prostego, tzn.

---

[115] Istnieje przynajmniej kilka powodów przemawiających za koniecznością nieproporcjonalnego przyporządkowania jednostek populacji do poszczególnych warstw w próbie (por. Dorofeev i in. 2006: 20). Przede wszystkim warto wskazać, iż nieproporcjonalne ulokowanie jednostek próby może wynikać ze względów organizacyjnych, np. z konieczności obniżenia kosztów badań, nierównych szans realizacji wywiadów w kolejnych warstwach populacji czy też niejednakowej jakości operatów wykorzystywanych do losowania prób badawczych z poszczególnych subpopulacji.

$\mathrm{DEFF}_s > 1$. Oczywiście nawet w doborze nieproporcjonalnym może być też tak, że $\mathrm{DEFF}_s \leq 1$, lecz sytuacja taka wystąpi wyłącznie przy spełnieniu pewnych warunków. Aby ukazać, od czego zależy kierunek owej nierówności, należy przyjrzeć się ponownie formule (IV.6''.), zapisując ją w nieco innej postaci:

(IV.6'''.) $$\mathrm{DEFF_S} \approx \frac{\sum_{j=1}^{H} \frac{h_j}{h_j} h_j \sigma_j^2}{\sigma^2}.$$

Zauważyć można teraz, że gdyby nie wyrażenie $w_j = \frac{h_j}{h_j}$ (które jest wagą stratyfikacyjną nadaną każdej jednostce wylosowanej z *j*-tej warstwy)[116], to formuła (IV.6'''.) byłaby równoważna (IV.8.). Ów iloraz jest kluczowy dla ustalenia czynników warunkujących efektywność losowania warstwowego z lokalizacją nieproporcjonalną. Wyrażenie to oznacza, że te warstwy, dla których liczebności obserwowane przewyższają oczekiwane, otrzymają wagę mniejszą od jedności, te zaś, w których liczebności obserwowane są mniejsze od oczekiwanych, uzyskają wagę większą od jeden. W konsekwencji, gdy w próbie badawczej znajduje się więcej – niż wynika to z proporcji populacji – elementów z warstw charakteryzujących się mniejszym zróżnicowaniem, a mniej elementów z tych warstw, które cechuje większa wewnątrzwarstwowa wariancja, to te pierwsze otrzymają wagę mniejszą, drugie zaś wagę większą od jedności. Co za tym idzie, licznik wyrażenia (IV.6'''.) będzie większy od mianownika, a zatem efektywność doboru warstwowego okaże się w takiej sytuacji mniejsza niż w przypadku doboru prostego. Zresztą w literaturze metodologicznej jest to zagadnienie doskonale rozpoznane. Dla przykładu Groves (1989) stwierdza, że:

> [w takim schemacie doboru próby – P.J.] bądź to większa, bądź też mniejsza wariancja estymatorów jest możliwa w porównaniu do losowania prostego. Kierunek […] zależy od relacji pomiędzy wewnątrzwarstwowym zróżnicowaniem wyników w określonej warstwie a frakcją losowania warstwy [proporcją elementów z warstwy w próbie – P.J.]. Jeżeli większa frakcja losowania została użyta w warstwach o większej wariancji, to zredukuje to wariancję [estymatorów – P.J.] względem lokalizacji proporcjonalnej. (por. Groves 1989: 256)

Jest to całkowicie zgodne z intuicjami. Wiadomo bowiem, że dla osiągnięcia określonej precyzji estymacji parametrów charakteryzujących się różnymi wariancjami większa liczebność próby badawczej potrzebna będzie tam, gdzie zmienność wyników w populacji jest większa. Odnosząc to do doboru warstwowego, można powiedzieć, że w losowaniu warstwowym nieproporcjonal-

---

[116] Tak ustalone wagi mają na celu przekształcenie szans selekcji wynikających z nieproporcjonalnego rozlokowania jednostek w próbie w odniesieniu do rzeczywistych prawdopodobieństw doboru wynikających z liczności warstw w populacji.

nym efektywniejsze jest zwiększenie próby w tych warstwach, w których poziom zróżnicowania wyników jest większy, nawet kosztem zmniejszenia jej tam, gdzie wariancja jest mniejsza.

Zresztą problematyka optymalnej lokalizacji próby w obrębie wybranych warstw populacji docelowej została doskonale opisana w literaturze światowej. Większość autorów odwołuje się przy tym do pewnego szczególnego typu lokalizacji nieproporcjonalnej, opracowanego przez polskiego statystyka Jerzego Neymana. Podstawy tej procedury opublikowane zostały w monografii *Zarys Teorji*[117] *i Praktyki Badania Struktury Ludności Metodą Reprezentacyjną* wydanej w 1933 roku nakładem Wydawnictwa Instytutu Problemów Społecznych z siedzibą w Warszawie, natomiast międzynarodowy rozgłos zyskała dzięki artykułowi zamieszczonemu w czasopiśmie „Journal of the Royal Statistical Society" (por. Neyman 1934: 558–625). Chociaż lokalizacja Neymana jest najbardziej znana, to pozostaje wyłącznie szczególnym przypadkiem jednego z wielu kryteriów optymalizacji losowania warstwowego. Charakterystykę głównych klas takich kryteriów odnaleźć można w podręczniku V. Barnetta (1982: 130–140). Kryteria te obejmują warunki: (1) minimalizacji wariancji estymatorów, przy ustalonym całkowitym koszcie badania oraz znanych kosztach losowania w warstwach, (2) minimalizacji kosztów, przy ustalonej maksymalnej wariancji estymatorów, a także (3) minimalizacji całkowitej wielkości próby badawczej, przy ustalonej maksymalnej wariancji estymatorów oraz określonych proporcjach warstw w całej próbie. Nie ma sensu charakteryzowanie rozwiązań optymalnych dla wszystkich tych kryteriów, warto jednak skupić się na rozwiązaniach pierwszego oraz trzeciego warunku optymalizacji. Ma to swoje uzasadnienie o tyle, że choć redukcja kosztów badania jest kryterium niezwykle ważnym, to jednak w świetle analiz ukierunkowanych na jakość badań sondażowych musi pozostać – siłą rzeczy – warunkiem drugorzędnym.

Warto zacząć od rozważenia kryterium trzeciego. Rozwiązanie optymalne z uwagi na ten warunek pozwala na ustalenie minimalnej wielkości próby badawczej potrzebnej do osiągnięcia określonej precyzji estymacji, przy ustalonych *a priori* liczebnościach próby w warstwach (tzn. określonych przez badacza $h_j^{'}$ dla każdej *j*-tej warstwy). W dalszej części wywodu znalezione będzie dla tego przypadku konkretne zastosowanie, teraz trzeba podać jedynie, że minimalną wielkość próby badawczej dla estymatorów wskaźników struktury można wyznaczyć z nierówności:

$$n \geq \frac{\sum_{j=1}^{H} \frac{h_j^2}{h_j^{'}} p_j(1-p_j)}{\frac{1}{\mathrm{Var}(\hat{p}_S)} + \frac{1}{N}\sum_{j=1}^{H} \frac{h_j^2}{h_j^{'}} p_j(1-p_j)}, \tag{IV.9.}$$

[117] Pisownia oryginalna.

natomiast dla estymatorów średniej arytmetycznej – za pomocą formuły:

(IV.10.) $$n \geq \frac{\sum_{j=1}^{H} \frac{h_j^2}{h_j} \sigma_j^2}{\frac{1}{\mathrm{Var}(\bar{X}_S)} + \frac{1}{N} \sum_{j=1}^{H} \frac{h_j^2}{h_j} \sigma_j^2}$$ (por. Barnett 1982: 138).

Z kolei pierwsze z podanych wcześniej kryteriów optymalnej lokalizacji jednostek w losowaniu warstwowym (odnoszące się do minimalizacji wariancji estymatorów) sprowadza się do wyznaczenia takich liczebności podprób badawczych $(n_1, n_2, \ldots, n_H)$, dla których $\mathrm{Var}(\hat{p}_s)$ oraz $\mathrm{Var}(\bar{X}_s)$ osiągać będą wartości minimalne przy określonej wielkości kosztów przeznaczonych na realizację projektu badawczego. Ponieważ strukturę takich kosztów można zapisać w postaci równania $C = c_0 + \sum_{j=1}^{H} c_j n_j$ (por. Groves 1989: 258), gdzie c_0 oznacza sumę kosztów organizacyjnych, natomiast – dla dowolnej *j*-tej warstwy – c_j jest kosztem selekcji, dotarcia oraz realizacji wywiadów z jednostkami wylosowanymi do próby, to dla estymatorów wskaźników struktury rozwiązaniem optymalnym jest formuła:

(IV.11.) $$n_j = (C - c_0) \frac{h_j \sqrt{\frac{p_j(1-p_j)}{c_j}}}{\sum_{j=1}^{H} h_j \sqrt{\frac{p_j(1-p_j)}{c_j}}}$$ (por. Barnett 1982: 153),

natomiast dla estymatorów średnich arytmetycznych – wyrażenie:

(IV.12.) $$n_j = (C - c_0) \frac{\frac{h_j \sigma_j}{c_j}}{\sum_{j=1}^{H} \frac{h_j \sigma_j}{c_j}}$$ ((por. Barnett 1982: 133).

Praktyczna implementacja owych procedur skutkować będzie tym, że w warstwach o dużej zmienności oraz niewielkich kosztach losowania, wielkość próby badawczej będzie proporcjonalnie dużo większa od wielkości próby dobieranej z innych warstw, zwłaszcza tych homogenicznych o znacznych kosztach realizacji wywiadów.

Jeżeli założy się teraz, iż koszty losowania są jednakowe w każdej warstwie, (tj. że $C = c_0 + c \cdot n$), to optymalizacje (IV.11.) oraz (IV.12.) sprowadzają się do wspomnianej wcześniej lokalizacji Neymana. Dla estymatorów parametrów wskaźników struktury można ją zapisać w postaci wyrażenia:

(IV.13.) $$n_j = \frac{h_j \sqrt{p_j(1-p_j)}}{\sum_{j=1}^{H} h_j \sqrt{p_j(1-p_j)}} \cdot n,$$

natomiast dla estymatorów parametrów średnich arytmetycznych jako:

(IV.14.) $$n_j = \frac{h_j \sigma_j}{\sum_{j=1}^{H} h_j \sigma_j} \cdot n.$$

Innymi słowy, równania (IV.13.) oraz (IV.14.) umożliwiają optymalne rozlokowanie jednostek przy ustalonej całkowitej wielkości próby. Jeżeli teraz rozpatrzy się pierwsze czynniki obu iloczynów (tzn. wyrażenia $\frac{h_j \sigma_j}{\sum_{j=1}^{H} h_j \sigma_j}$ oraz $\frac{h_j \sqrt{p_j(1-p_j)}}{\sum_{j=1}^{H} h_j \sqrt{p_j(1-p_j)}}$), nietrudno zauważyć, iż w lokalizacji „neymanowskiej" dużo większa liczba jednostek próby badawczej przydzielona zostanie do tych warstw, w których zróżnicowanie wartości zmiennych jest większe. Jeżeli zatem w danej warstwie występuje znaczne zróżnicowanie wartości badanej zmiennej, to w celu uzyskania określonej efektywności doboru próby konieczne byłoby wylosowanie z takiej warstwy większej liczby elementów, niż z warstwy o mniejszej zmienności (por. Groves i in. 2004: 117–118). Jest to zresztą zgodne z tym, o czym mówiono wcześniej – nieproporcjonalne rozlokowanie jednostek będzie tym bardziej efektywne, im więcej elementów próby przypadnie na warstwy bardziej zróżnicowane. Obserwację tę potwierdzają analizy V. Barnetta, który wskazuje iż „[…] zysk, jaki możemy osiągnąć, zastępując losowanie proporcjonalne losowaniem Neymana, zależy od zmienności wariancji warstwowych: im większe są te wariancje, tym większy zysk" (Barnett 1982: 139).

Z lokalizacją Neymana wiążą się jednak pewne istotne ograniczenia natury metodologicznej oraz praktycznej (por. Groves i in. 2004: 118). W pierwszej kolejności warto wskazać, iż optymalizacja neymanowska nie sprawdza się tak dobrze w odniesieniu do estymatorów parametrów wskaźników struktury jak w przypadku estymatorów średnich arytmetycznych, ponieważ daje zazwyczaj próby o znacznych liczebnościach. Po drugie, rozwiązanie optymalne względem jednej zmiennej nie musi być już optymalne względem innej, co wynika oczywiście z tego, iż efektywność rozwarstwienia odnosi się do konkretnej zmiennej, a nie wielu zmiennych jednocześnie. Po trzecie, w optymalizacji Neymana nie bierze się pod uwagę niedoskonałości operatów losowania, błędów wynikających z niedostępności wylosowanych osób, uchybień pomiarowych, czy też innych źródeł błędów, których oddziaływanie na całkowity błąd pomiaru może być uwarunkowane warstwą, w której prowadzony jest pomiar. Bardzo często bowiem decyzja o rozwarstwieniu populacji wynika z przyczyn technicznych, nie jest natomiast motywowana dążeniem do zminimalizowania wariancji estymatorów.

Typowym przykładem zastosowania strategii losowania warstwowego przeprowadzanego ze względów organizacyjnych są badania Europejskiego

Sondażu Społecznego. Przyglądając się polskiej części projektu ESS, można zauważyć, iż rozwarstwienie populacji dokonywane jest z uwagi na typ oraz wielkość miejscowości, z kolei liczności prób pozostają nieproporcjonalne do rzeczywistych wielkości warstw w populacji, czego głównym powodem są nierówne prawdopodobieństwa realizacji wywiadów w miejscowościach o różnym typie oraz wielkości[118] (por. Sawińska i in. 2009: 12–14). Pociąga to za sobą ciekawe konsekwencje metodologiczne, które jednak omówione zostaną w dalszej części rozdziału, poświęconej nierównym prawdopodobieństwom selekcji jednostek z populacji do próby badawczej. W tym momencie najistotniejsze jest to, że ponieważ w badaniach ESS realizowanych w Polsce liczba losowanych jednostek w warstwach populacji ustalana jest z góry, to optymalizacja Neymana nie znajduje zastosowania. Jeśli jednak proporcje warstw wyznaczone są *a priori*, to można dążyć do zminimalizowania całkowitej liczebności próby potrzebnej do osiągnięcia określonej precyzji estymacji. Służyć temu mogłyby omówione wcześniej nierówności (IV.9.) oraz (IV.10.).

Przyglądając się procedurom lokalizacji optymalnej, można też wskazać na problem ogólniejszy, a mianowicie na to, że jeśli badaczowi nie są znane wielkości wariancji warstwowych, lecz jedynie ich przybliżenia pochodzące z badań pilotażowych lub wcześniejszych badań sondażowych (co jest raczej normą niż wyjątkiem), to lokalizacja nieproporcjonalna (nawet ta zoptymalizowana) zawsze okazuje się ryzykowna. We wszystkich takich sytuacjach wielkości parametrów należałoby zastąpić ich estymatorami (por. Clark 2013: 6–23), które są, jak wiadomo, wyłącznie przybliżeniem prawdziwych wartości parametrów populacyjnych. Chociaż same w sobie podlegają one wszystkim typowym ograniczeniom wnioskowania indukcyjnego, to jednak „nawet oszacowania obarczone błędem mogą zwiększać efektywność [warstwowych schematów – P.J.] doboru prób" (Lissowski i in. 2008: 537).

Można oczywiście ograniczyć potencjalne niebezpieczeństwo wynikające z nieproporcjonalnego przyporządkowania jednostek próby do poszczególnych warstw populacji. Wymagałoby to jednak takiego podziału populacji, by wariancje wewnątrzwarstwowe pozostały jak najmniejsze, natomiast te międzywarstwowe – możliwie największe[119]. Z przynajmniej dwóch fundamentalnych powodów nie jest to jednak sprawa prosta. Po pierwsze, badacz posiada (wbrew pozorom) niezwykle ograniczoną możliwość manipulowania wyborem warstw w populacji i chociaż to, jaką liczbę warstw określi, nie odgrywa prak-

---

[118] W projekcie ESS prowadzonym w Polsce większe wskaźniki realizowalności próby odnotowuje się w badaniach prowadzonych na terenach wiejskich, mniejsze natomiast w dużych miastach. Jest to prawidłowość charakterystyczna dla badań sondażowych w ogóle (por. Grzeszkiewicz-Radulska 2009: 167–175; Goyder i in. 1992).

[119] Wewnątrz zdefiniowanych subpopulacji jednostki musiałyby charakteryzować się homogenicznością, natomiast warstwy (między sobą) – heterogenicznością.

tycznie żadnej roli (por. Cochran 1977: 132–134), to jednak sposób, w jaki warstwy są zdefiniowane, ma już znaczenie fundamentalne (por. Kozak i in. 2006: 159–160). Po drugie, jak już zostało wskazane, rozwiązanie optymalne dla jednej zmiennej nie musi być wcale optymalne dla innej (por. Groves 1989: 256). Co za tym idzie, zarówno definiowanie warstw, jak i rozlokowanie jednostek w kolejnych warstwach próby badawczej, nie są problemami, które by można w łatwy sposób rozwiązać.

IV.1.2. Losowanie zespołowe

Podział populacji na zespoły przypomina trochę operację jej rozwarstwienia, z tą oczywiście różnicą, że zespołów takich jest znacznie więcej i są one nieporównywalnie mniej liczne niż warstwy w populacji. Za tym pozornym podobieństwem kryją się jednak zasadnicze różnice o charakterze substantywnym. O ile bowiem losowanie warstwowe zakłada pomiar prowadzony na reprezentatywnym podzbiorze jednostek dobieranych w ramach każdej (bez wyjątku) warstwy populacji, o tyle losowanie zespołowe opiera się już na (możliwie) wyczerpującym badaniu całych zespołów, lecz nie wszystkich, ale tylko tych dobranych do próby. Różnice w obu schematach losowania niezwykle trafnie wyraził Barnett (1982) w piątym rozdziale opracowania *Elementy teorii pobierania prób*, stwierdzając:

> [w] doborze warstwowym losujemy z każdej warstwy. Tutaj natomiast, zamiast wybierać niektórych osobników z każdej warstwy, wybieramy niektóre warstwy [a dokładnie niektóre zespoły – P.J.] i chcemy możliwie w pełni [przebadać każdy wybrany zespół – P.J.]. (por. Barnett 1982: 165)

W taki sam sposób losowanie zespołowe definiują zresztą H.F. Weisberg (2005: 242) w monografii *The Total Survey Error Approach,* oraz P. Biemer i L. Lyberg (2003: 343) w dziewiątym rozdziale podręcznika *Introduction to Survey Quality*.

Jeśli zatem populacja składa się z M rozłącznych zespołów o liczebnościach równych N_1, N_2, …, N_M pokrywających badaną populację w stopniu zupełnym ($N = N_1 + N_2 + \ldots + N_M$), to schemat losowania zespołowego polega na doborze m spośród M zespołów, a także (a) na przeprowadzeniu pomiaru wszystkich jednostek wchodzących w skład tych zespołów – ma się wówczas do czynienia z losowaniem jednostopniowym całych zespołów (por. np. Lissowski i in. 2008: 538; Biemer i in. 2003: 346) lub też (b) na realizacji badań z jakąś reprezentacją jednostek tworzących owe zespoły – w takich sytuacjach mówi się o wiązkowanym losowaniu dwustopniowym (por. np. Lissowski i in. 2008: 542; Biemer i in. 2003: 346).

Niezwykle istotne jest to, że losowanie zespołowe (w przeciwieństwie do doboru próby z populacji rozwarstwionych) podejmowane jest prawie wyłącznie z uwagi na zoptymalizowanie procesu pomiaru, to znaczy z przyczyn organizacyjnych (por. Weisberg 2005: 242; Groves i in. 2004: 102; Groves 1989: 260)[120] lub też z powodu niskiej jakości operatów doboru próby zawierających dane jednostkowe (por. Weisberg 2005: 243) [121], nie jest natomiast motywowane dążeniem do uzyskania estymatorów bardziej efektywnych od tych z prób prostych (tak jak w proporcjonalnym doborze warstwowym). W konsekwencji te „inne metody [niż losowanie zespołowe – P.J.] mogą […] dawać bardziej efektywne estymatory, ale [jak się zazwyczaj okazuje – P.J.] przy wyższych kosztach i większym trudzie organizacyjnym" (Barnett 1982: 166). W takich przypadkach ma się zatem nadzieję – jak ukazuje V. Barnett w dalszej części swojego wywodu, dotyczącego motywów skłaniających ku wykorzystaniu schematów doboru zespołowego – że „straty wynikające ze zmniejszenia efektywności estymacji rekompensują mniejsze koszty i większa łatwość losowania" (Barnett 1982: 166).

Efektywność doboru zespołowego najłatwiej zobrazować, przyglądając się jednostopniowemu schematowi losowania zespołów o takiej samej liczbie elementów. Chociaż jest to przypadek szczególny, który rzadko kiedy występuje w praktyce, to jednak pozwala – mimo wszystko – na ukazanie pewnych uniwersalnych właściwości losowania zespołowego. Przyjęte zostanie zatem założenie, że populacja składa się z *M* zespołów, z których każdy ma taką samą liczbę *L* jednostek[122], a losowaniu – w sposób prosty zależny – podlega *m* zespołów oraz wszystkie jednostki wchodzące w ich skład. Ponieważ każdy zespół ma (przy takich założeniach) jednakową liczbę elementów, to nieobciążony estymator wskaźnika struktury[123]:

(IV.15.) $$\hat{p}_c = \frac{1}{m}\sum_{j=1}^{m}\hat{p}_j,$$

gdzie $\hat{p}_j$ oznacza wartość wskaźnika struktury w każdym z *m* zespołów (*de facto* wartość parametru zespołowego, tj. $\hat{p}_j = p_j$), będzie miał wariancję wyrażoną

---

[120] Np. w celu ograniczenia przestrzennego rozproszenia jednostek próby oraz redukcji kosztów dotarcia do respondentów.

[121] Jak wiadomo, adresowe próby gospodarstw domowych lub budynków mieszkalnych pozwalają na dotarcie do jednostek populacji poprzez wstępne wylosowanie ich zespołów.

[122] Co oznacza, że liczebność całej populacji wyrazić można wzorem $N = M \cdot L$.

[123] Bezsprzeczne wydaje się stwierdzenie, że dla innych wariantów losowania zespołowego estymatory określone wzorami (IV.15.) oraz (IV.17.) nie muszą być już statystycznie nieobciążone. Wystarczy przywołać sytuację doboru prostego zespołów o nierównych liczebnościach, czy też nieproporcjonalny do wielkości populacji dobór równolicznych wiązek respondentów z nierównolicznych zespołów, by ukazać oczywiste obciążenie estymatorów opisanych wzorami (IV.15.) i (IV.17.). Kwestie te podjęte zostaną w części poświęconej praktycznym komplikacjom wynikającym z konieczności oceny złożonych i wielostopniowych schematów doboru sondażowych prób badawczych.

wzorem:

(IV.16.) $\mathrm{Var}(\hat{p}_c) = (1-f)\frac{1}{m}\cdot\frac{\sum_{j=1}^{M}(p_j-p)^2}{M-1}$ (por. Barnett 1982: 173),

w której p_j oznacza wartość parametru wskaźnika struktury w obrębie każdego *j*-tego zespołu populacji, natomiast *p* jest wartością szacowanego parametru w całej populacji. Podobnie można podać, że nieobciążony estymator średniej arytmetycznej:

(IV.17.) $$\bar{X}_c = \frac{1}{m}\sum_{j=1}^{m}\bar{X}_j,$$

gdzie $\bar{X}_j$ oznacza oszacowanie (a w zasadzie – wartość parametru) średniej w każdym *j*-tym zespole, ma wariancję równą:

(IV.18.) $\mathrm{Var}(\bar{X}_c) = (1-f)\frac{1}{m}\cdot\frac{\sum_{j=1}^{M}(\mu_j-\mu)^2}{M-1}$ (por. Lissowski i in. 2008: 540),

gdzie μ_j oznacza wartość średniej arytmetycznej w *j*-tym zespole populacji, natomiast μ jest średnią populacyjną.

Wiadomo już, że miernik przyrostu wariancji będący konsekwencją losowania zespołowego definiuje się w sposób analogiczny do tego, który odnosi się do doboru próby z populacji rozwarstwionych, tzn. określa się go jako iloraz wariancji estymatorów z prób zespołowych oraz wariancji tych estymatorów, które otrzymałoby się z pomiaru prób prostych (por. rozdział II.1.2.)[124]. Przeprowadzając niezwykle proste przekształcenia arytmetyczne wzorów (IV.17.) oraz (IV.18.), miarę przyrostu wariancji dla estymatorów parametrów wskaźników struktury wyrazić można wzorem:

(IV.19.) $$\mathrm{DEFF}_c = \frac{\left(L-\frac{1}{m}\right)\sum_{j=1}^{M}(p_j-p)^2}{(M-1)p(1-p)},$$

natomiast dla estymatorów średnich arytmetycznych – jako:

(IV.20.) $$\mathrm{DEFF}_c = \frac{L\sum_{j=1}^{M}(\mu_j-\mu)^2}{(M-1)\sigma^2},$$

gdzie σ^2 oznacza wielkość parametru wariancji w całej populacji. Oczywiście, ponieważ wymaga się tutaj wiedzy o wartościach parametrów populacyjnych (tj. średnich zespołowych oraz zespołowych wskaźników struktury, średniej ogólnej oraz wskaźnika struktury w populacji, a także wariancji zmiennych

---

[124] Oczywiście porównuje się tu próby o jednakowych liczebnościach. Wielkość próby można wyrazić w postaci iloczynu liczności dobranych zespołów oraz ich liczby, tzn. wzorem $n = m \cdot L$. Właściwość tę wykorzystuje się we wzorach (IV.19.) oraz (IV.20.), ustalając wariancję estymatorów w próbach prostych.

poddanych pomiarowi), a wielkości te są zazwyczaj nieznane, to wartość miary $DEFF_c$ daje się jedynie oszacować na podstawie danych empirycznych.

Pomimo tych ograniczeń formuły (IV.19.) oraz (IV.20.) pozwalają określić czynniki warunkujące mniejszą lub większą efektywność losowania zespołowego w porównaniu z efektywnością prostego doboru indywidualnego. Wyprzedzając w tym momencie ustalenia o charakterze formalnym, można odwołać się do rozważań Grovesa (1989), który w odniesieniu do schematu doboru zespołowego stwierdza, iż:

> W praktyce próby zespołowe dają zazwyczaj większe błędy losowania [...] niż próby jednostkowe o tych samych liczebnościach [...]. Jednakże nie ma nic takiego w tym schemacie, aby ze względów statystycznych [przyrost wariancji – P.J.] był nieuchronny. Utrata precyzji estymacji wynika z socjologicznego faktu, że w sposób naturalny zespoły grupują jednostki [...] podobne do siebie pod względem wartości wielu zmiennych uwzględnionych w badaniu. (por. Groves 1989: 259–260).

Ten fenomen naturalnego – jak nazywa je Groves – grupowania w zespołach jednostek o cechach jednorodnych znajduje zastosowanie w szczególnym sposobie definiowania miernika $DEFF_c$. Zgodnie z oryginalnym pomysłem L. Kisha (1965: 170–173) wskaźnik ten wyraża się dość często za pomocą tzw. współczynnika korelacji wewnątrzzespołowej. Chociaż miara ta (w odniesieniu do doboru zespołowego) zostanie zdefiniowana dopiero w dalszej części rozdziału, to jednak w świetle przedstawionych wyżej definicji można już wskazać, że im większe będą zróżnicowania międzyzespołowych średnich oraz wskaźników struktury (lub estymatorów innych parametrów), tym mniejsza będzie efektywność schematu doboru zespołowego w porównaniu z losowaniem prób prostych o takiej samej liczebności. Wniosek ten jest odmienny od tego, który sformułowano w odniesieniu do losowania warstwowego. Jest to całkowicie zgodne z intuicjami, bowiem w schematach doboru z populacji rozwarstwionych chodzi o to, aby warstwy były jak najbardziej homogeniczne (wówczas efektywność doboru jest największa, a rozwarstwienie najbardziej uzasadnione), natomiast w losowaniu zespołowym jednorodność zespołów jest już czynnikiem obniżającym efektywność schematu doboru próby. Zauważyć można też, że im liczniejsze będą takie zespoły (a tym samym liczba dobieranych zespołów mniejsza), tym schemat ten będzie mniej efektywny. Zresztą do takich samych konkluzji dochodzi H. Weisberg (2005), który w przywoływanej już wcześniej monografii *The Total Survey Error Approach. A Guide to The New Science of Survey Research*, w części poświęconej doborowi prób badawczych stwierdza:

> efekt schematu doboru próby w losowaniu zespołowym zależy od różnicy pomiędzy zespołową przeciętną oraz przeciętną całej populacji, od heterogeniczności wiązek, a także od liczby wylosowanych zespołów. (por. Weisberg 2005: 243)

Omówione tu właściwości stają się jeszcze bardziej widoczne po zapisaniu miernika DEFF_c w nieco innej postaci. Bez utraty ogólności dla formułowanych wniosków można przyjrzeć się już wyłącznie estymatorom średnich arytmetycznych. Ponieważ daje się łatwo wykazać, że licznik wyrażenia (IV.20.) jest równoważny formule[125]:

$$\text{(IV.21.)} \qquad \mathrm{L}\sum_{j=1}^{M}\left(\mu_j-\mu\right)^2 \equiv \left(M-1\right)\sigma^2 + \left(L-1\right)M\left(\sigma^2-\bar{\sigma}^2\right),$$

(gdzie $\bar{\sigma}^2 = \frac{1}{M}\sum_{j=1}^{M}\sigma_j^2$ jest przeciętną z wewnątrzzespołowych wariancji), to w losowaniu prostym zespołów równolicznych miernik przyrostu wariancji przyjmuje postać równania:

$$\text{(IV.20'.)} \qquad \mathrm{DEFF}_c = 1 + \left(L-1\right)\frac{M}{M-1}\frac{\sigma^2-\bar{\sigma}^2}{\sigma^2}.$$

Formuła ta ukazuje już jednoznacznie, że estymator z próby zespołowej będzie miał większą (lub odpowiednio mniejszą) efektywność od estymatora z próby prostej o tej samej liczebności, jeżeli tylko przeciętna wewnątrzzespołowa wariancja będzie większa (lub mniejsza) od wariancji w całej populacji. Jeśli natomiast średnia zespołowa wariancja równa będzie wariancji populacyjnej (tzn. gdy cała zmienność ulokowana zostanie wewnątrz zespołów, a nie między nimi), to dobór zespołowy będzie miał efektywność losowania indywidualnego[126] i to niezależnie od liczebności zespołów, czy też w ogóle od liczby zespołów w całej populacji. Można również zauważyć, iż wylosowanie większej liczby zespołów prowadzi do większej precyzji estymacji, co pozostaje prostą konsekwencją tego, że iloraz $\frac{M}{M-1}$ jest asymptotycznie zbieżny do jedności (por. Barnett 1982: 169–170; Weisberg 2005: 243).

---

[125] Przekształcenie licznika wyrażenia (IV.20.) do postaci równania (IV.21.) wymaga przeprowadzenia kilku prostych operacji arytmetycznych na wariancji populacyjnej. Ponieważ ani te przekształcenia, ani też miernik DEFF_c w zaprezentowanej poniżej formule nie pojawiają się w żadnej z przywoływanych pozycji literaturowych, przytaczam poniżej wyprowadzenie formuły (IV.21.) oraz (IV.20'.). W pierwszej kolejności trzeba zauważyć, że w losowaniu prostym zespołów o jednakowych liczebnościach otrzymuje się: $L\sum_{j=1}^{M}\left(\mu_j-\mu\right)^2 \equiv (ML-1)\sigma^2 + M(L-1)\bar{\sigma}^2$, gdzie $\bar{\sigma}^2 = \frac{1}{M}\sum_{j=1}^{M}\sigma_j^2$. Dowód tego jest bezpośredni, tzn. z definicji wariancji wynika, iż $\sigma^2 \equiv \frac{1}{ML-1}\sum_{j=1}^{M}\sum_{i=1}^{L}\left(x_{ji}-\mu\right)^2 = \frac{1}{ML-1}\sum_{j=1}^{M}\sum_{i=1}^{L}\left(x_{ji}-\mu_j+\mu_j-\mu\right)^2 = \frac{1}{ML-1}\left[M(L-1)\bar{\sigma}^2 + L\sum_{j=1}^{M}\left(\mu_j-\mu\right)^2\right]$, co daje $L\sum_{j=1}^{M}\left(\mu_j-\mu\right)^2 \equiv (ML-1)\sigma^2 + M(L-1)\bar{\sigma}^2$. Po jednoczesnym odjęciu oraz dodaniu po prawej stronie tego równania wielkości $(M-1)\sigma^2$, wyrażenie $L\sum_{j=1}^{M}\left(\mu_j-\mu\right)^2$ można wyrazić już w postaci (IV.21.).

[126] Zresztą podobnie będzie, gdy $L = 1$. Jest to oczywiste, gdyż w takiej sytuacji dobór zespołowy jest *de facto* doborem indywidualnym. W nieco innym kontekście właściwość tę wykorzystuje się w schematach losowania dwustopniowego, a dokładnie w adresowych próbach osób z doborem jednego przedstawiciela każdego zespołu. Działanie takie pozwala wyeliminować efekt uzespołowienia próby, lecz jednocześnie skutkuje zróżnicowaniem szans selekcji (kwestie te podjęte będą w kolejnej części rozdziału).

Definiowanie miary przyrostu wariancji w schematach losowania zespołowego poprzez wyrażenie jej wzorem (IV.20'.) nie znajduje jednak szczególnego zastosowania praktycznego. Zauważyć wystarczy, że nawet jeśli znane są populacyjne wielkości *L* czy też *M*, to już parametry wariancji zespołowych oraz wariancji populacyjnej pozostają zazwyczaj niewiadome. Wielkości te można by naturalnie oszacować *post hoc* na podstawie wyników przeprowadzonego pomiaru (traktując oszacowanie całkowitej zmienności wyników jako estymator wariancji w prostej próbie losowej – por. Groves i in. 2004: 105), jednak nawet przy takim założeniu możliwość zastosowania formuły (IV.20'.) byłaby bardzo ograniczona. Zakłada ona bowiem równoliczność zespołów w całej populacji, te natomiast warunku takiego zazwyczaj nie spełniają.

Konieczne jest więc zdefiniowanie miernika przyrostu wariancji na skutek losowania zespołowego w formie umożliwiającej jego praktyczną implementację. Jak już wspomniano, L. Kish (1965: 162) zdefiniował miernik efektywności losowania zespołowego poprzez tzw. współczynnik korelacji wewnątrzzespołowej[127]. Miara DEFF_c przyjmuje wtedy postać doskonale znanego wyrażenia:

$$\mathrm{DEFF}_c = 1 + \left(l' - 1\right)\rho\,, \tag{IV.22.}$$

gdzie:

- ρ jest współczynnikiem korelacji wewnątrzklasowej, którego wielkość można oszacować za pomocą procedury analizy wariancji[128];
- natomiast l jest średnią liczebnością zespołu (por. Kish 1987; Kish i in. 1974: 7) lub też wielkością estymowaną w inny sposób, w zależności od

---

[127] O współczynniku tym wspomniano już w rozdziale II, analizując przyrost wariancji na skutek efektu ankieterskiego, czy też rozważając konsekwencje wynikające z uchybień w kodowaniu danych wynikowych. Wszystkie te współczynniki opierają się na propozycji L. Kisha (1965). Oryginalnie miara korelacji wewnątrzzespołowej podana została w formie właściwej dla losowania zespołów równolicznych (por. Kish 1965: 171); taką też można ją odnaleźć w wielu pozycjach literaturowych (por. np. Gabler i in. 2008: 194; Barnett 1982: 170). Z kolei formułę pozwalającą na wyznaczenie miary korelacji wewnątrzzespołowej dla przypadku zespołów o nierównych liczebnościach odnaleźć można w opracowaniu Maurice'a G. Kendalla oraz Alana Stuarta (1979). Współczynnik korelacji międzyzespołowej nie bierze pod uwagę zróżnicowania wszystkich par obserwacji (tak jak w klasycznym współczynniku korelacji), ale jedynie pary elementów wewnątrz zespołów (por. Dorofeev i in. 2006: 95).

[128] Sposoby oszacowania współczynnika korelacji wewnątrzzespołowej poprzez procedurę analizy wariancji omówiono już w rozdziale II (por. wzory II.22. – dla korelacji wyników pomiaru w obrębie ankieterów) oraz II.25'. – dla korelacji wyników w obrębie osób kodujących wyniki badań) powołując się w tym względzie na studia Gablera i in. (2008: 196), Ukoumunne (2002: 3760), a także Grovesa (1989: 363–364). Odpowiednich formuł nie warto ponownie przywoływać, ważne jest jednak to, że – jak podaje Gabler i in. (2008: 196) powołując się na symulacyjne analizy danych przeprowadzone przez Sudhira Paula i in. (2003: 507–523) – estymacja współczynnika korelacji wewnątrzklasowej poprzez procedurę *ANOVA* jest asymptotycznie nieobciążona, efektywna (tj. o najmniejszej wariancji) oraz zgodna (tj. asymptotycznie zbieżna do parametru ρ). Taki sposób estymacji współczynnika korelacji wewnątrzklasowej zastosowano zresztą w badaniach ESS-u (por. Gabler i in. 2008: 197).

charakteru działań podejmowanych w trakcie losowania próby, np. od stratyfikacji populacji lub ważenia danych (por. Dorofeev i in. 2006: 95; Gabler i in. 2006: 115–120; Lynn i in. 2005: 101–104; Gabler i in. 1999: 105–106). Praktyczne konsekwencje wynikające z zastosowania różnych metod estymacji wielkości *l'* przeanalizowane będą dokładniej w ostatniej części tego rozdziału poprzez odwołanie się do danych z polskiej części piątej rundy badań Europejskiego Sondażu Społecznego.

Powracając do oceny efektywności schematu losowania zespołowego o równolicznych zbiorach jednostek, można teraz – wykorzystując współczynnik korelacji wewnątrzzespołowej – zapisać formuły (IV.20.) oraz (IV.20'.) w postaci (por. Gabler i in. 2008: 194):

$$\text{(IV.20''.)} \qquad \text{DEFF}_c = \frac{ML-1}{L(M-1)}\left[1 + (L-1)\rho\right] \approx 1 + (L-1)\rho,$$

gdzie: $\rho = 1 - \frac{ML}{ML-1}\frac{\bar{\sigma}^2}{\sigma^2}$ (por. Kish 1965: 171). Formuła ta ukazuje, iż losowanie zespołowe będzie miało efektywność doboru prostego o tej samej liczebności próby, o ile tylko wielkość $\rho \approx 0$, lub, inaczej, gdy przeciętna wariancja międzyzespołowa będzie równa wariancji populacyjnej[129]. Z kolei, ponieważ wartość $\rho > 0$ świadczy o tym, że jednostki populacji są bardziej zróżnicowane między zespołami niż w ich obrębie[130], to losowanie zespołowe ma w takiej sytuacji efektywność mniejszą od doboru prostego. Natomiast jeśli tylko zespoły są bardziej zróżnicowane wewnętrznie niż między sobą, to wartości $\rho < 0$, a efektywność losowania zespołowego jest większa od prostego (por. Groves 1989: 261–262).

Wzajemne układy odniesienia wewnątrz-zespołowego i między-zespołowego zróżnicowania jednostek populacji, a także oddziaływanie tych zróżnicowań na wariancję estymatorów parametrów populacyjnych ukazuje, iż podstawową trudnością w schematach losowania wielostopniowego z doborem wiązek respondentów (jako reprezentantów zespołów)[131], a w mniejszym stopniu również w próbach opartych na rejestrach adresowych (gospodarstw domowych lub budynków mieszkalnych)[132], jest wyznaczenie takiej liczby zespołów oraz

---

[129] W losowaniu zespołów równolicznych DEFF_c=1 będzie tak, o ile tylko $\rho = -\frac{1}{ML-1}$ (por. Kish 1965: 171). W takiej sytuacji każdy zespół może zostać uznany za losową reprezentacją populacji (por. Lissowski i in. 2008: 541). Co oczywiste, również dla L=1 otrzyma się wartość DEFF_c=1. Nie ma się wtedy do czynienia z próbą zespołową, ale indywidualną. Podobnie jest zresztą w dwustopniowym losowaniu adresowych prób osób, co jest prostą konsekwencją tego, że l^*=1.

[130] W losowaniu zespołów równolicznych będzie to $\rho > -\frac{1}{ML-1}$.

[131] Jak już wiadomo, w losowaniu wiązek o takiej samej liczbie elementów wielkość próby można wyrazić iloczynem liczebności zespołów oraz liczności dobieranych wiązek.

[132] W próbach adresowych rozważyć można dwie główne strategie losowania reprezentantów zespołu. Pierwsza polega na doborze jednego członka w obrębie każdego gospodarstwa do-

ich liczności, aby precyzja estymacji osiągnęła pożądaną przez badacza wielkość. Problem ten przypomina wybór optymalnej lokalizacji próby w schemacie losowania stratyfikacyjnego, w takim jednak znaczeniu, że sprowadza optymalizację doboru zespołowego do kryterium maksymalizacji precyzji prowadzonego pomiaru przy ustalonych całkowitych kosztach badania. Można zresztą przypuszczać, że przy znaczącej homogenizacji zespołów (która, jak wiadomo, skutkuje przyrostem wariancji), bardziej zasadnym działaniem będzie dobór mniejszej liczby respondentów w każdej wiązce (co przekładać się będzie na konieczność losowania większej liczby takich wiązek), inaczej niż przy niewielkiej zespołowej homogenizacji jednostek, dla której bardziej efektywnym działaniem będzie dobór zespołów o większej liczebności (co pozwoli ograniczyć liczbę losowanych zespołów). Intuicyjne przypuszczenia daje się stosunkowo łatwo potwierdzić prostymi układami analitycznymi. Wprawdzie zostały one szczegółowo przedstawione w pracach L. Kisha (1965: 268–272), W. Cochrana (1977: 280–285), czy też R. Grovesa (1989: 262–263), to jednak warto – przynajmniej fragmentarycznie – odnieść się do ich podstawowych założeń oraz scharakteryzować zasadę działania procedur optymalizacyjnych w losowaniu zespołowym.

W pierwszej kolejności można wskazać, że jednym ze sposobów wyrażenia całkowitych kosztów badania w doborze zespołowym jest zapisanie ich za pomocą następującej funkcji:

(IV.23.) $C = c_1 m + c_2 ml$ (por. Groves 1989: 262),

gdzie:

- C jest całkowitym kosztem badania sondażowego z wyłączeniem stałych kosztów organizacyjnych,
- c_1 jest jednostkowym kosztem realizacji badania z całą wylosowaną wiązką respondentów,
- c_2 jest kosztem realizacji wywiadu z każdą wylosowaną osobą,
- m oznacza liczbę dobranych wiązek,
- l oznacza stałą liczbę jednostek w każdej wiązce.

---

mowego (dobór dwustopniowy), druga natomiast na przeprowadzeniu badań ze wszystkimi osobami zamieszkującymi takie gospodarstwa (jednostopniowy dobór całego zespołu). Oba działania prowadzą jednak do odmiennych skutków. Losowanie dwustopniowe zróżnicuje szanse selekcji jednostek do próby badawczej, natomiast realizacja wywiadów z każdym członkiem gospodarstwa domowego obniży precyzję estymacji na skutek wewnątrzzespołowej homogenizacji uzyskiwanych wartości pomiarowych. Kwestie te podjęte będą w następnej sekcji rozdziału, w ramach analizy empirycznych konsekwencji losowania prób z nierównymi prawdopodobieństwami selekcji. W tym momencie można już jednak wskazać, że „losowanie dwustopniowe będzie bardziej efektywne od losowania zespołowego, jeśli [współczynnik wewnątrzzespołowej homogenizacji – P.J.] będzie dodatni" (Aliaga i in. 2006: 6).

Chociaż nie jest to jedyny sposób ujmowania kosztów badań prowadzonych na próbach uzespołowionych (por. np. Aliaga i in. 2006: 15–17), to jednak w literaturze metodologicznej pojawia się zdecydowanie najczęściej. Optymalną liczebność wiązek dobieranych w ramach każdego zespołu wyznacza się wówczas z następującego układu równań[133]:

$$\begin{cases} l_{\text{opt}} = \sqrt{\frac{c_1(1-\rho)}{c_2 \cdot \rho}}, \text{jeżeli } \rho > 0 \\ l_{\text{opt}} = L, \text{jeżeli } \rho \leq 0 \end{cases}, \tag{IV.24.}$$

a z równania:

$$m_{\text{opt}} = \frac{c}{c_1 + c_2 \cdot l_{\text{opt}}} \tag{IV.25.}$$

ustala się optymalną liczbę wiązek. Przyjmując, że iloczyn kosztów c_1 oraz c_2 jest większy od jedności (por. Aliaga i in. 2006: 7), można zauważyć, iż wielkość l_{opt} zależy przede wszystkim od stopnia podobieństwa jednostek w obrębie zespołów. W przypadku ich znacznej wewnątrzzespołowej homogenizacji (tj. dla $\rho \approx 1$), optymalnym rozwiązaniem[134] okazuje się losowanie wyłącznie jednego przedstawiciela z każdego zespołu[135]. Z drugiej strony, im zespoły są bardziej zróżnicowane, tym korzystniejszy jest dobór wiązek respondentów o większej liczebności, czyli losowanie mniejszej liczby zespołów, ale bardziej licznych (por. Fahimi 2008: 98). W skrajnych przypadkach, tj. dla $\rho \leq 0$ (heterogeniczność zespołów) najlepszym rozwiązaniem jest badanie wszystkich elementów wylosowanych zespołów (por. Aliaga i in. 2006: 4).

Niezwykle wymownym potwierdzeniem tych zależności są ustalenia G. Lissowskiego i in. (2008: 544) zawarte w 10. rozdziale podręcznika *Podstawy statystyki dla socjologów*. Analizując schematy losowania zespołowego z doborem wiązek respondentów o jednakowych liczebnościach, autorzy wspomnianego opracowania zamieszczają jednoznaczną konkluzję dotyczącą optymalnej alokacji jednostek losowania pierwszego (zespołów) oraz drugiego stopnia (wiązek jednostek w obrębie zespołów). Wychodząc od rozważań dotyczących mię-

---

[133] Formuły (IV.24.) oraz (IV.25.) odnaleźć można w niezwykle ciekawym opracowaniu autorstwa Alfredo Aliagi oraz Ruilina Rena z 2006 roku pt. *Optimal Sample Sizes for Two-stage Cluster Sampling in Demographic and Health Surveys*. W tekście tym podano przykład praktycznej implementacji procedur optymalizacyjnych w losowaniu prób do badań zdrowia ludności. Autorzy zamieszczają też tablice statystyczne w których, dla różnych wielkości iloczynu c_1/c_2 oraz wartości ρ współczynnika korelacji wewnątrzzespołowej, podają optymalne liczebności wiązek respondentów (por. Aliaga i in. 2006: 7).

[134] O ile tylko iloczyn kosztów jest równy jedności lub gdy koszty nie są w ogóle kryterium uwzględnianym przy optymalizacji schematu losowania.

[135] Przykładem takiej strategii losowania są adresowe próby gospodarstw domowych. Liczba dobieranych zespołów równa jest liczebności próby.

dzyzespołowego oraz wewnątrzzespołowego zróżnicowania elementów populacji, stwierdzają:

> dodatkowo, wielkość wariancji estymatora zależy od liczby wylosowanych do próby zespołów i liczby elementów wylosowanych w każdym zespole […]. Jeżeli wariancja między zespołami jest większa od średniej wariancji w zespole, wówczas najkorzystniejsze jest wylosowanie większej liczby zespołów i jak najmniejszej liczby elementów w każdym zespole. Jeżeli jednak wariancja między zespołami jest mniejsza od wariancji w zespole (odpowiada to w przybliżeniu ujemnej wartości współczynnika korelacji wewnątrzzespołowej), najbardziej korzystna będzie sytuacja, w której badane będą wszystkie elementy w wylosowanych do próby zespołach […]. (Lissowski i in. 2008: 544)

Pamiętać należy jednak, że możliwość ustalenia optymalnej wielkości wiązki respondentów w losowaniu zespołowym ma swoje poważne ograniczenia praktyczne, co sprawia, że rzadko kiedy wykorzystywana jest w empirii[136]. Po pierwsze, ponieważ analizowane w surveyach zmienne charakteryzują się zazwyczaj bardzo zróżnicowanymi wielkościami współczynników korelacji wewnątrzklasowej (por. Sawiński 2011), „każdy z estymatorów będzie się cechował odmienną wielkością efektu schematu [doboru zespołowego – P.J.]" (Groves i in. 2004: 105), co jest przypadłością o charakterze uniwersalnym – problem ten został już zaznaczony w niniejszej książce w kontekście zagadnienia optymalizacji losowania stratyfikacyjnego. Innymi słowy, rozwiązanie optymalne dla jednej zmiennej nie musi już być najlepsze dla innej, a zatem minimalizacja wariancji jednego estymatora może skutkować przyrostem wariancji innego. Po drugie, ponieważ wielkości współczynników ρ estymowane są najczęściej dopiero *a posteriori* (por. Aliaga i in, 2006: 5–8), to ustalenie l_{opt} z przyjętych *a priori* wielkości wewnątrzzespołowej homogenizacji jednostek okazuje się zazwyczaj ryzykowne i prowadzi do niepewnych rozwiązań (por. Groves 1989: 263).

Problem polega też na tym, że wyłącznie w nielicznych przypadkach można oczekiwać, iż wszystkie – bez wyjątku – zespoły populacji, składać się będą z takiej samej liczby elementów. Oznacza to, że przedstawione dotąd schematy

---

[136] W podręczniku *Designing Household Survey Samples: Practical Guidelines* wydanym w 2005 roku przez Departament Spraw Gospodarczych i Społecznych (DESA) Organizacji Narodów Zjednoczonych w ramach serii wydawniczej poświęconej studiom metodologicznym badań sondażowych, odnaleźć można kilka praktycznych rad, których zastosowanie pozwala zminimalizować skalę przyrostu wariancji estymatorów w schematach doboru zespołowego. W rozdziale trzecim tego podręcznika, w części poświęconej minimalnym licznościom prób badawczych, zamieszczono notę informującą, iż najkorzystniejsze jest (a) losowanie jak największej liczby zespołów, (b) o możliwie najmniejszej liczbie elementów oraz (c) dobór jednakowej liczby jednostek w każdym zespole. Wnioski te wynikają wprost z opisanych procedur optymalnego doboru liczebności jednostek losowania pierwszego oraz drugiego stopnia. Podążanie za zasadami podanymi w przywołanym podręczniku będzie tym bardziej efektywne, im bardziej jednostki wchodzące w skład zespołów będą homogeniczne.

jednostopniowego losowania zespołów równolicznych są rzadko wykorzystywane w praktyce badawczej. Ponieważ jednak walory schematu doboru zespołowego – zwłaszcza w zakresie redukcji kosztów w badaniach prowadzonych techniką wywiadów bezpośrednich na populacjach o znacznym rozproszeniu terytorialnym – wydają się bezsprzeczne, to uzespołowienie populacji pozostaje jednym z częściej wykorzystywanych schematów doboru sondażowych prób badawczych. Wystarczy powołać się na dokumentację metodologiczną projektu ESS, by zauważyć, że elementy schematu losowania zespołowego oraz wiązkowania jednostek populacji (poprzedzone zazwyczaj stratyfikacją), znajdują zastosowanie w większości krajów biorących udział w badaniach ESS-u. Przyglądając się raportom metodologicznym z piątej rundy tych badań można przykładowo wskazać, że schemat losowania zespołowego wykorzystano w 19 z 26 krajów uczestniczących w tych badaniach (por. *ESS5-2010*, *Documentation Report.* ed. 2.0).

Z wielu możliwych do zastosowania wariantów wielostopniowych schematów doboru prób badawczych zawierających elementy losowania zespołowego warto zwrócić uwagę zwłaszcza na dwa przypadki, mające szersze zastosowanie praktyczne. Jest to: (a) losowanie proste zależne (bez zwracania) zarówno zespołów, jak i elementów z wylosowanych do próby zespołów oraz (b) losowanie zwrotne zespołów z prawdopodobieństwem proporcjonalnym do ich liczebności wraz z bezzwrotnym doborem elementów w obrębie zespołów. W podręczniku *Podstawy statystyki dla socjologów* wskazano, że choć w obu tych przypadkach liczby elementów (tj. wiązek) dobieranych w ramach zespołów mogą być dowolne, to jednak istnieją takie dwa sposoby wyznaczania ich liczebności, aby prowadzona estymacja była statystycznie nieobciążona. Odpowiednio zatem, w odniesieniu do pierwszego wariantu losowania, można podać, że liczebności jednostek dobieranych w każdym zespole (wylosowanym co najwyżej jeden raz) musiałyby być proporcjonalne do wielkości każdego zespołu, w drugim natomiast, w ramach zespołów losowanych jednokrotnie lub wielokrotnie z prawdopodobieństwem proporcjonalnym do ich liczebności należałoby dobierać taką samą liczbę elementów, zwielokrotnioną oczywiście o częstość wylosowania danego zespołu (por. Lissowski i in. 2008: 547–548). Ten drugi sposób dobierania próby zbliżony jest zresztą do schematów zastosowanych w wielu krajach uczestniczących w badaniach ESS-u[137] (por. Lynn i in. 2007: 116–122; Gabler i in. 2006: 117–118; Lynn i in. 2005: 103–104). W takich sytuacjach zakłada się, że z populacji liczącej M zespołów (np. miejscowości), z których każdy składa się z N_j (j = 1, 2, …, M) elementów (mieszkańców), do-

[137] W obrębie wylosowanych zespołów (najczęściej miejscowości) dobiera się wiązki respondentów (o takiej samej liczebności) z prawdopodobieństwem proporcjonalnym do liczebności zespołów.

biera się najpierw w sposób zwrotny[138] *m* zespołów (wsi lub miast), z prawdopodobieństwem proporcjonalnym do wielkości wylosowanego zespołu (tj. do wartości $\frac{N_j}{N}$), a następnie, w ramach każdego zespołu, losuje się w sposób prosty, bez zwracania, jednakową liczbę jednostek (tj. wiązki respondentów o takiej samej liczebności)[139]. Najbardziej charakterystycznym trudnościom związanym z szacowaniem mierników $DEFF_c$ oraz $DEFF_{TOTAL}$ w takich złożonych schematach doboru sondażowych prób badawczych poświęcona będzie ostatnia część tego rozdziału.

IV.1.3. Losowanie z nierównymi prawdopodobieństwami doboru

Chociaż równość prawdopodobieństw selekcji jednostek jest warunkiem *sine qua non* losowania zgodnego z zasadą doboru próby prostej, to jednak nawet w bardziej złożonych schematach losowania sondażowych prób badawczych udaje się bez większego wysiłku wybrać jednostki, zapewniając jednakowość szans ich selekcji (por. Biemer i in. 2003: 346–347). Wystarczy przy tym przywołać schemat losowania próby z populacji rozwarstwionych (w którym liczebności podprób są proporcjonalne do wielkości warstw w populacji) lub też nawet schemat doboru wielostopniowego zespołowego (z prawdopodobieństwem losowania zespołów proporcjonalnym do wielkości zespołowych populacji (por. Chromy 2008: 619–621), by zauważyć, iż problem nie tkwi w tym, że badacz napotyka na jakieś szczególne trudności w zapewnieniu doboru z równymi prawdopodobieństwami selekcji, ale w tym, że takie schematy losowania są rzadko przez niego pożądane.

Przyglądając się ustaleniom literaturowym oraz praktyce badawczej, można zauważyć, iż zróżnicowanie szans selekcji jednostek dobieranych do prób sondażowych jest przede wszystkim konsekwencją trzech klas działań, w tym: (a) wielostopniowego doboru próby z operatów zespołowych (adresowych) o nierównych liczebnościach zespołów, zamiast losowania indywidualnego[140],

---

[138] Zespoły mogą być wylosowane wielokrotnie, co w rzeczywistości oznacza, że każdemu z nich przypisuje się wiązki jednostek w liczbie odpowiadającej wielokrotnościom doboru.

[139] W sytuacji wielkokrotnego wylosowania danego zespołu (np. miejscowości), dobiera się za każdym razem *l* elementów w sposób prosty. Liczebność próby badawczej można wyrazić formułą $n = m \cdot l$. Wariancja nieobciążonego estymatora wartości średniej w populacji pozostaje wówczas sumą (a) wariancji wynikającej z doboru zespołowego oraz (b) wariancji dodatkowego składnika związanego z tym, że nie bada się wszystkich elementów wylosowanych zespołów, lecz jedynie ich próbę (por. Lissowski i in. 2008: 547).

[140] Typowym przykładem strategii badawczych mieszczących się w pierwszej klasie działań są takie schematy losowania, w których próba dobierana jest wielostopniowo z operatu zespołowego. Z taką sytuacją ma się do czynienia w adresowych próbach osób, gdzie losuje się gospodarstwa domowe lub budynki mieszkalne. W tym pierwszym przypadku respondentów dobiera się

(b) zastosowania pewnych specyficznych procedur ograniczających błędy niepełnego pokrycia populacji określonymi operatami doboru[141] oraz (c) losowania próby z populacji rozwarstwionych z nieproporcjonalnym rozlokowaniem jednostek w obrębie zdefiniowanych warstw populacji docelowej. Choć nie są to wszystkie możliwe działania skutkujące niejednakowymi szansami selekcji jednostek populacji do próby[142], to jednak losowanie zespołowe, warstwowe oraz działania podejmowane na rzecz poprawy jakości operatów odgrywają tutaj rolę najbardziej znaczącą. Zresztą w procesie doboru sondażowych prób badawczych w mniejszym stopniu chodzi o to, aby jednostki losowane były bezwzględnie z jednakowymi prawdopodobieństwami selekcji, ale by w ogóle owe szanse doboru były znane[143] (por. Dorofeev i in. 2006: 27). Ów, jak mogłoby się wydawać, zaskakujący wniosek ma jednak proste uzasadnienie. Otóż, ponieważ nierówne prawdopodobieństwa losowania daje się zrównoważyć poprzez ważenie przypisujące każdej jednostce wartość będącą odwrotnością szans jej selekcji (por. Biemer i in. 2003: 347) – co wprawdzie skutkuje często przyrostem wariancji estymatorów[144] (por. np.: Lee 2012: 17; Kish 1992: 190–192), ale jednocześnie eliminuje też błąd systematyczny estymatorów (por. Dorofeev i in. 2006: 27) – brak wiedzy o prawdopodobieństwach losowania

---

spośród wszystkich członków gospodarstwa domowego spełniających określone przez badacza kryteria, w drugim natomiast najpierw dobiera się gospodarstwo domowe, a następnie respondenta. Ponieważ szansa wylosowania jednostki (lub wylosowania gospodarstwa domowego, a następnie osoby) zależy od wielkości zespołu, to nierówne szanse selekcji należy zrekompensować poprzez ważenie danych.

[141] Spośród omówionych w poprzednim rozdziale procedur ograniczających błędy niepełnego pokrycia populacji operatami doboru próby wskazać należy na dwie metody, które w sposób oczywisty różnicują szanse doboru jednostek. Pierwszą z nich jest sieciowanie (każda jednostka ma takie szanse doboru do próby, jakie wynikają z liczby członków jego najbliższej rodziny mieszkających w obszarze objętym badaniem), drugą natomiast jest procedura wielokrotnych operatów (jednostki populacji będą miały tyle szans selekcji, w ilu operatach się znajdują).

[142] W zasadzie nie sposób wymienić wszystkich działań, które potencjalnie mogą zróżnicować szanse selekcji jednostek. Ciekawą próbę systematyzacji źródeł takiego zróżnicowania odnaleźć można jednak w artykule Lynna i in. (2007: 113) poświęconym metodologicznym aspektom pierwszej rundy badania *Europejskiego Sondażu Społecznego*. Problemy te rozpatruje także Leslie Kish w opracowaniu *Weighting for Unequal* P_i (por. Kish 1992: 185–188).

[143] Niezwykle ważne jest to, iż wiedza o rzeczywistych szansach selekcji pozwala wyznaczyć skalę przyrostu wariancji jeszcze przez fazą doboru próby oraz wylosować próbę o liczebności uwzględniającej przewidywaną utratę precyzji estymacji. W wielu sytuacjach jednak (np. w wyniku zastosowania adresowych prób osób, czy też procedury operatów wielokrotnych w celu redukcji błędu niepełnego pokrycia jakiegoś operatu) szanse selekcji poszczególnych jednostek znane są najczęściej dopiero w fazie realizacji badań terenowych lub też po jej zakończeniu (czyli same wymagają empirycznego oszacowania). W takich sytuacjach wyznaczanie wielkości próby odpowiadającej zakładanej efektywności doboru musi być oparte na danych historycznych.

[144] O ile oczywiście nie ma się do czynienia z losowaniem warstwowym z optymalną lokalizacją próby w warstwach, która choć różnicuje szanse selekcji, to jednak pozwala na zredukowanie wariancji estymatorów właśnie poprzez zoptymalizowane (często nieproporcjonalne) przyporządkowanie jednostek do warstw.

uniemożliwia przeprowadzenie operacji ważenia. W takim przypadku doszłoby zawsze do kierunkowego zniekształcenia uzyskiwanych wyników, gdyż estymatory parametrów populacyjnych byłyby statystycznie obciążone[145] (por. Berger i in. 2009: 39–40).

Jeśli zatem schemat losowania próby badawczej (lub operat wykorzystany do jej doboru) nie zapewnia równości prawdopodobieństw selekcji, to wszystkim wylosowanym jednostkom – których rzeczywiste szanse doboru wynoszą π_i – należałoby przyporządkować wagi równe:

(IV.26.) $$w_i = \frac{1}{\pi_i}, \text{ gdzie } i \in \{1, 2, \dots, n\},$$

lub alternatywnie:

(IV.26'.) $$\widetilde{w}_i = n\frac{w_i}{\sum_{i=1}^{n} w_i}.$$

W tym drugim przypadku „surowe" wartości wag w_i zostają znormalizowane do wielkości próby, tzn. są określone tak, że suma wszystkich $\widetilde{w}_i$ daje liczebność próby badawczej[146]. Warto zaznaczyć, że zastosowanie obu procedur daje te same wartości estymatorów ważonych oraz w taki sam sposób wpływa na precyzję estymacji.

Przy tak zdefiniowanych wagach estymator przeciętnej populacyjnej będzie nieobciążony, o ile tylko przyjmie postać zaproponowaną przez Horwitza-Thomsona (1952). Zwracano na to uwagę w rozdziale III, rozpatrując konsekwencje wynikające z działań ograniczających błędy pokrycia populacji poprzez wykorzystanie wielu operatów lub sieciowanie jednostek. Wskazywano wówczas także na pewne nietrywialne komplikacje związane z koniecznością ustalania rzeczywistych szans selekcji jednostek, które tylko w nielicznych przypadkach dawało się wyznaczyć w sposób jednoznaczny. Abstrahując w tym momencie od prowadzonej wówczas dyskusji, należy jednak wskazać, iż skalę przyrostu wariancji, będącej efektem ważenia rekompensującego nierówne szanse selekcji jednostek z populacji do próby badawczej, można wyznaczyć przy pomocy miernika *VIF* zdefiniowanego już w rozdziale II (zob. wzór II.18.). Znacznie częściej korzysta się jednak z nieco innych (lecz równoważnych) formuł o postaci wygodnej w szacowaniu przyrostu wariancji dla danych skategoryzowanych[147] (por. Lynn i in. 2007: 112–113; Gabler i in. 1999: 105):

---

[145] Inaczej mówiąc, ich wartości oczekiwane nie byłyby równe wartościom estymowanych parametrów.

[146] Procedurę (IV.26'.) wykorzystuje się między innymi w badaniach Europejskiego Sondażu Społecznego (por. *The ESS Sample Design Data File (SDDF)* 2013: 3)

[147] Znajduje ona swoje zastosowanie w takich schematach doboru próby, w których respondentów daje się podzielić na *k* rozłącznych klas z jednakowymi wartościami wag.

(IV.27.) $$\mathrm{DEFF}_p = \frac{n\sum_{j=1}^{k} n_j w_j^2}{\left(\sum_{j=1}^{k} n_j w_j\right)^2},$$

lub nieskategoryzowanych (por. *The ESS Sample Design Data File (SDDF)* 2013: 4):

(IV.27'.) $$\mathrm{DEFF}_p = \frac{n\sum_{i=1}^{n} w_i^2}{\left(\sum_{i=1}^{n} w_i\right)^2},$$

gdzie przez *n* oznacza się zakładaną (lub zrealizowaną) wielkość próby badawczej, n_j jest liczbą jednostek z przypisaną wagą w_j, natomiast $j = \{1, 2, \ldots, k\}$ odpowiada podziałowi próby na *k* różnych klas jednostek z jednakowymi wartościami wag w obrębie każdej klasy[148].

Właściwość miernika DEFF_p warto zobrazować kilkoma przykładami empirycznymi. Przyglądając się w pierwszej kolejności danym metodologicznym Europejskiego Sondażu Społecznego, można zauważyć, iż w odniesieniu do schematów losowania próby z populacji rozwarstwionych decyzja o doborze jednostek w sposób nieproporcjonalny do liczebności warstw w populacji motywowana jest zarówno względami merytorycznymi (np. koniecznością wylosowania większej – niż wynika ze struktury populacji – liczby jednostek z tych warstw, których wielkości są zbyt małe, by ich proporcjonalny udział pozwalał na przeprowadzenie jakichkolwiek sensownych porównań między warstwami)[149], jak też kwestiami organizacyjnymi, na przykład przewidywanymi wskaźnikami realizacji próby w warstwach populacji (por. Lynn i in. 2007: 116). Co ciekawe, choć w obu przypadkach stosuje się schematy losowania z lokalizacją nieproporcjonalną, to jednak prowadzą one do zupełnie odmiennych skutków i należy je rozpatrywać rozłącznie. Otóż, o ile rzeczywiście następstwem pierwszego z tych działań jest zróżnicowanie szans selekcji oraz (potencjalnie) znaczny przyrost wariancji estymatorów, o tyle konsekwencje doboru próby z założeniem niejednakowych wskaźników realizowalności wywiadów w poszczególnych warstwach populacji okazują się już marginalne. Jest to prostą konsekwencją tego, że chociaż próba dobierana jest w sposób niepro-

---

[148] W schematach losowania z populacji rozwarstwionej, liczba *k* klas odpowiada liczbie zdefiniowanych warstw, natomiast wielkość n_j wyznaczana jest *a priori* w oparciu o schemat lokalizacji optymalnej lub jakieś inne kryterium alokacji jednostek populacji do jej warstw w próbie. Z kolei waga w_j, przypisana wszystkim jednostkom z danej warstwy, jest ilorazem proporcji tej warstwy w populacji oraz w próbie badawczej (por. Jabkowski 2011: 36–37). W schematach losowania zespołowego (np. w doborze gospodarstw domowych oraz jednostek w obrębie wylosowanych mieszkań), wartości wag przypisane jednostkom wylosowanym z *j*-tego zespołu równe są odwrotnościom liczebności owego zespołu (por. Weisberg 2005: 221).

[149] Np. w Izraelu losuje się więcej, niż wskazują na to liczebności populacji, przedstawicieli mniejszości arabskiej. Z kolei w Niemczech nadreprezentuje się obywateli zamieszkujących obszar dawnej NRD.

porcjonalny do liczebności warstw w populacji, to jednocześnie ów nierównomierny „odpad" jednostek powoduje, iż proporcje warstw w próbie zrealizowanej pozostają, mimo wszystko, proporcjonalne do rzeczywistych liczności w populacji[150]. Uwidacznia to bardzo interesującą właściwość takiego schematu doboru próby, ukazuje bowiem, iż losowanie stratyfikacyjne nieproporcjonalne może (przy spełnieniu pewnych warunków) zostać sprowadzone – *de facto* – do niezwykle efektywnego schematu warstwowego z lokalizacja proporcjonalną. Doskonałym tego potwierdzeniem są wnioski z analiz pierwszej rundy projektu ESS1–2002 zawarte w artykule P. Lynna i in. (2007: 107–124), które ukazują, iż:

> [...] warstwy mogą być nadreprezentowane proporcjonalnie do wskaźników realizowalności próby, co w zasadzie nie powinno mieć jednak wpływu na wariancję estymatorów, ponieważ doprowadza to do jednakowych prawdopodobieństw losowania, o ile tylko założone wskaźniki realizacji próby okazują się adekwatne [do wartości uzyskanych w badaniach – P.J.]. (Lynn i in. 2007: 113)

Wniosek ten jest niezwykle ważny, uświadamia bowiem, że korzyści wynikające z warstwowego doboru próby (z losowaniem jednostek nieproporcjonalnym do wielkości warstw w próbie, ale proporcjonalnym do przewidywanego *a priori* odsetka realizacji tej próby w kolejnych warstwach), będą tym większe, im bardziej przyjęte oszacowania poziomu realizacji próby w warstwach okażą się zbliżone do ich wielkości rzeczywistych. Fakt ten potwierdza po raz kolejny, iż jakość doboru sondażowych prób badawczych jest w dużej mierze uwarunkowana posiadaną wiedzą o badanej populacji.

Przechodząc ponownie do zagadnień związanych ze zróżnicowaniem szans selekcji jednostek w doborze zespołowym, należy rozpocząć od przypomnienia ustaleń z rozdziału III, a mianowicie od wskazania tego, iż dobór elementów populacji poprzez wstępne wylosowanie ich zespołów jest dość często wykorzystywanym schematem doboru próby w sytuacji niedostępności (lub niedostatecznej jakości) rejestrów indywidualnych. Schemat ten wiąże się jednak, jak już

---

[150] Ciekawą analizę zależności pomiędzy przewidywanymi wskaźnikami realizacji próby w warstwach populacji (wyodrębnionymi z uwagi na typ miejscowości) a licznością próby dobieranej w obrębie takich warstw odnaleźć można w artykule Jana Pickery oraz Ann Carton pt. *Oversampling in Relation to Differential Regional Response Rates* z 2008 roku. Rozważając schemat doboru próby w badaniach poświęconych zmianom społeczno-kulturowym we flamandzkim regionie Belgii, autorzy tego artykułu stwierdzają, iż nadreprezentacja pewnych warstw populacji w wylosowanej próbie badawczej (odwrotnie proporcjonalnie do wskaźników realizacji próby) przynosi znacznie większe korzyści dla jakości pomiaru, niż stosowanie post-stratyfikacyjnego ważenia danych, którego celem jest skorygowanie struktury próby zrealizowanej dobranej zgodnie ze schematem warstwowym z lokalizacją proporcjonalną. Wnioski te są zbieżne z ustaleniami Lynna i in. (2007: 107–124).

wiadomo, ze zróżnicowaniem szans selekcji[151], co przejawia się zazwyczaj nadreprezentacją osób zamieszkujących mało liczne gospodarstwa domowe z budynków jednorodzinnych oraz niedoreprezentacją osób z wielolicznych mieszkań w dużych blokach (por. Dorofeev i in. 2006: 30). W literaturze metodologicznej odnaleźć można dwie propozycje rozwiązania tego problemu. Pierwsza sugeruje przeprowadzenie losowania wymaganej liczby respondentów ze zbioru wszystkich osób zamieszkujących gospodarstwa domowe dobrane z operatu zespołowego (zamiast losowania tej samej liczby jednostek wewnątrz każdego zespołu). Druga natomiast, prowadzenie wywiadów ze wszystkimi – bez wyjątku – osobami zamieszkującymi każde z wylosowanych gospodarstw. Wprawdzie obie te metody pozwalają wyeliminować zróżnicowanie prawdopodobieństw selekcji jednostek, jednakże nie pozostają bez wpływu na jakość sondażu i rzadko kiedy wykorzystywane są w empirii. Pierwsza z nich wiąże się z komplikacjami natury praktycznej oraz organizacyjnej (tj. wydłużeniem czasu badań o okres potrzebny na spisanie wszystkich jednostek), powodując wzrost kosztów badań (por. Dorofeev i in. 2006: 30). Druga skutkuje znacznie niższymi wskaźnikami realizacji wywiadów oraz powoduje przyrost wariancji wynikający z wiązkowania elementów próby[152] (por. Barnett 1982: 182–187).

Do niezwykle pouczających wniosków w zakresie oceny efektywności pewnych określonych strategii doboru respondentów w adresowych próbach osób prowadzą zresztą analizy Roberta G. Clarka oraz Davida G. Steela (2002: 289–314), zamieszone w artykule *The Effect of Using Household as a Sampling Units*. Głównym celem przywoływanych studiów było znalezienie odpowiedzi na pytanie, czy bardziej zasadny będzie dobór oparty na losowaniu jednego przedstawiciela z każdego gospodarstwa domowego, czy też prowadzenie badań ze wszystkimi jego członkami (lub też z ustaloną – stałą – liczbą osób w każdym z wylosowanych zespołów)[153]. Konfrontując te procedury, autorzy

---

[151] Szansa doboru jest odwrotnie proporcjonalna do liczebności zespołu, zatem gdyby wszystkie zespoły miały jednakową liczbę członków, to prawdopodobieństwo losowania jednostek pozostałoby takie samo w całej badanej populacji.

[152] Jeśli bowiem badaniu miałyby podlegać wszystkie osoby wchodzące w skład wylosowanego gospodarstwa domowego, to wewnątrzzespołowa homogenizacja wartości zmiennych obniżałaby precyzję estymacji. Podobnie, negatywny wpływ miałaby kategoryczna odmowa udziału w badaniu wyrażona przez jednego przedstawiciela gospodarstwa domowego (np. „głowy" rodziny). Oznaczałaby ona w wielu przypadkach brak możliwości przeprowadzenia wywiadów również z innymi członkami tego gospodarstwa. Podobnie zresztą, błędne lub nieaktualne dane adresowe uniemożliwiałyby nawiązanie kontaktu ze wszystkimi przedstawicielami wylosowanych gospodarstw domowych. Odsetek zrealizowanej części próby badawczej może być zatem zdecydowanie mniejszy, niż w schemacie losowania jednego przedstawiciela całego zespołu.

[153] Autorzy przywołanego tekstu rozpatrywali także warianty pośrednie, w których zakładano losowanie dwóch, trzech oraz czterech przedstawicieli z każdego zespołu. Oszacowanie wariancji estymatorów średnich arytmetycznych dla doboru zespołowego ze stałą liczba elementów dobieranych spośród jednostek każdego zespołu odnaleźć można w podręczniku V. Barnetta (1982: 182–187). Zauważyć można jednak, iż w adresowych próbach gospodarstw domowych lub

artykułu wyszli od wyrażonej wprost obserwacji, iż „losowanie jednej osoby […] może wyeliminować efekt wiązkowania […], jednakże oznacza zróżnicowanie szans selekcji [i odwrotnie – P.J.]" (Clark i in. 2002: 297). Rezultaty badań R.G. Clarka oraz D.G. Steela, choć przewidywalne, okazały się niezwykle interesujące. Ukazują one bowiem przewagę schematu opartego na losowaniu jednego przedstawiciela z każdego gospodarstwa domowego nad prowadzeniem badań ze wszystkimi osobami zamieszkującymi wylosowane gospodarstwa (por. Clark i in. 2002: 307). Innymi słowy, potwierdziło się intuicyjne przypuszczenie, że zespołowa homogenizacja jednostek w obrębie gospodarstw domowych będzie miała dużo większy wpływ na przyrost wariancji, niż zróżnicowanie szans selekcji jednostek. Do podobnych wniosków prowadzą zresztą studia empiryczne innych autorów. Można przywołać ustalenia S. Dorofeeva oraz P. Granta, którzy w odniesieniu do schematów doboru próby z operatów zespołowych stwierdzają:

> bardzo często najlepszym rozwiązaniem jest zaakceptowanie zróżnicowania prawdopodobieństw selekcji oraz uwzględnienie ich na etapie analizy danych, pod jednym wszak warunkiem, iż prawdopodobieństwa doboru są znane. (Dorofeev i in. 2007: 31)

Doskonałą egzemplifikacją konsekwencji wykorzystania prób adresowych w doborze jednostek z nierównymi prawdopodobieństwami selekcji są wyniki *Diagnozy Społecznej*. Warto zastrzec, że choć głównym przedmiotem tych badań nie są jednostki, ale gospodarstwa domowe, to jednak z uwagi na reprezentatywny charakter próby badawczej w odniesieniu do populacji gospodarstw domowych uzyskane rozkłady liczebności zespołów pozwalają uwidocznić, z jakim potencjalnym przyrostem wariancji miałoby się do czynienia, gdyby próba osób dobierana była z operatu adresowego, zamiast losowania jej z rejestru imiennego.

W kolejnych kolumnach tabeli IV.2. zamieszczone zostały rozkłady liczby osób w wieku 15 lat i więcej zamieszkujących w wylosowanych gospodarstwach domowych. W wierszu oznaczonym symbolem S^2 przedstawiono estymatory wariancji liczebności gospodarstw domowych, natomiast w wierszu ostatnim zawarto informacje o wartościach miernika $DEFF_p$ w kolejnych latach[154]. Wyniki analiz ukazują, iż przyrost wariancji wynikający z doboru zespołowego może przyjąć dość znaczną skalę. Wedle przeprowadzonych szacunków dwustopniowe losowanie jednostek populacji byłoby w Polsce o około

---

budynków mieszkalnych nie da się spełnić założeń schematu losowania zakładającego dobór tej samej liczby jednostek (większej niż jedna) w obrębie każdego zespołu.

[154] Przyjęto przy tym założenie, że rozkłady liczności gospodarstw domowych w próbie odpowiadają rozkładom liczności gospodarstw domowych w populacji, a także, iż gospodarstwa domowe dobierane są do próby z jednakowymi szansami doboru.

25 procent mniej efektywne od prostego doboru z operatu imiennego[155]. Oznaczałoby to konieczność znacznego zwiększenia liczebności próby oraz wiązałoby się ze wzrostem kosztów prowadzonych badań. Wyniki potwierdzają też, iż przyrost wariancji będący efektem ważenia rekompensującego nierówne szanse losowania z operatów zespołowych pozostaje wprost proporcjonalny do stopnia zróżnicowania liczebności zespołów. Mówiąc inaczej, im mniej są one jednorodne pod względem liczby członków, tym utrata precyzji estymacji jest większa.

Tabela IV.2. Oszacowanie przyrostu wariancji na skutek zróżnicowania szans selekcji w doborze dwustopniowym ($DEFF_p$) na podstawie danych z repozytorium badań „Diagnoza Społeczna 2000–2011"

Liczba osób w wieku 15+ w gospodarstwach domowych	**DS-2000**[i)]	**DS-2003**	**DS-2005**	**DS-2007**	**DS-2009**	**DS-2011**
	Rozkład procentowy (dane w %)					
1 osoba	20,1	18,7	18,1	17,6	18,1	17,1
2 osoby	50,0	45,9	43,0	40,3	39,0	35,8
3 osoby	16,7	18,2	18,8	19,0	20,0	19,3
4 osoby	9,2	11,4	12,9	14,2	13,9	16,0
5 osób	2,7	3,8	4,6	5,7	5,9	7,0
6 osób	0,9	1,3	1,6	2,0	2,0	3,0
7 osób	0,3	0,4	0,6	0,7	0,7	1,1
8 osób	0,1	0,2	0,2	0,3	0,2	0,5
9 osób	0,02	0,04	0,1	0,1	0,05	0,1
10 osób i więcej	0,02	0,04	0,1	0,1	0,02	0,1
$\boldsymbol{S^2}$	**1,17**	**1,41**	**1,62**	**1,78**	**1,70**	**2,09**
$\boldsymbol{DEFF_p}$	**1,22**	**1,24**	**1,25**	**1,26**	**1,25**	**1,27**

Źródło: opracowanie własne na podstawie *Diagnoza Społeczna: Zintegrowana Baza Danych* [data pobrania: 10.11.2012.]

[i)] Oznaczenie edycji badania Diagnozy Społecznej.

IV.2. Schematy doboru prób badawczych – komplikacje praktyczne

Praktyka badawcza pozostaje dużo bardziej złożona niż zaprezentowane zostało to w poprzednich częściach tego rozdziału. Wystarczy zauważyć, że przeprowadzone dotąd analizy czynników warunkujących efektywność trzech

[155] Powołując się na analizy dotyczące projektu ESS, można wskazać za ustaleniami P. Lynna i in. (2007: 117), że podobną skalę przyrostu wariancji na skutek losowania próby z operatu gospodarstw domowych uzyskano w Austrii ($DEFF_p$=1,25), Czechach (1,25), Hiszpanii (1,22) oraz Wielkiej Brytanii (1,21). Znacznie mniejszy efekt losowania zespołowego odnotowano natomiast w Irlandii (1,04), z kolei dużo wyższy w Portugalii (1,83).

głównych schematów doboru próby (uwarstwowienia, uzespołowienia oraz zróżnicowania szans selekcji) wykonane były w zasadzie we wzajemnej izolacji tych schematów od siebie, co było przydatne do określenia czynników warunkujących ich większą lub mniejszą efektywność, lecz nie odpowiadało prawie w ogóle rzeczywistości empirycznej. W praktyce badawczej na schematy doboru próby składają się bowiem wielostopniowe oraz złożone procedury, co przekłada się na konieczność zastosowania takich estymatorów parametrów populacyjnych oraz miar szacowania ich efektywności, które przyjmują często postać dość skomplikowanych formuł matematycznych[156]. Główną trudnością z tym związaną nie jest naturalnie ich złożoność obliczeniowa, ale potrzeba opracowania takich metod szacowania wariancji estymatorów, które odpowiadałyby wykorzystanym schematom doboru respondentów.

Rozwiązanie tego problemu nie jest wcale trywialne. Badacze stosują wprawdzie uproszczenie w sposobie estymacji efektywności zastosowanego przez nich schematu losowania próby – które polega na niezależnym od siebie szacowaniu przyrostu wariancji wynikającej z doboru warstwowego ($DEFF_s$), zespołowego ($DEFF_c$) i/lub z nierównymi prawdopodobieństwami selekcji jednostek ($DEFF_p$), a następnie na wyznaczeniu iloczynu tych pojedynczych miar, jako całkowitego efektu schematu doboru respondentów[157] i chociaż działanie takie jest praktyką stosowaną na szeroką skalę, choćby w badaniach ESS-u (por. Lynn i in. 2007: 114), to należy zauważyć, iż ma ono uzasadnienie tylko i wyłącznie wtedy, gdy: (1) wewnątrzwarstwowe wagi nie będą skorelowane z wartościami analizowanych zmiennych, (2) relatywne wariancje wag (tj. komponent miary *VIF*) pozostaną równe w każdej warstwie populacji, a także wtedy, gdy (3) we wszystkich warstwach populacji równe są współczynniki korelacji wewnątrzzespołowej. Uchylenie któregoś z tych warunków pociąga za sobą konieczność estymacji miernika przyrostu wariancji w inny sposób, o czym przypomniał ostatnio Hyunshik Lee (2012: 16–20).

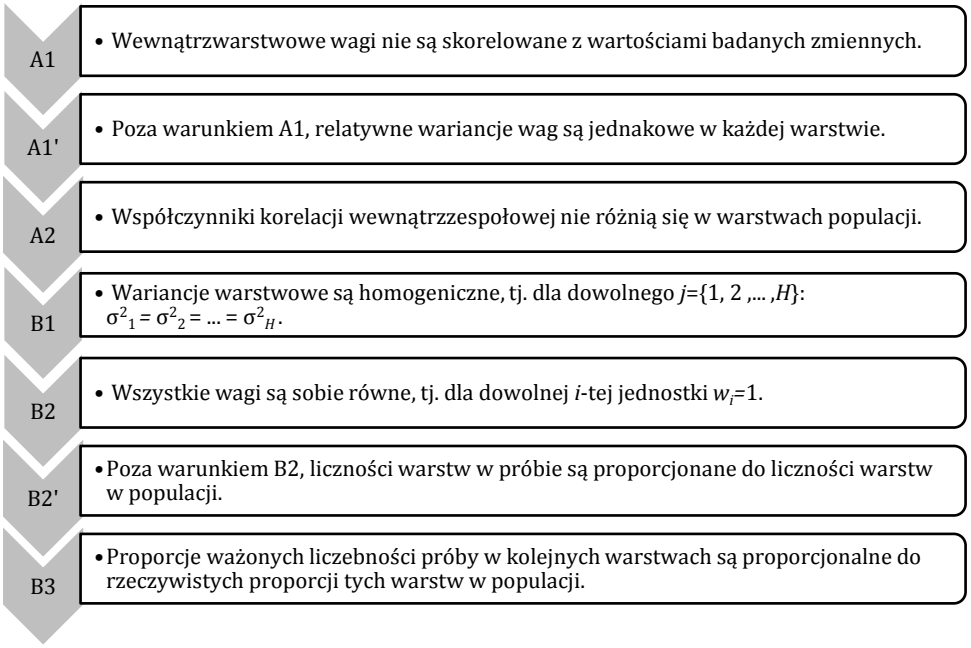
Zresztą poza wspomnianymi już kryteriami uprawomocniającymi możliwość estymacji $DEFF_{TOTAL}$ jako iloczynu efektu wynikającego z rozwarstwienia, uzespołowienia oraz ważenia danych, w literaturze metodologicznej pojawiają się również analizy innych kryteriów warunkujących określony sposób wyznaczania wartości miernika $DEFF_{TOTAL}$ (zob. ryc. IV.1.)[158]. Należy w tym względzie odwołać się do dwóch niezwykle pouczających artykułów, opublikowanych

---

[156] Świetnym przykładem takich estymatorów o złożonej postaci jest oszacowanie parametru średniej arytmetycznej w próbie zespołowej dobranej warstwowo ze zróżnicowanymi szansami wylosowania jednostek (por. wzór (7) w Lee (2012: 17)).

[157] Zresztą w rozdziale II pracy zdefiniowany został miernik całkowitego efektu schematu losowania próby jako iloczyn $DEFF_s \times DEFF_c \times DEFF_p$ (por. wzór II.7. w rozdziale II)

[158] Spełnienie warunków opisanych w punktach A1, A1' oraz A2 umożliwia estymację miernika $DEFF_{TOTAL}$ jako iloczynu trzech jego podstawowych komponentów, tj. $DEFF_s$, $DEFF_p$ oraz $DEFF_c$



Ryc. IV.1. Wybrane kryteria warunkujące sposób estymacji $DEFF_{TOTAL}$

Źródło: opracowanie własne

w czasopiśmie „Survey Methodology". W opracowaniu *Design Effects for the Weighted Mean and Total Estimators Under Complex Survey Sampling*, autorstwa Parka oraz Lee (2004), odnaleźć można ciekawe przykłady estymacji miary $DEFF_p$ dla przypadku skorelowania prawdopodobieństw selekcji (oraz w konsekwencji skorelowania zastosowanych wag) ze zmiennymi substantywnymi. Z kolei w artykule Gablera i in. (2006) *Design Effects for Multiple Design Samples* autorzy przyjmują założenie o homogeniczności wariancji warstwowych oraz o heterogeniczności współczynników korelacji wewnątrzzespołowej w warstwach populacji, a następnie, nakładając dodatkowe ograniczenia na właściwości wag korygujących szanse selekcji jednostek do próby badawczej, przedstawiają odmienne warianty estymacji mierników efektu schematu doboru próby. Jeden z takich estymatorów (por. wzór [7] w Gabler i in. 2006: 117) wy-

(por. Lee 2012: 19). Z kolei kryterium B1 – odnoszące się do homogeniczności warstwowych wariancji – uzasadnia estymację $DEFF_{TOTAL}$ z pominięciem efektu rozwarstwienia (por. Lee 2012: 19, Lynn i in. 2007: 114, Gabler i in. 2006: 115). Na konsekwencjach wynikających z zanegowania warunku opisanego w kryterium A1, czyli na takich schematach, w których prawdopodobieństwa selekcji pozostają skorelowane z analizowanymi zmiennymi, zwracają uwagę Park i in. (2004: 183-193), natomiast na różnych wariantach estymacji $DEFF_{TOTAL}$ wynikających z przyjęcia kryteriów B1, B2, B2' oraz B3 koncentrują uwagę Gabler i in. (2006: 116–117).

korzystany zostanie zresztą w dalszej części tego rozdziału w celu ukazania implikacji wynikających z przyjęcia różnych metod estymacji miary przyrostu wariancji w odniesieniu do schematu doboru próby w polskim komponencie piątej rundy Europejskiego Sondażu Społecznego.

Zanim jednak zaprezentowane zostaną analizy empiryczne, uszczegółowienia wymaga jeszcze jedna kwestia. Zwrócono już zresztą na nią uwagę w sekcji IV.1.2. tego rozdziału rozważając sposób definiowania miernika $DEFF_c$ przy pomocy współczynnika korelacji wewnątrzzespołowej (por. wzór IV.22.). W podanym wówczas wzorze pojawił się parametr *l'* odpowiadający liczności wylosowanych jednostek. W doborze wiązek o jednakowej liczbie elementów sprawa jest oczywista, jednak w przypadku ich nierówności pojawia się konieczność przyjęcia jakiegoś oszacowania wielkości parametru *l'*. Dla przykładu wskazywano już, że zgodnie z propozycją L. Kisha (1987) wykorzystać można średnią arytmetyczną liczebności zespołów[159]. Wówczas parametr *l'* daje się wyrazić jako:

(IV.28.) $$\bar{l} = \frac{1}{m}\sum_{j=1}^{m} l_j,$$

gdzie:

- *m* oznacza liczbę wylosowanych zespołów,
- natomiast $l_1, l_2, \ldots, l_m$ symbolizuje liczbę jednostek w każdym *j*-tym zespole.

Oszacowanie takie jest oczywiście proste i intuicyjne, chociaż okazuje się ryzykowne. W artykule pt. *A model based justification of Kish's formula for design effect for weighting and clustering*, opublikowanym w 1999 roku przez S. Gabler, S. Häder oraz P. Lahiri, wykazane zostało bowiem, że w schemacie losowania zespołowego z nierównymi szansami selekcji jednostek wykorzystanie średniej arytmetycznej liczebności zespołu w oszacowaniu miernika $DEFF_c$ może prowadzić do jego znacznego niedoszacowania. Autorzy tego artykułu wykazali, że bardziej dokładne przybliżenie $DEFF_c$ uzyska się przez zdefiniowanie parametru *l'* w postaci średnich liczebności ważonych (por. Gabler i in. 1999: 105)[160], tzn. jako:

---

[159] Zresztą właśnie taki sposób oszacowania parametru *l'*, tj. poprzez średnią arytmetyczną liczebność zespołów, przyjmowany jest w polskiej części projektu ESS (por. *Sampling design in* ESS5-PL). W odniesieniu do badań Europejskiego Sondażu Społecznego należy jednak zauważyć pewną nieścisłość w rekomendacjach dotyczących sposobów estymacji parametru *l'*. Wystarczy wskazać, że część dokumentów ESS-u zawiera sugestię, by szacunki $DEFF_c$ oprzeć na średnich liczebnościach zespołów, tj. na $\bar{l}$ (por. *Sampling for the European Social Survey Round V: Principles and Requirements* 2010: 13), inne natomiast, aby operować współczynnikiem l^* (por. Ganninger 2008: 4).

[160] Analizy podjęte przez Gablera i in. (1999) kontynuowane były w pracy Lynna i in. (2005: 101–104). W tym bardzo ciekawym opracowaniu skupiono się na wzajemnych relacjach pomiędzy parametrami $\bar{l}$, l^* oraz $\bar{l}_w$, a także na uwarunkowaniach ich wartości zarówno od liczebności zespo-

(IV.28'.) $$l^* = \frac{\sum_{j=1}^{m}\left(\sum_{i=1}^{l_j} w_{ji}\right)^2}{\sum_{j=1}^{m}\sum_{i=1}^{l_j} w_{ji}^2}$$

lub też – choć nie równoważnie – w postaci formuły:

(IV.28''.) $$\bar{l}_w = \frac{\sum_{j=1}^{m} l_j \sum_{i=1}^{l_j} w_{ji}^2}{\sum_{j=1}^{m}\sum_{i=1}^{l_j} w_{ji}^2},$$

gdzie dla każdego *j*-tego zespołu (lub *j*-tej wiązki respondentów) o liczebnościach równych $l_1, l_2, \ldots, l_m$, w_{ji} oznacza wagę[161] nadaną *i*-tej jednostce z *j*-tego zespołu. Owe alternatywne sposoby definiowania parametru *l'* wykorzystane będą w części analitycznej tego rozdziału do porównania metod estymacji mierników $DEFF_{TOTAL}$ na przykładzie danych z projektu ESS.

IV.2.1. Porównanie procedur szacowania wielkości $DEFF_{TOTAL}$ – analizy empiryczne na przykładzie danych ESS5-PL (ed. 2010)

Przedstawione w poprzednich częściach tego rozdziału rozważania metodologiczne zobrazowane zostaną kilkoma przykładami empirycznymi[162]. Analizy oparte będą na danych pochodzących z polskiej części piątej rundy Europejskiego Sondażu Społecznego z 2010 roku (ESS5-PL)[163] i chociaż trudno

---

łów, jak i przyjętych procedur ważenia danych. Na podstawie analiz tych autorów wykazane będzie zresztą dalej, że w odniesieniu do polskiej części ESS5-PL (ed. 2010) musi być tak, że $l^* > \bar{l}$ oraz $l^* = \bar{l}_w$.

[161] Oczywiście w zamian „surowej" wagi w_{ji} można wykorzystać wagi znormalizowane $\widetilde{w}_{ji}$.

[162] Fragmenty analiz zaprezentowanych w tej części rozdziału opublikowane zostały przez autora tej monografii w artykule pt. *How (Not) to Estimate the Design Effect of a Complex Sampling Scheme: A Case Study of the Polish Section of the European Social Survey, Round 5* (por. Jabkowski 2013: 55–77).

[163] W ramach tego projektu istnieje możliwość połączenia danych wynikowych z bazą dokumentującą działania podejmowane w trakcie losowania próby. Istotnie, w repozytorium *Europejskiego Sondażu Społecznego* udostępniany jest plik *ESS Sample Design Data File* (w skrócie ESS-SDDF). Począwszy od piątej rundy badań ESS-u dane dostępne są również dla polskiego komponentu badań. Wspomniany zbór zawiera między innymi jednostkowe informacje o prawdopodobieństwach selekcji w ramach każdego etapu losowania próby (pozwalających na wyznaczenie wag będących odwrotnościami owych szans doboru), numerach porządkowych wiązek respondentów (pozwalających na oszacowanie wielkości współczynników korelacji wewnątrzzespołowej oraz wyznaczenie liczebności wiązek), jak też kodach identyfikujących warstwy populacji. Pomimo tak bogatego zestawu informacji (niedostępnego w żadnym innym projekcie badawczym) zakres publikowanych danych okazał się jednak niewystarczający. Akurat w odniesieniu do analiz dedykowanych rozwarstwieniu populacji, dane zawarte w pliku SDDF nie dawały możliwości pełnej identyfikacji warstw, a jedynie tę ich część, w ramach której zastosowano losowanie zespołowe. Autor wyraża wdzięczność prof. Pawłowi B. Sztabińskiemu oraz dr. Dariuszowi Przybyszowi

będzie uciec od studiów dedykowanych pewnym specyficznym właściwościom schematu doboru próby wykorzystanemu w tych badaniach, to jednak przeprowadzone analizy pozwolą – mimo wszystko – na sformułowanie ważnych wniosków o charakterze ogólnym.

Zacząć wypada od krótkiej prezentacji schematu losowania próby badawczej w ESS5-PL (zresztą na niektóre z tych działań zwracano uwagę już wcześniej). Z kwestii najważniejszych trzeba przypomnieć, że w badaniach tych wykorzystuje się rozwarstwienie populacji. Populacja dzielona jest na warstwy wyodrębniane z uwagi na typ oraz wielkość miejscowości, w tym na: (1) wsie, (2) miasta do 9,9 tys. mieszkańców, (3) miasta od 10 do 19,9 tys. mieszkańców, (4) miasta od 20 do 49,9 tys. mieszkańców, (5) miasta od 50 do 99,9 tys. mieszkańców, (6) miasta od 100 do 199,9 tys. mieszkańców, (7) miasta od 200 do 499,9 tys. mieszkańców, (8) miasta od 500 do 999,9 tys. mieszkańców oraz (9) miasto st. Warszawa (pow. 1 mln osób). Ponieważ w liczbie jednostek dobieranych z każdej warstwy uwzględnia się przewidywany stopień realizowalności wywiadów (w tym między innymi odsetek błędów adresowych, nienawiązanych kontaktów, odmów itd., to rozkład warstw w próbie zrealizowanej odpowiada, w znacznym stopniu, rzeczywistemu rozkładowi warstw w populacji. Dość istotną właściwością schematu losowania w polskich badaniach ESS-u jest to, że w części warstw (a dokładniej pięciu z nich, tj. w miastach o liczebności co najmniej 50 tys. mieszkańców), dobór respondentów przeprowadzany jest w sposób jednostopniowy, z prawdopodobieństwem proporcjonalnym do wielkości populacji wszystkich 86 miast z tych warstw, natomiast w warstwie wiejskiej oraz trzech warstwach miejskich o liczbie mieszkańców nieprzekraczającej 49,9 tys. osób stosuje się już dobór zespołowy. W tej wiązkowanej części próby badawczej losowanie przeprowadza się w sposób trójstopniowy, przy czym jednostkami doboru pierwszego stopnia, w ramach każdej warstwy, są gminy, a w ich obrębie – miejscowości (drugi stopień doboru), dobierane ze zwracaniem oraz z prawdopodobieństwem proporcjonalnym do wielkości populacji tych miejscowości, a następnie (trzeci etap doboru) w obrębie wylosowanych już wsi oraz miast dobierane są w sposób prosty wiązki 4 respondentów, przy czym, jeśli daną miejscowość wylosowano wielokrotnie, to zwiększa się też liczba dobieranych w niej zespołów. W ramach działań związanych z doborem próby określane są również prawdopodobieństwa selekcji jednostek z populacji do próby badawczej oraz wagi odpowiadające (znormalizowanym do wielkości próby[164]) odwrotnościom tych prawdopodobieństw[165] (por.

---

z Instytutu Filozofii i Socjologii PAN w Warszawie za udostępnienie pełnego zakresu informacji, bez których niemożliwe byłoby przeprowadzenie zaprezentowanych tu analiz.

[164] Por. wzór IV.26'. z części IV.1.3. tego rozdziału.

[165] Konstrukcja wag przebiega w następujący sposób. W pierwszej kolejności ustalane są liczebności osób, które mają być wylosowane z każdej z dziewięciu warstw populacji, przy czym są

Sampling Design in ESS5-PL 2010: 2–3). Podsumowując, należy powiedzieć, że schemat losowania próby w ESS5-PL przyjmuje postać doboru wielostopniowego, z losowaniem jednostek prowadzonym z populacji rozwarstwionej, a także z częściowym uzespołowieniem próby badawczej. Innymi słowy, w szacowaniu efektywności tego schematu należy uwzględnić konsekwencje rozwarstwienia, uzespołowienia (w część warstw) oraz losowania z nierównymi prawdopodobieństwami selekcji.

Wykorzystanie danych ESS5-PL tylko i wyłącznie w celu prezentacji technicznej strony estymacji mierników efektu schematu losowania próby badawczej pozbawione byłoby jednak większego sensu. Studia takie miałyby charakter czysto warsztatowy i siłą rzeczy same w sobie nie wniosłyby wiele do prowadzonej dyskusji. Szczególnie interesująca w kontekście analizy danych ESS-*u* wydaje się jednak możliwość ich wykorzystania w celu znalezienia odpowiedzi na kilka ważnych pytań metodologicznych. Można je podzielić na trzy grupy tematyczne dotyczące: (1) empirycznej oceny poziomu wewnątrzzespołowej homogenizacji jednostek, (2) weryfikacji kryteriów warunkujących sposób estymacji mierników $DEFF_{TOTAL}$ w odniesieniu do ESS5-PL (por. kryteria opisane na Ryc. IV.1.), a także (3) porównania różnych wariantów szacowania mierników $DEFF_{TOTAL}$. Część z tych problemów ma charakter ogólny, inne dotykają konkretnych zagadnień powiązanych ze schematem losowania próby w polskiej części badań piątej rundy Europejskiego Sondażu Społecznego. W ramach pierwszego ze sformułowanych tu celów znaleziona będzie odpowiedzieć na dwa pytania:

> P1: Czy zawarta w dokumentacji metodologicznej ESS-u rekomendacja, aby w ocenie efektywności schematu doboru wiązek respondentów przyjmować oszacowanie a priori współczynnika korelacji wewnątrzzespołowej na poziomie = 0,02 (por. Lynn i in. 2007: 114) odpowiada rzeczywistości empirycznej, tj. wielkościom tego współczynnika w próbie?

---

one (a) (prawie) proporcjonalne do wielkości populacji 15+ w danej warstwie (bowiem relatywnie więcej osób losowanych jest w tych warstwach, w których próba jest wiązkowana) oraz (b) odwrotnie proporcjonalne do przewidywanego poziomu *response rate* w tych warstwach. Prawdopodobieństwa losowania ustala się dla każdego z trzech etapów doboru próby. W ramach pierwszego z tych etapów prawdopodobieństwo doboru jest szansą wylosowania określonej gminy z danej warstwy. Mówiąc dokładniej, dla miast pow. 50 tys. mieszkańców prawdopodobieństwo to wynosi 1 (wszystkie miasta są uwzględnione), natomiast dla gmin miejskich oraz wiejskich prawdopodobieństwo wyznacza się jako iloraz liczby osób w danej gminie z określonej warstwy do liczby osób w tej warstwie. Na drugim etapie wyznacza się prawdopodobieństwo wylosowania wsi z danej gminy wiejskiej (liczba osób zamieszkujących w danej wsi do liczby osób w danej gminie wylosowanej na pierwszym etapie). Dla miast prawdopodobieństwo to wynosi 1. Na trzecim etapie prawdopodobieństwo doboru wyznacza się z kolei jako iloraz liczby osób 15+ wybranych z wylosowanych miejscowości do wielkości populacji 15+ tych miejscowości. Całościowe prawdopodobieństwa doboru wykorzystywane do wyznaczenia wag w zbiorze danych są iloczynem szans selekcji odpowiadającej każdemu z trzech etapów losowania próby.

P2: Czy w zbiorze pytań zawartych w kwestionariuszu ESS5 da się wyodrębnić jakieś moduły lub rodzaje pytań, dla których poziom korelacji wewnątrzzespołowej jest mniejszy lub większy? Od czego zależą takie różnice oraz czy da się w tym zakresie wskazać na jakieś prawidłowości?

Drugi cel oraz pytanie z tym związane dotyczy kryteriów warunkujących sposób estymacji mierników $DEFF_{TOTAL}$ w badaniach ESS5-PL:

P3: Które z podanych kryteriów (odnoszących się do rozwarstwienia, uzespołowienia oraz ważenia danych) znajdują potwierdzenie w ramach ESS5-PL?

Ustalenia związane z odpowiedzią na pytanie 3 wykorzystane będą do porównania różnych wariantów szacowania mierników $DEFF_{TOTAL}$. Warto przy tym poszukać odpowiedzi na następujące pytanie:

P4: Czy zastosowanie odmiennych (często alternatywnych) procedur estymacji efektu schematu doboru próby prowadzi do zasadniczych różnic w uzyskiwanych wielkościach tych miar? Jeśli tak, to jakie czynniki decydują o odmienności lub zbieżności wielkości $DEFF_{TOTAL}$ wyznaczanych w oparciu o różne wersje estymacji tego współczynnika?

Sformułowanie tego pytania wymaga doprecyzowania – wydaje się zresztą, że dotyka problemu niezwykle istotnego dla działań mających na celu ocenę efektywności doboru próby. Wspominano wcześniej, że podstawową trudnością wynikającą z wykorzystania złożonych schematów losowania respondentów (takich jak w ESS5-PL) okazuje się konieczność estymacji $DEFF_{TOTAL}$ w oparciu o taki miernik, który byłby zgodny ze schematem losowania oraz z kryteriami, jakie ów schemat spełnia. Wskazywano też, że w wielu przypadkach wykorzystuje się mierniki uproszczone, to znaczy takie, które oparte są na idealizacyjnych założeniach o właściwościach próby oraz sposobach jej doboru. Przeformułowując nieco pytanie czwarte można doprecyzować zakres analiz, stawiając pytanie w zmienionej formie:

P4': Czy wykorzystanie procedur estymacji $DEFF_{TOTAL}$, które nie mają uprawomocnienia metodologicznego w kryteriach warunkujących metodę oceny efektywności schematu doboru próby, prowadzi do istotnego przeszacowania (lub też niedoszacowania) rzeczywistych wielkości $DEFF_{TOTAL}$?

W poniższej tabeli zestawiono oraz scharakteryzowano porównywane warianty estymacji miernika $DEFF_{TOTAL}$ (wraz z kryteriami uprawomocniającymi ich zastosowanie). Będą one przedmiotem analiz w odniesieniu do polskiego zbioru danych z projektu ESS5–2010.

Tabela IV.3. Specyfikacja porównywanych wariantów estymacji mierników DEFF_{TOTAL} w ESS5-PL (ed. 2010)

Wariant estymacji miernika DEFF_{TOTAL}	**Kryteria warunkujące określony wariant estymacji DEFF_{TOTAL}** (por. Ryc. IV.1.).
1	2
<u>*Wariant I*</u>: Procedura estymacji DEFF_{TOTAL} wykorzystywana pierwotnie do szacowania efektu schematu losowania próby w ESS5-PL. Ogólna postać estymatora: $\text{DEFF}_{TOTAL} = \text{DEFF}_p \cdot \sum_{j=1}^{H} h'_j \text{DEFF}_c^j$, gdzie w każdej warstwie h'_j oznacza nieważoną proporcję jednostek znajdujących się w *j*-tej warstwie. Ponieważ w ESS5-PL stosuje się losowanie warstwowe z podziałem próby na dwie części: I. wiązkowaną (1 warstwa wiejska oraz 3 warstwy miejskie – do 49,9 tys. mieszkańców), z założeniem równości współczynnika ρ w warstwach. Wówczas: $\text{DEFF}_c^{\text{I}} = 1 + (\bar{l}' - 1)\rho$, gdzie $h'_{\text{I}} = \sum_{j=1}^{4} h'_j$ określa sumę proporcji kolejnych warstw z wiązkowanej części próby, parametr $\bar{l}' = \bar{l}$; II. niewiązkowaną (5 warstw miejskich zawierających miasta o populacji 50 tys. mieszkańców i więcej, rozpatrywanych łącznie w estymacji DEFF). Wówczas: $\text{DEFF}_c^{\text{II}} = 1$; $h'_{\text{II}} = \sum_{j=5}^{9} h'_j$ – określa sumę proporcji kolejnych warstw z niewiązkowanej części próby, to efekt uzespołowienia dotyczy tylko części próby. Estymator wykorzystany do oszacowania tego efektu przyjmuje następującą postać: $\text{DEFF}_c = h'_{\text{I}} \text{DEFF}_c^{\text{I}} + h'_{\text{II}}$. Z kolei $\text{DEFF}_{TOTAL} = \text{DEFF}_p \cdot \text{DEFF}_c$ (por. Lynn i in. 2007: 114 oraz *Sampling design in ESS5-PL* 2010: 2).	A1 A1' A2 B1 B2 B2'
<u>*Wariant II*</u>: Uwzględniona nieproporcjonalna ważona liczebność warstw w próbie (w stosunku do liczności warstw w populacji), przy założeniu równości współczynnika ρ we wszystkich warstwach wiązkowanej części próby. Ogólna postać estymatora: $\text{DEFF}_{TOTAL} = \sum_{j=1}^{H} \frac{\hat{h}_j^2}{h'_j} \text{DEFF}_p^j \text{DEFF}_c^j$ (por. wzór [8] w Lee 2012: 18 oraz wzór [5] w Gabler i in 2006: 116), gdzie $\hat{h}_j = \sum_{i=1}^{n_j} w_{ji} / \sum_{i=1}^{n} w_i$ jest oszacowaniem frakcji każdej *j*-tej warstwy w populacji ustalonym w oparciu o wagi zdefiniowane jako odwrotności szans losowania jednostek (por. Lee 2012: 18). Wariant ten różni się od pierwszego w dwóch kwestiach. Po pierwsze, DEFF_p estymowane jest niezależnie w każdej warstwie i mnożone przez DEFF_c zanim wyznaczona zostanie warstwowa wielkość DEFF_{TOTAL}. Po drugie, warstwowe wagi wyznaczane są jako proporcja wag będących odwrotnościami prawdopodobieństw selekcji oraz frakcji respondentów w warstwie. Co więcej, oszacowanie to nie bierze pod uwagę tej części DEFF_p, która wynika ze zróżnicowań prawdopodobieństw selekcji pomiędzy warstwami. W konsekwencji DEFF_{TOTAL} może być niedoszacowane. Warto jednak wskazać, że taka formuła estymacji DEFF_{TOTAL} sugerowana jest w literaturze metodologicznej jako właściwy sposób estymacji tej wielkości dla schematu losowania w którym wewnątrzwarstwowe wagi są równe (por. scenariusz 2 w Gabler i in. 2006: 116). Dla danych z ESS5-PL otrzymuje się podział próby na dwie części: wiązkowaną – 4 warstwy rozpatrywane łącznie ($\text{DEFF}_c^{\text{I}} = 1 + (\bar{l}' - 1)\rho$), parametr $\bar{l}' = l^*$, niewiązkowaną – 5 warstw rozpatrywanych łącznie, dla których efekt wiązkowania $\text{DEFF}_c^{\text{II}} = 1$.	A1 A2 B1 ~B2 ~B3

cd. tab. IV.3.

1	2
Wariant III: Uwzględniona nieproporcjonalna ważona liczebność warstw w próbie (w stosunku do liczności warstw w populacji), przy założeniu zróżnicowania współczynników ρ w warstwach wiązkowanej części próby. Ogólna postać estymatora tak jak w II (różni się od II wyłącznie założeniem o odmienności współczynników korelacji wewnątrzzespołowej w kolejnych warstwach). Dla danych z ESS5-PL podział próby na pięć części, tzn. cztery odpowiadające warstwom populacji z wiązkowaniem respondentów w próbie (rozpatrywane oddzielnie, tj. dla każdej wyznacza się $\mathrm{DEFF}_c^j = 1 + (l_j^{'} - 1)\rho_j$, w każdej warstwie parametr $l^{'} = l^{*}$, oraz część niewiązkowana (5 warstw rozpatrywana łącznie dla których $\mathrm{DEFF}_c^{\mathrm{II}} = 1$).	A1 ~A2 B1 ~B2 ~B3

Źródło: opracowanie własne

Pierwszy z tych wariantów – odpowiadający metodzie wykorzystywanej w szacowaniu efektu schematu doboru próby w polskiej części projektu ESS – przyjmuje postać stosunkowo prostej formuły obliczeniowej. Możliwość jej zastosowania pozostaje jednak obłożona dość silnymi obwarowaniami metodologicznymi. Analizy empiryczne rozpocząć należy zatem od odpowiedzi na pytanie trzecie, a więc od oceny kryteriów warunkujących sposób estymacji miary DEFF_{TOTAL}. W pierwszej kolejności warto zauważyć, że w procedurze opisanej w wariancie I pomija się efekt rozwarstwienia, co oznacza, iż działanie takie będzie właściwe, o ile tylko spełnione pozostanie kryterium B1, czyli warunek homogeniczności wariancji warstwowych[166]. W przeciwnym razie pominięcie efektu losowania warstwowego prowadzić może do niedoszacowania DEFF_{TOTAL} (w takiej sytuacji rzeczywista efektywność schematu doboru próby będzie większa, niż wskazuje na to wyznaczona wielkość tego współczynnika) lub też skutkować może przeszacowaniem tej miary (czego konsekwencją będzie wylosowanie próby zbyt licznej w stosunku do rzeczywistej efektywności schematu doboru respondentów)[167]. Po drugie, ponieważ miarę DEFF_{TOTAL} wyznacza się

[166] W tabeli A3. zamieszczonej w aneksie (w kolumnie oznaczonej symbolem B1), podane zostały wyniki testu Levene'a, który wykorzystany został do zweryfikowania hipotezy o równości wariancji warstwowych dla kolejnych zmiennych zwartych w kwestionariuszu piątej rundy ESS. Z danych tych można odczytać, że kryterium homogeniczności wariancji należy odrzucić w przypadku 33 spośród 90 analizowanych tutaj pytań.

[167] W badaniach piątej rundy projektu ESS zrealizowanych w Polsce ma się do czynienia z sytuacją, w której pominięcie efektu losowania próby z populacji rozwarstwionej prowadzi częściej do przeszacowania (niż do niedoszacowania) miary DEFF_{TOTAL}. Choć oczywiście DEFF_s jest charakterystyką przypisywaną konkretnemu pytaniu, to uśredniona wielkość tej miary (będąca wypadkową ze wszystkich pojedynczych mierników DEFF_s) ukazuje, że rozwarstwienie ma w ESS5-PL nieznacznie większą efektywność niż dobór prowadzony z całej populacji (przeciętna wielkość DEFF_s wyniosła bowiem 0,997). Innymi słowy, pominiecie efektu losowania warstwowego w ocenie efektywności schematu doboru próby prowadzi rzeczywiście do nieznacznego przesza-

(w wariancie wykorzystywanym przez polskich koordynatorów badań), jako iloczyn efektu uzespołowienia oraz ważenia danych, to wymaga się tu spełnienia kryteriów A1, A1' oraz A2[168]. Po trzecie wreszcie, procedura ta w zasadzie wymaga, aby w całym zbiorze danych wagi były sobie równe (kryterium B2), a losowanie warstwowe prowadzone było w sposób proporcjonalny (warunek B2')[169].

Wyprzedzając w tym momencie dokładne analizy empiryczne danych ESS5-PL, można już wskazać, że w ramach badań realizowanych w Polsce nie są spełnione (z całą pewnością) dwa fundamentalne kryteria warunkujące wykorzystanie tej wersji estymacji miary $DEFF_{TOTAL}$, tzn. warunki B2 oraz B2'. Podstawowe zastrzeżenie budzi tutaj sposób ustalania wielkości $DEFF_c$, a dokładniej wykorzystanie w tym estymatorze parametru h'_j, czyli zwykłej proporcji

---

cowania miary $DEFF_{TOTAL}$. Przyglądając się dokładnym wielkościom $DEFF_s$ dla każdego z 90 rozpatrywanych tu pytań (por. tabela aneksowa A4.), można zauważyć, że zróżnicowanie wartości mierników $DEFF_s$ jest niewielkie (dla większości pytań poziom efektywności losowania warstwowego oscyluje wokół wielkości 1). Rozwarstwienie okazało się najbardziej korzystne w odniesieniu do pomiaru dochodów respondentów, tj. w pytaniu F41 ($DEFF_s$=0,956), z kolei najmniej korzystne było dla pomiaru zmiennej odpowiadającej ocenie szybkości działania policji, tzn. dla pytania D14 ($DEFF_s$=1,031). W sumie dla 45 pytań, tj. połowy ze wszystkich rozpatrywanych zmiennych, wartość $DEFF_s$ była mniejsza od 1, dla 7 była równa jedności, natomiast dla 38 większa od jeden.

[168] Zarówno kryterium A1 (mówiące o tym, iż wewnątrzwarstwowe wagi nie są skorelowane z wartościami analizowanych zmiennych), jak i A2 (mówiące, iż relatywne wariancje wag są jednakowe w każdej warstwie populacji) są spełnione w schemacie losowania próby w ESS5-PL. Wynika to ze sposobu wyznaczania wielkości wag, które są odwrotnościami prawdopodobieństw selekcji osób z populacji do próby badawczej. Konstrukcja wag jest bowiem taka, że praktycznie w każdej warstwie wszystkie jednostki z takiej warstwy mają przypisaną taką samą wagę. Nieco bardziej problematyczne okazuje się kryterium A2, mówiące o równości współczynników korelacji wewnątrzzespołowej w kolejnych warstwach populacji. Analizy empiryczne danych ESS5-PL (por. tabela A3.) wskazują na międzywarstwowe zróżnicowanie tych mierników. Istotnie można zauważyć, że uśredniona wielkość współczynnika ρ osiągnęła w całej próbie (a w zasadzie w tych warstwach, w których dobór był zespołowy, tzn. na wsiach oraz w miastach do 50 tys. mieszkańców) wartość równą 0,15, natomiast w kolejnych warstwach z wiązkowanej części próby badawczej wartości współczynników ρ osiągnęły poziom równy: (a) 0,168 dla warstwy wiejskiej, (b) 0,161 dla miast do 9,9 tys. mieszkańców, (c) 0,133 dla miast od 10 do 19,9 tys. mieszkańców, a także (d) 0,093 dla miast od 20 do 49,9 tys. mieszkańców. Widać zatem, że w ostatniej z tych warstw wielkość ρ odbiega znaczniej od wartości współczynników wewnątrzzespołowej korelacji jednostek w pozostałych warstwach badanej populacji. Choć różnice nie są być może znaczące, to nadal estymacja efektywności schematu doboru uwzględniająca, lub też pomijająca, wewnątrzwarstwowe zróżnicowania ρ prowadzić może do znaczących różnic w wartościach miary $DEFF_{TOTAL}$. Wielkość ta nie jest bowiem uwarunkowana wyłącznie poziomem wewnątrzwiązkowej homogenizacji jednostek, ale również liczebnością wiązek, frakcją jednostek próby z danej warstwy populacji, czy też wagami przypisanymi kolejnym respondentom. Problem ten podjęty zostanie w analizach dedykowanych empirycznej ocenie efektywności doboru próby w ESS5-PL, w ramach porównań II oraz III wariantu estymacji $DEFF_{TOTAL}$ (por. tabela IV.3.).

[169] Należy zresztą wskazać, że ów sposób estymacji miernika $DEFF_{TOTAL}$ wraz z podanymi tu obwarowaniami omówiony został przez Gablera i in. (2006: 116) (por. wzór [6] oraz [9] w tym artykule).

jednostek próby należących do *j*-tej warstwy populacji. Ponieważ w ESS5-PL zbiór danych jest ważony (odrzucić trzeba kryterium B2 oraz B2'), a ważone liczebności warstw w próbie nie są proporcjonalne do liczebności warstw w populacji (zanegowanie warunku B3), to w szacowaniu $DEFF_{TOTAL}$ należałoby, zamiast parametru h'_j, wykorzystać współczynnik uwzględniający ważoną nieproporcjonalną strukturę warstw w próbie (w opracowaniu Gabler i in. 2006: 116–117 odnaleźć można wyrażoną wprost sugestię, by efekt schematu doboru w polskiej części ESS estymować w oparciu o procedurę zgodną z wariantem II lub III). Innymi słowy, metoda szacowania efektywności schematu doboru próby w wersji wykorzystanej przez zespół polskich koordynatorów projektu ESS nie jest dobrana w sposób odpowiadający w pełni zastosowanemu schematowi losowania respondentów, co każe postawić pytanie, czy ta nieprawidłowa procedura estymacji miary $DEFF_{TOTAL}$ prowadzi do znaczącego niedoszacowania czy też przeszacowania rzeczywistej wielkości efektu schematu doboru próby.

Procedura opisana w wariancie pierwszym zestawiona będzie z takimi dwoma metodami estymacji miary $DEFF_{TOTAL}$, które są bardziej właściwe dla schematu losowania zastosowanego w ESS5-PL (por. wzór [8] w Lee 2012: 18; wzór [5] w Gabler i in 2006: 116). Przy podtrzymaniu założenia o homogeniczności wariancji warstwowych (warunek B1)[170] oraz o braku zależności pomiędzy wagami i wartościami analizowanych zmiennych (kryterium A1)[171], a także zanegowaniu założeń B2, B2' oraz B3, porównane zostaną dwie procedury estymacji efektu schematu losowania próby w ESS5-PL, różniące się założeniem o wartościach międzywarstwowych współczynników korelacji wewnątrzzespołowej. W wariancie II przyjęto bowiem równość takich współczynników w kolejnych warstwach populacji z wiązkowanej części próby (kryterium A2), natomiast w wariancie III – założenie o międzywarstwowym zróżnicowaniu wartości ρ (zanegowanie warunku A2). W dalszej części analiz wykazane zostanie, że obie te metody są uprawomocnione w odniesieniu do badań polskiej części piątej rundy ESS. Zestawienie ich ze sobą ma służyć ocenie tego, czy uwzględnienie warstwowego zróżnicowania ρ wpływa w jakiś znaczący sposób na estymację wielkości $DEFF_c$ oraz $DEFF_{TOTAL}$.

Zanim przedstawione zostaną wyniki analiz, konieczne jest zaprezentowanie zestawu zmiennych oraz opisanie kluczowych charakterystyk zbioru danych. Rozpocząć warto od tego, że z ponad 200 pytań zawartych w kwestionariuszu ESS5 do analiz wybrano jedynie 90 itemów spełniających dwa kryte-

---

[170] Oznaczający możliwość pominięcia efektu rozwarstwienia.

[171] Jest to założenie konserwatywne. Prowadzi raczej do przeszacowania, niż niedoszacowania wielkości miary $DEFF_p$. W rzeczywistości skorelowanie wag z badanymi zmiennymi jest korzystne i może zwiększyć poziom precyzji estymacji (por. Little i in. 2005).

Tabela IV.4. Charakterystyki zbioru danych ESS5-PL (ed. 2010) oraz opis parametrów wykorzystanych w szacowaniu mierników DEFF

Parametr	Warstwy populacji									Ogółem
	Wieś	Miasto do 10 tys.	Miasto 10–19 tys.	Miasto 20–49 tys.	Miasto 50–99 tys.	Miasto 100–199 tys.	Miasto 200–499 tys.	Miasto 500–999 tys.	Warszawa	
Frakcja warstw w populacji: h_j	37,94	6,02	6,99	10,99	8,50	8,14	9,47	7,31	4,64	100,0
		61,94					38,06			
Estymator frakcji warstw w populacji (liczebności ważone): $\hat{h}_j$	41,58	6,62	7,22	10,82	7,51	7,43	8,61	6,04	4,17	100,0
		66,24					33,76			
Frakcja warstw w próbie: h'_j	36,61	6,68	7,34	10,91	7,59	8,74	10,11	7,08	4,94	100,0
		61,54					38,46			
Liczebność próby: n_j	641	117	129	191	133	153	177	124	86	1751
		1078					673			
Liczba wiązek w próbie: l_j	219	38	38	58	–	–	–	–	–	353
Średnia liczebność wiązki: $\bar{l}_j$	2,927	3,079	3,395	3,293	–	–	–	–	–	3,054
Średnie ważone liczebności wiązek: $l^*_j = \bar{l}_{w_j}$	3,234	3,547	4,473	4,047	–	–	–	–	–	3,507
Współczynnik korelacji wewnątrzzespołowej: $\hat{\rho}_j^{ANOVA}$	0,168	0,162	0,133	0,093	–	–	–	–	–	0,150
Wskaźnik przyrostu wariancji (wagi jako szanse losowania): $DEFF_{pj}=VIF_j$	1,000	1,000	1,000	1,003	1,001	1,000	1,000	1,000	1,000	1,014
		1,005					1,005			

Źródło: Obliczenia własne na podstawie repozytorium danych ESS5-PL oraz ESS5-SDDF-PL

ria[172]. Po pierwsze, uwzględniono wyłącznie te pytania, które zadawane były wszystkim respondentom. Po drugie, zestaw pytań zawężono do takich zmiennych, których poziom pomiaru pozwalał na scharakteryzowanie ich wartości za

[172] Wybrane do analiz pytania obejmują znaczną część tematyki poruszanej w ESS5. Spośród modułów stałych, tzn. powtarzanych we wszystkich kolejnych rundach ESS, wybrano w module A – 6 pytań, w B – 22 pytania, w C – 5 pytań, w F – 2 pytania. Z kolei w modułach rotacyjnych, tzn. w D, G oraz H, w analizach uwzględniono odpowiednio 31, 3 oraz 21 pytań. Wykluczono jedynie wszystkie pytania z modułu I (zestaw pytań weryfikujących zgodność odpowiedzi udzielanych we wcześniejszych pytaniach) oraz z modułu J, zawierającego pytania skierowane do ankieterów (opis przebiegu wywiadu). Wykaz zmiennych, wraz z ich charakterystyką, zawarty został w tabeli aneksowej A2.

pomocą średniej arytmetycznej. Sformułowanie tego drugiego warunku było konieczne z uwagi na sposób wyznaczania wartości współczynników korelacji wewnątrzzespołowej prowadzonej w oparciu o procedurę analizy wariancji.

W tabeli IV.4. zestawione zostały najważniejsze charakterystyki próby badawczej z polskiej części piątej rundy projektu ESS. Z kwestii zupełnie podstawowych należy ponownie przypomnieć, że wiązkowanie próby badawczej dotyczyło tylko czterech (z dziewięciu) warstw populacji (tj. wsi oraz miast o liczebności populacji do 49,9 tys. mieszkańców), do których należało łącznie nieco ponad 61,5% wszystkich respondentów (co stanowiło 1078 osób) przydzielonych do 353 różnych zespołów (tj. wiązek respondentów). Pozostałe 673 osoby dobierane były już schematem losowania indywidualnego, dla którego $DEFF_c$=1.

Do ciekawych wniosków prowadzi analiza wartości parametrów charakteryzujących wielkość wiązek respondentów. W zbiorze danych ESS5-PL średnia arytmetyczna liczebność wiązki wyniosła 3,054 osoby[173], natomiast przeciętna liczebność ważona (wyrażona parametrami l^* i $\bar{l}_w$) przyjęła wartość równą 3,507 osoby[174]. Innymi słowy, ponieważ $l^* = \bar{l}_w > \bar{l}$, to $DEFF_c$ wyznaczane w oparciu o parametr średniej arytmetycznej (czyli $\bar{l}$) – tak jak czyni się w polskim komponencie tego projektu – będzie mniejsza niż wartość tej miary dla liczebności ważonych (l^* lub $\bar{l}_w$). Choć oczywiście nie zawsze musi być tak, że $l^* > \bar{l}$ oraz $l^* = \bar{l}_w$, to jednak w odniesieniu do ESS5-PL owe zależności nie są przypadkowe, a dokładniej – wynikają wprost ze schematu doboru próby. Relacje pomiędzy tymi parametrami przeanalizowane zostały w artykule P. Lynna i in. (2005: 101–104). Choć autorzy przywołanego opracowania rozważali wyniki badań pierwszej rundy ESS z 2002 roku, to jednak ich ustalenia odnieść można również do piątej odsłony tego projektu (co do zasady, w Polsce nie zmienił się schemat doboru próby). W części empirycznej tego artykułu odnaleźć można konstatację wskazującą na to, iż:

> w pewnych krajach [tu wymieniana jest także Polska – P.J.] jednostki populacji losowane były (wewnątrz wiązek) z jednakowym prawdopodobieństwem doboru. W takich przypadkach spełniony jest warunek [równości wag przypisanym jednostkom w obrębie tych samych zespołów- P.J.]. (Lynn i in. 2005: 104)

Ze stwierdzenia faktu o jednakowych wartościach wag w każdej wiązce respondentów wynika już wprost, że l^* musi być równe $\bar{l}_w$, nadal jednak nie tłumaczy to nierówności $l^* > \bar{l}$. Doprecyzowanie tej drugiej kwestii odnaleźć

---

[173] W ramach poszczególnych pytań w ESS5-PL liczebności wiązek mogły różnić się między sobą, co wynikało przede wszystkim z odmów odpowiedzi na konkretne pytania. Zróżnicowania te pominięto w dalszych analizach przyjmując wartości przeciętne dla całej próby.

[174] W ESS5-PL nie ma zatem znaczenia, czy będzie to l^*, czy też $\bar{l}_w$, bowiem oba te parametry przyjmują jednakowe wartości.

można jednak we wcześniejszej części przywołanego artykułu. Jego autorzy wskazują bowiem, że jeżeli w obrębie każdego zespołu wagi są jednakowe, a liczebności zespołów różne (tak jak w polskiej próbie ESS), to l^* będzie większe od $\bar{l}$, o ile tylko kowariancja pomiędzy liczebnością zespołów oraz ich liczebnością ważoną będzie większa lub równa zero (por. Lynn i in. 2005 102)[175]. Proste operacje arytmetyczne przeprowadzone na danych z ESS5-PL pozwalają wykazać spełnienie tego warunku[176], stąd też $l^* > \bar{l}$. Podsumowując, należy wskazać, że w ramach polskiego zbioru danych wynikowych wykorzystanie parametru l^* w miejsce $\bar{l}$ ukazuje potrzebę losowania próby o liczności większej niż wynika to z obliczeń opartych na zwykłej średniej arytmetycznej liczebności wiązek[177]. Pociąga to za sobą wzrost kosztów badawczych, jednak dla schematów losowania z nierównymi prawdopodobieństwami doboru działanie takie jest metodologicznie bardziej uzasadnione (por. Gabler i in. 1999: 105–106).

Przyglądając się na koniec wartościom mierników *VIF* (przyrost wariancji na skutek ważenia danych), zauważyć można jeszcze jedną ciekawą prawidłowość schematu doboru próby ESS w Polsce. Otóż wartości wag pozostają jednakowe nie tylko w obrębie wiązek respondentów, ale również w obrębie całych warstw populacji (warstwowe wartości *VIF* są równe 1)[178]. Jest to o tyle znaczące, że w ESS5-PL ważenie danych nie powoduje (prawie w ogóle) przyrostu wariancji w obrębie pomiaru wewnątrz warstw populacji, skutkuje jednak utratą precyzji estymacji w całej próbie badawczej. Oznacza to również spełnienie kryterium A1', czyli warunku mówiącego o równości relatywnych wariancji wag w każdej warstwie badanej populacji.

Wewnątrzzespołowa homogenizacja jednostek w ESS5-PL

Analizy poświęcone poziomowi wewnątrzzespołowej homogenizacji jednostek w polskiej części *Europejskiego Sondażu Społecznego* warto rozpocząć od wskazania tego, że w dokumentacji piątej rundy projektu ESS (por. *Sampling for the ESS Round V* 2010)[179], jak również w omówieniach wcześniejszych odsłon

---

[175] P. Lynn i in. podają wzór służący wyznaczeniu wielkości owej kowariancji (por. Lynn i in. 2005: 101). Dopasowując ich propozycje do stosowanych w tej pracy oznaczeń, można zapisać ów wskaźnik w postaci formuły: $\mathrm{Cov}\left(l_j, l_j \overline{w}_j^2\right) = \frac{1}{m}\sum_{j=1}^{m} l_j^2 \overline{w}_j^2 - \frac{1}{m}\bar{l}\sum_{j=1}^{m} l_j \overline{w}_j^2$, gdzie dla każdej *j*-tej wiązki respondentów o liczebnościach $l_1, l_2, \ldots, l_m$ przez $\bar{l}$ oznaczono średnią arytmetyczną z liczebności wszystkich wiązek, natomiast przez $\overline{w}_j$ średnią arytmetyczną z wartości wag w *j*-tym zespole.

[176] Współczynnik kowariancji w zbiorze wyników ESS5-PL wynosi 1,609.

[177] Dokładne wartości estymatorów przyrostu wariancji wynikającej z uzespołowienia próby dla parametrów l^* oraz $\bar{l}$ zamieszczone zostały w tabeli aneksowej A4.

[178] Jedynym wyjątkiem od tego są dwie warstwy miejskie (miast o wielkości populacji od 20 do 49,9 tys. mieszkańców oraz od 50 do 99,9 tys. osób), w których wartości współczynnika *VIF* były nieznacznie większe od jedności.

[179] Mowa tutaj o przekazywanej krajowym koordynatorom badań instrukcji opisującej schematy doboru prób badawczych (por. *Sampling for the ESS Round V* 2010).

tych badań (por. np. *ESS1 Sampling Report* 2002: 5–6*, ESS2 Sampling Report* 2004: 61), odnaleźć można ciekawą rekomendację pociągającą za sobą ważne konsekwencje praktyczne. W dokumentach tych, w części poświęconej estymacji $DEFF_c$, zamieszczono następujące stwierdzenie:

> jeśli nie ma się w ogóle dostępu do danych empirycznych, na których można by oprzeć szacunki ρ, sugerujemy, aby [w estymacji $DEFF_c$ – P.J.] wykorzystać wartość *roh* równą 0,02. (*Sampling for the ESS Round V* 2010: 14)

Jest to znamienne o tyle, że nawet w tak zaawansowanym projekcie metodologicznym, jakim niewątpliwie jest badanie ESS, niewiele w sumie wiadomo o konsekwencjach wynikających z uzespołowienia próby badawczej. Wprawdzie zalecana tu wielkość współczynnika korelacji wewnątrzzespołowej wyznaczona została (jeszcze przed pierwszą falą ESS z 2002 roku) w oparciu o wyniki wcześniejszych badań (pochodzących z kilku krajów) o tematyce zbliżonej do tej, która miała być poruszana w projekcie ESS (por. Lynn i in. 2007: 114), jednak siłą rzeczy takie aprioryczne założenie pozostaje obarczone znacznym ryzykiem i w konsekwencji skutkować może niedoszacowaniem efektu losowania zespołowego. Chociaż koordynatorzy badań krajowych zachęcani byli do podjęcia wysiłków na rzecz estymacji wielkości ρ (tak, aby szacunki tych parametrów odpowiadały rzeczywistości empirycznej danego kraju), to jednak w wielu państwach rezygnowano z dodatkowych analiz, polegając wyłącznie na rekomendacjach projektowych (por. Sawiński 2011: 2)[180]. Zarzut ten ma charakter fundamentalny, nie dotyczy jednak polskiej części badania ESS. Istotnie, przed losowaniem próby badawczej do badań ESS5-PL (ed. 2010) przeprowadzone zostały dodatkowe studia metodologiczne oparte na danych ESS4-PL (ed. 2008), których celem było oszacowanie poziomu wewnątrzwiązkowej homogenizacji jednostek w badaniach ESS realizowanych w Polsce. Studia te wykazały, iż przeciętna (tj. uśredniona z mierników ρ dla 183 zmiennych) wartość tego współczynnika wyniosła 0,12 (por. *Sampling Design in ESS5-PL* 2010: 3), co oznacza, że była znacznie większa od wielkości rekomendowanej w dokumentacji projektowej.

Znajduje to również potwierdzenie w analizach danych piątej rundy ESS z 2010 roku. Z przeprowadzonych studiów wynika bowiem, że przeciętna war-

[180] W referacie wygłoszonym w ramach *4th Conference of the European Survey Research Association* (Lozanna, 18–22 lipca 2011 roku) Z. Sawiński wygłosił tezę, że taki stan rzeczy wynika głównie z tego, iż „wszelkie eksperymenty prowadzone w tym obszarze są [dla badaczy – P.J.] ryzykowne, gdyż im większa jest wartość roh, tym bardziej zwiększona powinna być liczebność próby badawczej" (Sawiński 2011: 2). Kryje się za tym niewyrażona wprost sugestia, że chociaż przyjęcie rekomendacji zawartej w dokumentacji projektowej ESS-u może odbiegać od rzeczywistych wielkości współczynnika korelacji wewnątrzzespołowej w próbie badawczej, to jednak czyni się tak z powodów czysto instrumentalnych, unikając wzrostu kosztów związanych z realizacją większej liczby wywiadów.

tość współczynnika korelacji wewnątrzzespołowej była równa 0,15 (por. wyniki zaprezentowane w tabeli IV.4.), co oznacza, iż w ESS5-PL poziom wewnątrzwiązkowej homogenizacji jednostek był nawet nieco wyższy niż w ESS4-PL. Wartości tych (na poziomie ogólnym) nie da się jednak bezpośrednio porównać, co w sumie czyni takie zestawienia mało interesującymi. Po pierwsze, kwestionariusze wywiadów z czwartej oraz piątej rundy badań ESS różnią się między sobą pytaniami w modułach rotacyjnych (całe części D oraz G w narzędziach badawczych). Po drugie, takie zestawienia longitudinalne wskaźników ogólnych ograniczone są także z uwagi na fakt, iż analizy dla badań ESS4-PL prowadzone były w odniesieniu do znacznie większej liczby pytań[181] (183 zmienne) niż zaprezentowane tu studia nad ESS5-PL (90 zmiennych). Jednak pomimo tych ograniczeń – w odniesieniu do pojedynczych pytań – warto porównać wyniki uzyskane w ramach realizacji obu fal ESS-u oraz sformułować w tym względzie wnioski ogólne.

Należy przy tym wskazać, że w niezwykle pouczającej prezentacji *Intracluster homogeneity in Survey samples: a neglected tool*, zaprezentowanej przez Z. Sawińskiego (2011: 1–16) w ramach czwartej konferencji *European Survey Research Association*, uszczegółowione oraz usystematyzowane zostały wyniki analiz metodologicznych poświęconych estymacji mierników korelacji wewnątrzzespołowej w ESS4-PL. Autor ten skoncentrował się zarówno na podobieństwie jednostek wynikającym z uzespołowienia próby, jak i na homogenizacji odpowiedzi udzielanych przez respondentów. Ten drugi wątek rozważał w świetle wpływu ankieterskiego na odpowiedzi udzielane przez respondentów. Chociaż taka dwoistość analizy jest w pewnej mierze uzasadniona dla badań realizowanych w Polsce (w każdej wylosowanej wiązce wywiady prowadzone były przez tego samego ankietera, co oznacza, że podobieństwo jednostek w obrębie zespołów mogło rzeczywiście wynikać zarówno z doboru zespołowego, jak i z wpływu ankieterskiego), to jednak sposób terenowej realizacji próby (tzn. przydzielanie jednemu ankieterowi całej wiązki respondentów) nie daje możliwości estymacji efektu ankieterskiego wyrażonego miarą $DEFF_{ANK}$[182]. Mimo tych ograniczeń, wnioski sformułowane przez Sawińskiego w zakresie oddziaływania ankieterów na uzyskiwane wyniki są bardzo pouczające i znajdują również potwierdzenie w odniesieniu do badań piątej rundy ESS.

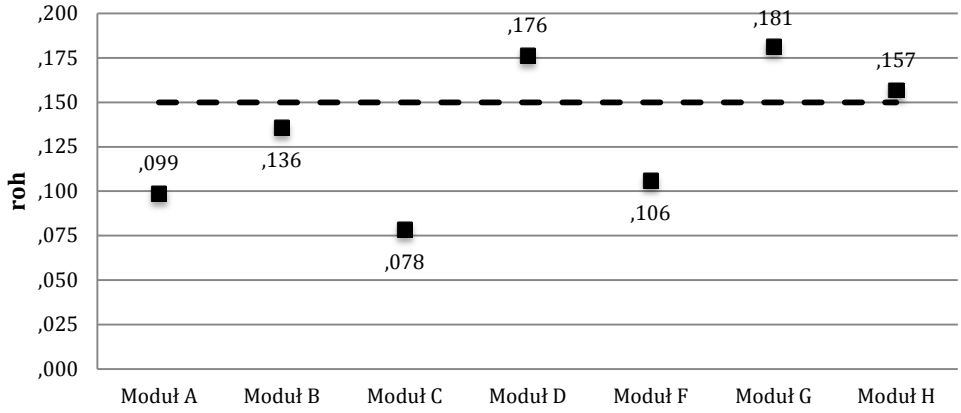
Dla przykładu, w obu edycjach ESS-u okazało się, że znacznie mniejszym poziomem wewnątrzwiązkowej homogenizacji charakteryzują się moduły kwe-

---

[181] Wskazywano już, że analizy ESS5-PL przeprowadzone zostały w obrębie tych pytań, które były zadawane wszystkim z respondentów, a wartości zmiennych dało się zobrazować przy pomocy średniej arytmetycznej. Ponieważ w analizach dla ESS4-PL osłabiono warunek drugi, to poziom wewnątrzzespołowej homogenizacji wyznaczany był również dla zmiennych porządkowych oraz nominalnych (por. Sawiński 2011: 2–3).

[182] Było to już przedmiotem dyskusji w części II.2.2. drugiego rozdziału tej pracy.

stionariusza ESS umieszczone w początkowej części narzędzia badawczego, większym natomiast te, które zawierają pytania zadawane już pod koniec wywiadu. Z. Sawiński (2011: 4–5) tłumaczył taki stan rzeczy efektem ankieterskim, co w jego opinii pozostawało zgodne z przekonaniem, iż efekt ten powinien odgrywać większą rolę pod koniec kwestionariusza, czyli wtedy, gdy respondenci są bardziej zmęczeni i podatni na sugestie ze strony ankietera.



Ryc. IV.2. Wielkości współczynników korelacji wewnątrzzespołowej (*roh*) na przykładzie wyników badań ESS5-PL (ed. 2010)

Źródło: Obliczenia własne na podstawie repozytorium danych ESS5-PL oraz ESS5-SDDF-PL

Znajduje to również potwierdzenie w analizach danych ESS5-PL. Istotnie można zaobserwować, że pierwsze trzy części tego narzędzia, tj. moduł A – media, zaufanie do ludzi, B – polityka oraz C – poczucie dobrobytu, wykluczenie społeczne, religia (zawarte zarówno w ESS4, jak i w ESS5) charakteryzują się niższymi (od średniej w całym zbiorze analizowanych pytań) wartościami korelacji wewnątrzzespołowych, natomiast moduły rotacyjne D – zaufanie do wymiaru sprawiedliwości oraz G – praca, rodzina, dobrobyt, a także moduł z pytaniami skalowymi o ludzkie wartości (w ESS5 oznaczony jako H, w ESS4 jako G), cechują się już większym poziomem homogenizacji odpowiedzi udzielanych przez respondentów z tych samych zespołów.

Zaproponowana przez Sawińskiego hipoteza o zależności wielkości ρ od umiejscowienia pytania w narzędziu badawczym jest jednak tylko z pozoru przekonująca. Trzeba bowiem zauważyć, iż analiza szczegółowych wartości mierników korelacji wewnątrzzespołowej dla pytań ESS5-PL (por. wyniki z tabeli aneksowej A.IV.2.) oraz ESS4-PL (por. Sawiński 2011: 11–16) ukazuje już, że wielkości tych miar są jednak dość wyraźnie zróżnicowane w obrębie

tych samych modułów kwestionariusza. Ponadto, wcale nie jest tak, że najwyższym poziomem wewnątrzwiązkowej homogenizacji charakteryzują się pytania z ostatniego lub nawet przedostatniego modułu kwestionariusza. Wydaje się raczej, że wartość ρ pozostaje w większym stopniu uzależniona od kontekstu pytania, niż od jego umiejscowienia w narzędziu badawczym. Do wniosków takich doszedł zresztą Sawiński, rozpatrując zbiór wyników ESS4-PL. Autor ten wskazał, że:

> pytania, które dotyczą zagadnień ważnych oraz kontrowersyjnych dla respondentów […] mają zazwyczaj niższą wartość [współczynnika korelacji wewnątrzzespołowej – P.J.] […]. Jest to zgodne z panującym przekonaniem, że im respondent jest bardziej zaangażowany w przebieg wywiadu, tym mniej staje się podatny na oddziaływanie ze strony ankietera. (Sawiński 2011: 5)

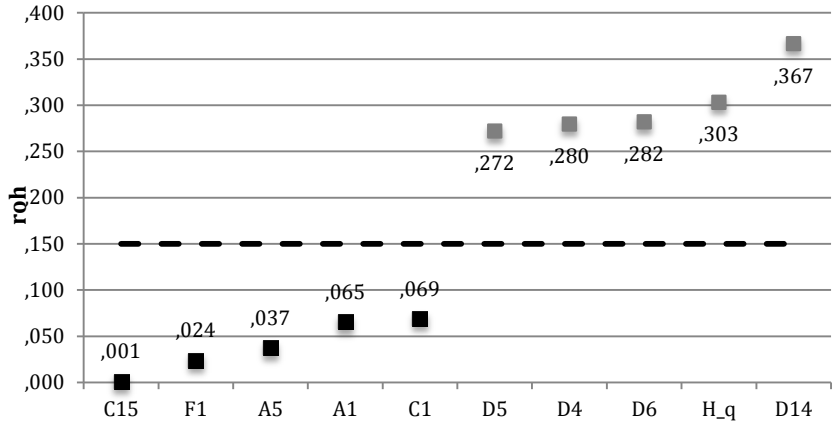
Na poparcie tej tezy Sawiński przytacza oszacowania wielkości ρ dla następujących pytań: B1 – zainteresowanie polityką (ρ = 0,02), B23 – umiejscowienie na skali lewica-prawica (ρ = 0,01), C15 – samoocena stanu zdrowia (ρ = 0,01), przeciwstawiając je znacznie większym wielkościom współczynników korelacji wewnątrzzespołowej dla dwóch grup pytań dotyczących prestiżu zawodów[183] (w jego ocenie pytania te dotyczą kwestii mało istotnych dla badanych, w przeciwieństwie do spraw poruszanych w pytaniach B1, B23 oraz C15).

Przypuszczenie, iż niższy poziom homogenizacji wewnątrzzespołowej dotyczyć będzie pytań ważnych i kontrowersyjnych, mniejszy natomiast pytań ogólnych, tj. dotykających kwestii mało istotnych, nie znajduje jednak jakiegoś szczególnego potwierdzenia w badaniach piątej fali ESS-u. Przyglądając się danym zamieszczonym w tabeli aneksowej A.IV.2., można bowiem zauważyć, że z przywołanych przez Sawińskiego pytań charakteryzujących się w zbiorze danych ESS4-PL niewielkim poziomem wewnątrzzespołowej homogenizacji jedynie w ramach pytania C15 wielkość ρ ukształtowała się w ESS5-PL na podobnie niskim poziomie. Dla pozostałych dwóch pytań wybranych przez Sawińskiego do uprawomocnienia jego przypuszczenia otrzymano w ESS5-PL już znacznie większe wartości współczynników ρ (odpowiednio dla B1 – $\rho = 0.098$, natomiast dla B23 – $\rho = 0.121$). Sprawa staje się o tyle bardziej kłopotliwa, że wśród pięciu zmiennych o najwyższym poziomie wewnątrzzespołowej homogenizacji w ESS5-PL znalazły się trzy pytania (D4, D5, oraz D6), które z całą pewnością były dla respondentów „kontrowersyjne" (wszystkie dotyczyły tego, jak bardzo prawdopodobne jest to, że respondent zostanie złapany oraz ukarany za hipotetycznie popełnione przez siebie wykroczenie lub przestępstwo), a także pytanie, które dotykało problemów „ważnych" dla badanych osób (py-

[183] Pierwsza grupa zmiennych charakteryzuje się przeciętą wielkością ρ = 0,20 (22 itemy), druga natomiast poziomem ρ = 0,22 (10 itemów) (por. Sawiński 2011: 5).

tanie D14, w którym respondenci wyrażali opinię na temat szybkości interwencji podejmowanych przez policję). To ostatnie pytanie cechowało się zresztą najwyższym poziomem wewnątrzwiązkowej homogenizacji w zbiorze wyników ESS5-PL.

Porównując obie serie pomiarów badań ESS z 2008 oraz 2010 roku, można sformułować przypuszczenie akcentujące nieco inny charakter zależności zachodzącej pomiędzy rodzajem pytania a wielkością współczynnika korelacji wewnątrzzespołowej. Otóż, można wskazać, że w zakresie pytań odnoszących się



Ryc. IV.3. Maksymalne oraz minimalne wielkości współczynników korelacji wewnątrzzespołowej (*roh*) w zbiorze wyników badań ESS5-PL (ed. 2010)

Źródło: obliczenia własne na podstawie repozytorium danych ESS5-PL oraz ESS5-SDDF-PL

do bezpośrednich doświadczeń życiowych respondentów, ich codziennej aktywności oraz faktów z ich życia, współczynniki ρ okazują się mniejsze niż w ramach pomiaru opinii, czy też przekonań o charakterze ogólnym. Istotnie, w zbiorze danych ESS5-PL zmienne charakteryzujące się najniższymi wielkościami wewnątrzzespołowej homogenizacji dotyczyły właśnie pytań o samoocenę stanu zdrowia (C15), liczbę osób zamieszkujących gospodarstwa domowe (F1), czas poświęcany w typowym dniu tygodnia na czytanie gazet (A5) oraz oglądanie telewizji (A1), jak też o poczucie własnego szczęścia (C1). Te zaś, dla których otrzymano najwyższe wielkości ρ, dotyczyły pytań o opinie oraz przekonania. To, co wydaje się szczególnie ważne w kontekście tej hipotezy, to fakt, że wszystkie te pytania, które w zbiorze ESS5-PL charakteryzował najniższy poziom wewnątrzzespołowej homogenizacji, cechowała również niewielka (lub mówiąc dokładnie, niższa od przeciętnej) wartość ρ w pomiarze ESS4-PL.

Efektywność schematu doboru próby w ESS5-PL

Przyglądając się schematowi doboru próby w ESS5-PL pod kątem kryteriów warunkujących możliwość estymacji miary $DEFF_{TOTAL}$ zgodnie z procedurą oryginalnie wykorzystywaną przez zespół polskich koordynatorów tego projektu (por. wariant I w tab. IV.3.), zwracano już uwagę na dość restrykcyjne wymogi, jakie muszą być spełnione, aby ową procedurę można było w ogóle zastosować. Abstrahując w tym momencie od przywoływanych już wcześniej argumentów należy przypomnieć, że takich kryteriów schemat losowania próby w ESS5-PL nie spełniał, co stawiało pod znakiem zapytania trafność przeprowadzanych szacunków $DEFF_{TOTAL}$. Do dyskusji tej nie warto już powracać, choć trzeba przypomnieć, iż zaproponowano dwie alternatywne procedury estymacji tego współczynnika (nazwane umownie wariantem drugim oraz trzecim), dla których spełnione były już formalne kryteria umożliwiające ich wykorzystanie w ESS5-PL. Głównym celem tej ostatniej części rozdziału będzie zatem znalezienie odpowiedzi na pytanie, czy zastosowanie odmiennych procedur oceny efektywności schematu doboru próby prowadzi do jakiś znaczących różnic w uzyskiwanych wartościach współczynników $DEFF_{TOTAL}$.

Tabela IV.5. Porównanie wartości $DEFF_c$ oraz $DEFF_{TOTAL}$ na przykładzie wybranych danych z badań ESS5-PL (ed. 2010)

ESS5–2010	**Wariant estymacji**[i)]					
	$DEFF_c^I$	$DEFF_c^{II}$	$DEFF_c^{III}$	$DEFF_{TOTAL}^I$	$DEFF_{TOTAL}^{II}$	$DEFF_{TOTAL}^{III}$
Moduł A (6 pytań)	1,124	1,186	1,179	1,140	1,191	1,181
Moduł B (22 pytania)	1,172	1,252	1,256	1,188	1,258	1,257
Moduł C (5 pytań)	1,099	1,150	1,153	1,114	1,155	1,154
Moduł D (31 pytań)	1,223	1,325	1,334	1,240	1,331	1,336
Moduł F (2 pytania)	1,144	1,177	1,161	1,160	1,182	1,162
Moduł G (3 pytania)	1,229	1,334	1,336	1,246	1,340	1,337
Moduł H (21 pytania)	1,198	1,290	1,292	1,214	1,296	1,293
Ogółem (90 pytań)	**1,189**	**1,277**	**1,281**	**1,205**	**1,283**	**1,283**

Źródło: obliczenia własne na podstawie repozytorium danych ESS5-PL oraz ESS5-SDDF-PL

[i)] Warianty estymacji $DEFF_c$ oraz $DEFF_{TOTAL}$ zgodne z opisem w tabeli IV.3. Obliczenia $DEFF_c^I$ oraz $DEFF_{TOTAL}^I$ z wykorzystaniem parametru $\bar{l}$, z kolei $DEFF_c^{II}$ i $DEFF_c^{III}$, a także $DEFF_{TOTAL}^{II}$ i $DEFF_{TOTAL}^{III}$ z wykorzystaniem parametru l^*.

W tabeli A.4. znajdującej się w aneksie książki odnaleźć można szczegółowe dane o wielkościach miar $DEFF_c$ oraz $DEFF_{TOTAL}$ wyznaczonych dla każdego z analizowanych pytań z kwestionariusza ESS5. Przedstawiona tam też tabela IV.5. zawiera z kolei uśrednione wartości tych dwóch współczynników obliczone dla kolejnych modułów kwestionariusza wywiadu, jak też dla wszystkich pytań (wynik ogółem).

Analizując dane zawarte w tabeli, należy zwrócić uwagę na fakt, iż w badaniu ESS5-PL na wielkość całkowitego efektu schematu doboru próby zdecydowanie większy wpływ wywierało jej uzespołowienie (tj. wiązkowanie respondentów w części warstw), niż ważenie danych, będące konsekwencją niejednakowych prawdopodobieństw selekcji jednostek do próby badawczej. Porównując wartości $DEFF_c$ oraz $DEFFF_{TOTAL}$ (estymowane każdą z trzech rozważanych tu procedur), można bowiem zauważyć, iż efekt uzespołowienia stanowił (niezależnie od zastosowanego wariantu estymacji) mniej więcej 99% całkowitej wielkości miary $DEFFF_{TOTAL}$[184]. Ujmując to inaczej, w badaniu ESS5-PL nierówne prawdopodobieństwa selekcji mają zupełnie marginalny wpływ na utratę precyzji estymacji. Wskazywano już na to uwagę w literaturze przedmiotu, w studiach poświęconym wcześniejszym odsłonom tego projektu (por. Vehovar 2007: 343; Gabler i in. 2006: 116).

Istotne pozostaje jednak to, do jakich różnic w ocenie efektywności schematu doboru próby prowadzi zastosowanie odmiennych procedur szacowania wielkości $DEFF_c$ oraz $DEFF_{TOTAL}$. Jak już wspomniano, wariant pierwszy odpowiada procedurze uproszczonej, czyli takiej, w której zakłada się, że struktura warstw w próbie i populacji jest taka sama, a także, że miarę $DEFF_{TOTAL}$ da się wyznaczyć jako iloczyn efektu uzespołowienia oraz ważenia danych. W wariancie drugim uwzględnia się nieproporcjonalną strukturę warstw w próbie badawczej oraz zakłada równość współczynników korelacji wewnątrzzespołowej we wszystkich warstwach populacji w wiązkowanej części próby badawczej. Z kolei w wariancie trzecim efekt uzespołowienia wyznacza się niezależnie dla każdej warstwy, czyli uwzględnia się międzywarstwowe zróżnicowania wielkości ρ.

W pierwszej kolejności porównany będzie drugi oraz trzeci wariant estymacji tych współczynników. Zdiagnozowane będzie więc, czy międzywarstwowe zróżnicowanie poziomu wewnątrzwiązkowej homogenizacji jednostek wpływa w jakiś sposób na szacunki $DEFF_c$ oraz $DEFF_{TOTAL}$. Zestawiając ze sobą wartości uogólnione (tzn. średnie wyniki dla wszystkich zmiennych), należy stwierdzić, że rozbieżności obu miar są niewielkie. Dla $DEFF_c$ wynoszą jedynie 0,006, natomiast dla $DEFF_{TOTAL}$, po zaokrągleniu do części tysięcznych, otrzymano nawet tę samą wartość miary. Przyczyna tego stanu rzeczy jest niezwykle prosta

[184] Nie ma sensu dalej rozważać międzymodułowych zróżnicowań wartości różnych wariantów estymacji współczynników $DEFF_c$ oraz $DEFFF_{TOTAL}$. Wszystkie te dysproporcje są bowiem prostym odzwierciedleniem różnic w analizowanych już wcześniej wielkościach współczynników korelacji wewnątrzzespołowej.

i wynika z niewielkiego zróżnicowania współczynników ρ w warstwach populacji. Wskazano już na to w części analiz dedykowanych wewnątrzwiązkowej homogenizacji jednostek. Innymi słowy, duża zbieżność wartości $DEFF_c$, jak i $DEFF_{TOTAL}$ (otrzymanych porównywanymi tu wariantami estymacji tych miar), jest konsekwencją specyfiki próby ESS5-PL. Gdyby różnice w międzywarstwowych wielkościach współczynników korelacji wewnątrzzespołowej były większe, to trzeci wariant estymacji dawałby bardziej trafne oszacowania tych miar niż wariant określony jako drugi.

Zestawienia tabelaryczne ukazują również, że zastosowanie pierwszej, czyli uproszczonej procedury estymacji miar $DEFF_c$ i $DEFF_{TOTAL}$, wskazuje na mniejszy przyrost wariancji, niż wynika to z wartości otrzymanych drugą oraz trzecią procedurą szacowania efektywności schematu doboru próby. Porównując wyniki na poziomie ogólnym, widać, że w pierwszym wariancie estymacji, wartość miary $DEFF_c$ okazała się o mniej więcej 9 pp. mniejsza niż uzyskana w wariancie drugim i trzecim. Dla $DEFF_{TOTAL}$ różnica była nieco mniejsza i wynosiła mniej więcej 8 pp. Jest to wniosek niezwykle ważny. Ukazuje bowiem, że zastosowanie uproszczonej wersji estymatora $DEFF_c$ oraz $DEFF_{TOTAL}$ może skutkować niedoszacowaniem stopnia, w jakim dany schemat doboru próby oddziałuje na precyzję estymacji. Z drugiej strony, gdyby w wariancie pierwszym zamiast parametru $l' = \bar{l}$ przyjąć, tak jak w *II* i *III*, $l' = l^*$, to wartości estymowane tymi trzema procedurami byłyby już bardzo podobne. Oznaczałoby to, że przyjęcie nieuprawnionej metody estymacji $DEFF_c$ i $DEFF_{TOTAL}$ nie miałoby już jakichś szczególnie poważnych konsekwencji w szacowaniu efektywności próby ESS5-PL. Warto poszukać zatem odpowiedzi na dodatkowe pytanie: jakie czynniki decydowałyby o tak wysokiej zbieżności wartości tych miar? Odpowiedź jest niezwykle prosta. W ESS5-PL te niewielkie różnice byłyby konsekwencją znacznego podobieństwa proporcji warstw w próbie oraz w populacji. Wystarczy zauważyć, że dla danych ESS5-PL indeks niepodobieństwa empirycznego rozkładu warstw w próbie oraz rozkładu tych samych warstw w populacji wynosi jedynie 2,5%[185]. Przy tak bardzo podobnej strukturze warstw próby oraz populacji, z jaką ma się do czynienia w badaniach ESS5-PL, konsekwencją pominięcia (w pierwszym wariancie estymacji miary $DEFF_{TOTAL}$) efektu nieproporcjonalnej lokalizacji jednostek w warstwach próby (tzn. przyjęcia błędnego założenia o jej proporcjonalności), byłoby niedoszacowanie wielkości $DEFF_{TOTAL}$ o krotność tej miary równą 1,014 (czyli o 1,4%)[186]. Oczywiście gdyby te dysproporcje były większe, to różnice w wartościach miar byłyby bardziej znaczące.

---

[185] Indeks niepodobieństwa (oryg. *dissimilarity index*) jest powszechnie stosowaną miarą umożliwiającą porównywanie dwóch różnych rozkładów tej samej zmiennej. Jego wartości mieszczą się w przedziale [0;1]. Im wartość jest bliższa zeru, tym rozkłady są bardziej zbieżne (por. Kuha et al. 2011: 376, Sawiński 2010: 192, Mulekar et al. 2008: 2099).

[186] Wartość ta wyznaczona została jako iloraz wielkości $DEFF^{II}_{TOTAL}$ i $DEFF^{I}_{\mathrm{TOTAL}}$, z pominięciem $DEFF_p$ oraz $DEFF_c$. Przy takich założeniach $DEFF^{I}_{TOTAL}=1$, natomiast $DEFF^{II}_{TOTAL} = 1{,}014$.

* * *

Abstrahując od pewnych specyficznych właściwości przedstawionego w tym rozdziale schematu doboru próby badawczej z polskiej części piątej rundy ESS (ed. 2010), a także od porównywanych metod szacowania efektywności tego schematu doboru próby, należy przede wszystkim uwypuklić trudności, przed jakimi stają badacze przystępujący do losowania prób reprezentatywnych. Wyzwaniem nie jest bowiem tylko i wyłącznie ustalenie optymalnej strategii doboru respondentów (nie wspominając już nawet o komplikacjach związanych z wyborem najlepszego operatu doboru próby), ale także takie zdefiniowanie estymatorów efektu schematu doboru próby, które pozwoliłyby na poprawną ocenę efektywności wykorzystanego schematu losowania respondentów. Nie jest to sprawa prosta głównie z tego powodu, że nie da się określić wszystkich schematów, według których losowania może być przeprowadzane, da się jedynie wskazać na pewne ogólne klasy takich działań (por. Groves 1989: 253). Scharakteryzowane w empirycznej części tego rozdziału warianty estymacji współczynnika *design effect* są tym samym odpowiednie dla schematu doboru próby w ESS5-PL, nie będą miały już jednak zastosowania w innych schematach losowania. Warto jednocześnie pamiętać, że uogólnionym sposobem określania efektywności złożonych schematów doboru próby może być szacowanie wielkości błędów losowych z wykorzystaniem – nieomówionych w tej pracy – technik replikacyjnych, w tym niezwykle popularnej procedury balanced repeated replication, zwłaszcza w tak zwanym wariancie Faya (por. np. Särndal i in. 1992).

ROZDZIAŁ V

Terenowa realizacja sondażowych prób badawczych

Niepełna realizacja próby sondażowej to jedno z największych wyzwań, z jakimi przychodzi zmierzyć się w praktyce badawczej. Doniosłość problemów związanych z niedostępnością jednostek oraz z wynikającym z tego ryzykiem wypaczenia wyników pomiaru potwierdza obszerna literatura przedmiotu. O ile jednak w studiach dotyczących teorii próbkowania uwaga badaczy koncentruje się zazwyczaj na sposobach definiowania wielkości błędu braku danych oraz na statystycznym opisie mechanizmów kształtujących charakter niedostępności wylosowanych jednostek, czy też na metodach służących eliminowaniu lub ograniczaniu negatywnych skutków niepełnej realizacji próby, o tyle w studiach socjologicznych (lub raczej w metodologii badań surveyowych) uwaga przesunięta zostaje w stronę identyfikacji społeczno-demograficznych determinant niedostępności jednostek, prób empirycznego oszacowania wielkości błędów braków danych, wypracowania modeli teoretycznych opisujących wpływ czynników społecznych (i pozaspołecznych) na szanse nawiązania kontaktu z wylosowanymi osobami oraz na szanse realizacji z nimi wywiadów, ustalenia motywów skłaniających wylosowane osoby do partycypacji lub też do odmowy udziału w sondażu, czy też wreszcie wypracowania procedur terenowych umożliwiających podjęcie skutecznych działań na rzecz zwiększenia odsetka realizacji prób badawczych oraz gromadzenia dodatkowych informacji (o jednostkach niedostępnych) wykorzystywanych w modelowaniu jednostkowych skłonności do udziału w badaniu. Podejścia te – statystyczne oraz socjologiczne – są oczywiście komplementarne, nie ma bowiem większego sensu konstruowanie modeli teoretycznych (opisujących problem praktyczny, jakim – bez wątpienia – jest niepełna realizacja próby) bez osadzenia ich w rzeczywistości empirycznej. Podobnie zresztą nie da się podejmować efektywnych działań badawczych oraz postbadawczych służących ograniczaniu negatywnych skutków niepełnej realizacji próby, nie rozpoznając wpierw możliwego przełożenia owych przedsięwzięć na wielkości błędów systematycznych.

Idea ta przyświeca również studiom przedstawionym w tym rozdziale. Ich głównym celem będzie zaprezentowanie spójnej koncepcji analizy reprezentatywności próby w sytuacji jej niepełnej realizacji.

V.1. Postrealizacyjna klasyfikacja jednostek próby badawczej

W literaturze poświęconej zagadnieniom niepełnej realizacji próby można zauważyć wysoką zbieżność sposobów definiowania przez badaczy kategorii jednostek niedostępnych. Metodologowie badań społecznych przyjmują najczęściej, że niedostępne są te osoby wylosowane do próby, z którymi nie udało się przeprowadzić badania, a o których wiadomo, iż przynależą (lub przynajmniej przypuszcza się, że mogą przynależeć) do badanej populacji[187] (por. na przykład Brick i in. 2009: 163; Grzeszkiewicz-Radulska 2009: 24; Domański 2006: 31; Goyder 1987: 9; Bethlehem i in. 1985: 287). Niedostępność jednostek jest przy tym – przede wszystkim – efektem odmówienia udziału w badaniu oraz skutkiem braku możliwości dotarcia do respondentów (tj. nawiązania z nimi kontaktu), choć oczywiście może wynikać też z czegoś innego, na przykład ze złego stanu psychofizycznego dobranej osoby, wykluczającego jej udział w badaniu, czy też być konsekwencją barier językowych uniemożliwiających przeprowadzenie wywiadu[188] (por. Martin 2011: 18; Stoop i in. 2010: 13–29; Sztabiński i in. 2009: 67–95, Billiet i in 2007: 138; Dorofeev i in. 2006: 32–35; Weisberg 2005: 160; Groves i in. 2004: 169–170; Stoop 2005: 64–80; Stoop 2004: 38–44; Biemer i in. 2003: 80–81; Atrostic i in. 2001: 211; Groves i in. 1998: 12–13; Groves 1989: 137; Cochran 1977: 359).

---

[187] Nawet jeśli w literaturze przyjmuje się nieco mniej precyzyjną definicję traktującą jednostki niedostępne (nie–respondentów) jako zbiór wszystkich elementów próby, z którymi nie udało się przeprowadzić wywiadu (nie uszczegóławiając, iż poza grupą jednostek niedostępnych znajdują się osoby, co do których ustalono, iż do populacji nie należą, to znaczy znalazły się w niej w wyniku błędów operatu) (por. Groves 1989: vi), to wskaźniki realizacji próby wyznaczane są już z pominięciem jednostek nienależących do populacji (por. Groves 1989: 136–145). Warto wskazać, iż na wielkość proporcji respondentów w zbiorze wszystkich (bez wyjątku) osób wylosowanych do próby używa się często w polskiej literaturze pojęcia *wskaźnika kompletności próby* (por. Grzeszkiewicz-Radulska 2009: 25, Domański 1999: 69), *odsetka uzyskanych materiałów* (por. Daniłowicz i in. 1992: 126) lub też *podstawowego wskaźnika realizacji próby* (por. Jabkowski 2010: 34).

[188] W dalszej części analiz wykazane będzie, iż taki podział jednostek niedostępnych uzasadniony jest zarówno względami praktycznymi, to znaczy umożliwia podjęcie konkretnych działań skutkujących zwiększeniem poziomu realizacji próby, jak też ma uprawomocnienie merytoryczne, to znaczy wynika z tego, że wyróżnione kategorie jednostek niedostępnych różnią się między sobą. Co więcej, w odmienny sposób oddziałują one na błędy braków danych. Innymi słowy, traktowanie jednostek niedostępnych jako grupy homogenicznej byłoby zbyt dużym uproszczeniem tego zjawiska i nie miałoby żadnego przełożenia na praktykę badawczą.

Naturalnie z takim postrealizacyjnym podziałem jednostek wylosowanej próby badawczej związana jest konieczność gromadzenia informacji o przebiegu badania, w tym o kolejnych próbach nawiązania kontaktu z respondentem oraz o efektach takich działań[189]. W projektach naukowych wydaje się to obecnie czymś zupełnie oczywistym, wszak wiedza o powodach niezrealizowania wywiadów jest wykorzystywana zarówno do podjęcia działań zwiększających szanse dotarcia do osób trudno dostępnych lub realizacji wywiadu z osobami niechętnymi do udziału w badaniu (por. na przykład Bates i in 2008: 591–612; Billiet i in. 2007: 135–162; Stoop 2005: 50), jak i też w postrealizacyjnym ważeniu danych wynikowych (por. na przykład Billiet i in. 2009: 27–28; Kreuter i in. 2009: 203–226). Warto przy tym zauważyć, że chociaż postulat ujednolicenia zasad postterenowej klasyfikacji jednostek wylosowanych do próby badawczej oraz stosowanie tych samych miar do opisu poziomu jej realizacji jest szczególnie ważny w przedsięwzięciach o charakterze porównawczym (por. Jowell 1998: 168–177) – i to zarówno w wymiarze międzykrajowym, jak i też wzdłużczasowym (por. Stoop 2005: 50) – to jednak konieczność standaryzacji mierników poziomu realizacji próby wydaje się istotna w badaniach sondażowych w ogóle (por. Groves 1989: 140). Wypracowanie w tym zakresie jednolitych wzorców postępowania jest niezwykle trudne, nie chodzi wszak tylko o to, by upowszechnić te same standardy w obrębie jakiegoś wybranego projektu, lecz by przekonać badaczy do stosowania tych samych (lub przynajmniej dających się porównać) norm w różnych badaniach (por. Kasprzyk i in. 2003: 353; Smith 2002: 27–40). Na potrzebę wprowadzenia do świata praktyki badawczej systemu jednoznacznych zasad wyznaczania wielkości wskaźników realizacji próby wskazywała ostatnio w polskiej literaturze K. Grzeszkiewicz-Radulska (2009: 25–36), a wcześniej postulat taki zgłaszali Paweł Daniłowicz oraz F. Sztabiński (1992: 122–141).

O konieczności wypracowania w tym względzie wspólnych standardów mówiono już jednak w literaturze światowej w latach 70. XX wieku (por. Kviz 1977; Wiseman i in. 1979). W późniejszym okresie zarzut zbyt daleko posuniętej dowolności w tej dziedzinie formułowany był między innymi przez Grovesa (1989: 136–145)[190], Lesser i in. (1992: 103–129) oraz Hidiroglou i in. (1993:

---

[189] Dobrym przykładem takich praktyk są formularze doboru respondentów wykorzystywane w Europejskim Sondażu Społecznym. Dla szóstej rundy badań ESS dostępne są one pod następującym adresem internetowym: http://www.europeansocialsurvey.org/docs/round6/fieldwork/source/ESS6_source_contact_forms.pdf

[190] R. Groves w czwartym rozdziale monografii *Survey Errors and Survey Costs* wyraził to w następujący sposób: „chociaż istnieją rekomendacje co do sposobów wyznaczania wskaźników realizacji próby [tutaj przywoływany jest dokument z 1982 roku opracowany przez organizację CASRO, czyli stowarzyszenie zrzeszające instytucje badawcze, głownie z USA – P.J.], to […] zamiast stosować się do jednolitych wytycznych, różni badacze wykorzystują różne wskaźniki, dla różnych potrzeb, otrzymując odmienne mierniki realizacji próby badawczej" (Groves 1989: 140).

81–94). Taka dezaprobata dla braku jednolitych wzorców postępowania połączona z ukazaniem wyraźnej dysfunkcyjności, wynikającej z dowolności stosowanych standardów przyczyniła się (w Stanach Zjednoczonych) do wypracowania zasad ustalania wielkości wskaźników realizacji próby, a także do przyjęcia jednoznacznych norm kodyfikacji jednostek do zbioru respondentów oraz osób niedostępnych. W 1998 roku pod auspicjami The American Association for Public Opinion Research (AAPOR) wydany został po raz pierwszy dokument zawierający dyspozycje określające reguły klasyfikacji elementów próby do jednego z czterech podzbiorów jednostek, to jest do grupy respondentów, osób niedostępnych, jednostek o nieustalonym statusie przynależności do populacji oraz osób błędnie sklasyfikowanych w operatach doboru próby jako elementy populacji. Dokument ten zawierał również definicje mierników realizacji próby badawczej, w tym wskaźniki ogólnego poziomu jej realizacji (*response rate*), gotowości respondentów do udziału w badaniu (*cooperation rate*), skali odmów udziału w badaniu (*refusal rate*) oraz frakcji jednostek, z którymi udało się nawiązać kontakt[191] (*contact rate*).

Schemat klasyfikacyjny jednostek próby odpowiadający kodyfikacji AAPOR z 2011 roku przedstawiony został na rycinie V.1. Chociaż propozycja ta poświęcona jest w sposób szczególny próbom opartym na adresowych operatach gospodarstw domowych (z dwustopniowym doborem jednostek), to jednak rozciągnięcie jej na próbę adresową budynków mieszkalnych, czy też nawet na próbę imienną, okazuje się niezwykle proste. Dla analiz podejmowanych w tej pracy (opartych w znacznej mierze na danych metodologicznych Europejskiego Sondażu Społecznego) nie bez znaczenia pozostaje też fakt, iż stosowane w ESS standardy postrealizacyjnej klasyfikacji jednostek dobranych do próby, a także procedury wyznaczania wielkości wskaźników jej realizacji, odwołują się wprost do norm zgodnych z propozycją AAPOR[192] (por. Stoop i in. 2010: 12, 61, 70–74).

Odnosząc się do zaprezentowanego schematu, warto rozpocząć od wskazania, iż zbiór respondentów, czyli osób, które zgodziły się na udział w badaniu i odpowiedziały na zadane im pytania, można podzielić na dwie subkategorie:

---

[191] Niezależnie od tego, czy następnie wywiad został zrealizowany, czy też nie.

[192] Wykorzystywany w ESS wskaźnik realizacji próby wyznaczany jest jako stosunek liczby respondentów do sumy elementów w dwóch zbiorach, to znaczy jednostek niedostępnych oraz jednostek o nieustalonym statusie (por. *ESS5: Specification for participating countries* 2009: 15–16). Taki sposób ustalania *response rate* odpowiada miernikowi *RR2* w klasyfikacji AAPOR (por. *Standard Definitions: Final Disposition of Case Codes and Outcome Rates for Surveys. 7th edition. AAPOR* 2011: 44). Należy jedynie zgłosić zastrzeżenie, iż w wielu krajach uczestniczących w badaniach ESS nie stosuje się prób prostych, ale schematy wymagające ważenia rekompensującego nierówne prawdopodobieństwa doboru jednostek populacji do prób badawczych, a zatem wydaje się, że bardziej zasadne byłoby jednak wykorzystywanie ważonych mierników realizacji prób badawczych (por. Brick i in. 2009: 167–168).

Zbiór respondentów	(1.1) Wywiad przeprowadzony w całości (1.2) Wywiad przeprowadzony częściowo		
Zbiór jednostek niedostępnych należących do populacji docelowej	(2.1) Odmowa lub przerwanie wywiadu		(2.1.1) Odmowy udziału w badaniu (2.1.1.1) Odmowa na poziomie gospodarstwa (2.1.1.2) Odmowa uzyskana od respondenta (2.1.2) Wywiad przerwany
	(2.2) Nie nawiązano kontaktu z respondentem		(2.2.3) Brak możliwości wejścia do budynku (2.2.4) Nie zastano nikogo w mieszkaniu (2.2.5) Respondent wyjechał/jest nieosiągalny
	(2.3) Inne przypadki niedostępności		(2.3.1) Śmierć respondenta (2.3.2) Fizyczna / umysłowa niezdolność respondenta do udziału w badaniu (2.3.3) Problemy językowe w przeprowadzeniu wywiadu (2.3.3.1) Problemy językowe na poziomie gosp. (2.3.3.2) Problemy językowe respondenta (2.3.3.3) Nie można wysłać ankietera posługującego się językiem respondenta (2.3.4) Niedostępność z innych przyczyn
Zbiór jednostek niedostępnych o nieustalonym statusie przynależności do populacji docelowej	(3.1) Nie wiadomo, czy budynek jest zamieszkany (3.1.1) Nie podjęto próby kontaktu / nie wysłano ankietera (3.1.7) Ankieter nie był w stanie dotrzeć / niebezpieczne miejsce (3.1.8) Nie udało się zlokalizować adresu (3.2) Budynek zamieszkany / nie wiadomo, czy znajduje się osoba z populacji (3.9) Inne przyczyny nieustalenia statusu respondenta		
Zbiór jednostek nienależących do populacji docelowej oraz błędy operatu doboru próby	(4.1) Wylosowana jednostka nie należy do populacji (4.5) Budynek nie zawiera jednostek mieszkalnych (4.5.1) Biurowiec, budynek administracji (4.5.2) Instytucje na przykład więzienia, sanatoria (4.5.3) Obiekty zbiorowego zakwaterowania (4.6) Pustostany (4.6.1) Budynki niezamieszkane (4.6.2) Budynki wakacyjne / czasowo zamieszkane (4.6.3) Pustostany – inne przypadki (4.7) W gospodarstwie nie ma osób spełniających kryteria doboru (4.9) Inne przypadki kwalifikujące wylosowane jednostki poza populację		

Ryc. V.1. Postrealizacyjna klasyfikacja jednostek wylosowanych do próby badawczej – próba adresowa (budynków lub gospodarstw domowych) oraz próba imienna

Źródło: opracowanie własne na postawie: The American Association for Public Opinion Research. 2011. Standard Definitions: Final Disposition of Case Codes and Outcome Rates for Surveys. 7th edition. AAPOR; kody zgodne z klasyfikacją AAPOR; por. tab. 2. Final Disposition Codes for In-Person, Household Surveys, s. 58

(1) jednostek, z którymi wywiad został przeprowadzony w całości, oraz (2) tych osób, które odpowiedziały wyłącznie na część zadanych im pytań. W praktyce taka dychotomia okazuje się jednak problematyczna. Niezwykle rzadko zdarza się bowiem, by respondent udzielił odpowiedzi na wszystkie zadane mu pytania (por. Koch i in. 2009: 51–54)[193], a tym samym istnieje potrzeba ustalenia dodatkowych reguł pozwalających na klasyfikację każdego rozpoczętego wywiadu jako zrealizowanego w całości lub częściowo. Co oczywiste, nieliczne przypadki braków danych na poziomie pojedynczych pytań (*item nonresponse*) traktować będzie się jako wywiad przeprowadzony w całości, chodzi jednak o to, aby ustalić, jaka frakcja braków danych w ogóle (lub w odniesieniu do jakich pytań) sprawi, iż wywiad częściowy uznany zostanie przez badacza za niezrealizowany w całości (*unit nonresponse*) (por. Dilmann i in. 2002: 12). Co więcej, reguły takie wymagają wprowadzenia dystynkcji pomiędzy wywiadami zrealizowanymi w części, a tymi częściowymi, których niekompletność jest efektem przerwania wywiadu przez respondenta[194]. W tym drugim przypadku – o ile przerwanie nastąpiło we wczesnej fazie realizacji badania – elementy próby należy traktować jako jednostki niedostępne[195] (por. de Leeuw i in. 2003: 158) i zgodnie z klasyfikacją AAPOR przypisać do wspólnej kategorii z osobami odmawiającymi udziału w badaniu[196].

---

[193] Achim Koch wraz z Michaelem Blohmem (2009: 46–45) prześledzili skalę zjawiska braków odpowiedzi w pojedynczych pytaniach we wszystkich krajach uczestniczących w badaniu ESS-u. Przedmiotem analiz tych autorów było 75 pytań zadawanych w każdej z trzech pierwszych odsłon tego projektu (z wyłączeniem zmiennych społeczno-demograficznych). Wyniki tych studiów ukazują, że nieco ponad połowa respondentów odpowiedziała na wszystkie pytania (odpowiednio było to 54,3% wszystkich przeprowadzonych wywiadów w ESS1, 56,1% w ESS2 oraz 58,2% w ESS3). Zaobserwowano przy tym znaczne zróżnicowania międzykrajowe. Dla przykładu, ponad ¾ wszystkich respondentów w Norwegii odpowiedziało na wszystkie pytania, podczas gdy w Portugalii było to już mniej więcej 33%. Wskaźnik braków danych na poziomie pojedynczego pytania (dla każdego respondenta obliczany jako uśredniona liczba pytań z brakami) wskazuje też, że Polska plasuje się w grupie państw o wyższych wartościach współczynnika *item nonresponse*. Studia te ukazują zatem, iż definiowanie wywiadu przeprowadzonego w całości jako takiego, w którym respondent odpowiedział na wszystkie bez wyjątku zadane mu pytania, nie jest wcale najlepsze.

[194] Reguły określające klasyfikacje wywiadów jako zrealizowanych całościowo, częściowo lub przerwanych przez respondenta odnaleźć można w opracowaniu *AAPOR: Standard Definitions*... 2011: 20. W polskiej literaturze do tych standardów odwołuje się między innymi K. Grzeszkiewicz-Radulska (2009: 29).

[195] W artykule *Prevention and Treatment of Item Nonresponse* opublikowanym przez Edith D, de Leeuw, Joop Hox oraz Martin Huisman w 2003 roku odnaleźć można następującą sugestię: „jeżeli przerwanie wywiadu nastąpiło we wczesnej fazie jego realizacji i tylko kilka pytań zostało zadanych, to [przerwanie – P.J.] traktuje się jako wywiad niezrealizowany w całości. Jeżeli natomiast wywiad został przerwany pod koniec jego trwania i większość pytań została zadana, to pozostałe niezadane pytania są zazwyczaj traktowane jako przypadki *item nonresponse*" (de Leeuw i in. 2003: 158).

[196] W literaturze metodologicznej niezwykle trudno odnaleźć studia empiryczne dedykowane tej szczególnej sytuacji niezrealizowania wywiadu, jaką są jego przerwania. Co więcej, przypadki

Krótkiego komentarza wymaga także rozróżnienie na te jednostki niedostępne, z którymi nie udało się nawiązać kontaktu, choć ustalono, iż należą do badanej populacji, a także na te niedostępne osoby, do których nie udało się dotrzeć i ustalić statusu ich przynależności do populacji. W dokumentacji AAPOR odnaleźć można dokładne charakterystyki przypadków kwalifikujących wylosowane osoby do tych rozłącznych zbiorów jednostek (por. *AAPOR: Standard Definitions…* 2011: 21–23). Niezwykle ważne wydają się jednak praktyczne konsekwencje takiego podziału elementów próby. Otóż w drugiej z tych grup znajdą się jednostki należące, jak i też nienależące do badanej populacji, a ponieważ o takich osobach nie ma żadnej informacji, to w konsekwencji nie wiadomo też, jaki odsetek z nich rzeczywiście wchodzi (a jaki nie wchodzi) w zakres populacji[197]. A zatem to, w jaki sposób potraktuje się tę drugą kategorię jednostek, będzie miało przełożenie na wielkość frakcji osób niedostępnych w próbie badawczej oraz – w konsekwencji – na wartość wskaźników realizacji próby. Jednym ze sposobów postępowania w tym względzie jest traktowanie wszystkich jednostek o nieustalonym statusie tak, jakby należały do badanej populacji[198]. Wprawdzie podejście takie skutkuje mniejszymi wskaźnikami realizacji próby, to jest jednak bezpieczne w tym oczywiście znaczeniu, że w zbiorze jednostek niedostępnych nie zostanie pominięta żadna jednostka, która powinna się tam znaleźć, niezależnie od tego, że pewne osoby, klasyfikowane jako niedostępne, będą w rzeczywistości elementami spoza populacji. Przykładem takiego postępowania są badania Europejskiego Sondażu Społecznego (por. Stoop i in. 2010: 90).

Zestawiając ze sobą klasyfikację AAPOR oraz kodyfikację wykorzystywaną w ESS, trzeba jednak odnotować pewną istotną różnicę w obu systemach porządkujących jednostki próby. Otóż w standardach AAPOR prawie wszystkie przypadki niezrealizowania wywiadu z powodu śmierci wylosowanej osoby

---

takie rozważane są zazwyczaj w kontekście badań prowadzonych techniką wywiadu telefonicznego lub ankiety internetowej (por. na przykład Peytchev 2011: 33–47; Peytchev 2009: 74–97; Sakshaug i in. 2010: 907–933; Keeter i in. 2006: 759–779; Peytchev i in. 2006: 559–607), co oznacza, iż ustalenia w tym zakresie mają niewielkie przełożenie na praktykę realizacji wywiadów osobistych prowadzonych na próbach imiennych lub adresowych. Trudno zatem orzec, czy traktowanie osób przerywających wywiady jako kategorii jednostek podobnych do osób odmawiających udziału w badaniu ma silne podstawy merytoryczne. Z nielicznych studiów wiadomo, że osoby przerywające wywiad różnią się znacznie od respondentów (por. Keeter i in. 2006: 774–776) i są podobne w swych charakterystykach do jednostek niedostępnych w ogóle (por. Sakshaug i in. 2010: 920).

[197] Aby ograniczyć niepewność w tym względzie, w badaniu ESS rekomenduje się, aby odsetek jednostek „nieskontaktowanych" o nieustalonym statusie przynależności do populacji nie przekraczał 3 pp. (por. Stoop i in. 2010: 60; Stoop 2005: 51).

[198] Przykłady procedur pozwalających na szacowanie frakcji jednostek należących do populacji w grupie osób o nieustalonym statusie odnaleźć można w opracowaniu Bricka i in. (2009: 169) oraz Toma W. Smitha (2009).

przypisuje się do zbioru jednostek niedostępnych z uwagi na brak kontaktu, podczas gdy w ESS sytuacje takie kodyfikuje się już jako przypadki jednostek spoza populacji docelowej (por. Stoop i in. 2010: 63). Pomijając tę jedną różnicę, system postterenowej klasyfikacji jednostek próby stosowany przez AAPOR oraz ESS jest w zasadzie taki sam. Wracając zatem do analizy przypadków związanych ze śmiercią osoby wylosowanej do próby badawczej oraz tego, do której grupy taki przypadek należałoby przypisać, można łatwo przywołać argumenty przemawiające zarówno za standaryzacją AAPOR, jak i te, które świadczą na rzecz propozycji ESS-u. Zacząć warto od wskazania, iż za tą drugą przemawia przede wszystkim fakt, że zgony są jednym z dwóch podstawowych (poza narodzinami) czynników kształtujących liczebność populacji, a co za tym idzie, śmierć wylosowanej osoby – w sposób zupełnie oczywisty – wyłącza ją z zakresu populacji docelowej. Z drugiej jednak strony można wskazać, iż dobierając próbę z jakiegoś operatu (np. imiennego), określa się nie tylko cechy osób, które do populacji należą, ale też ramy czasowe, do których populację się odnosi. Innymi słowy, to czy dany przypadek zgonu potraktować należy jako niedostępność, czy też jako przypadek wykluczenia z populacji, zależy tylko i wyłącznie od przyjętych ograniczeń czasowych. Dla przykładu, jeżeli dobrana osoba żyła w momencie losowania próby, a zmarła przed wizytą ankietera, to można ją – zgodnie z propozycją AAPOR – przypisać do kategorii jednostek niedostępnych z uwagi na inne przyczyny niezrealizowania wywiadów (por. AAPOR*: Standard Definitions…* 2011: 21–22). Chociaż dyskusję taką łatwo sprowadzić do poziomu absurdu, to jednak jeszcze jeden argument przemawia na rzecz stosowania w tym względzie kodyfikacji AAPOR. Można bowiem zauważyć, że z zupełnie analogiczną sytuacją będzie się miało do czynienia w przypadku tych osób, które udzieliły wywiadu, lecz zmarły przed zakończeniem terenowej fazy badań. Istotnie, tych zmarłych respondentów nie będzie się wyłączać poza zakres populacji, badacze nie mają bowiem w zwyczaju sprawdzać, czy ich respondenci jeszcze żyją – chyba że takie sytuacje zidentyfikuje się w trakcie kontroli pracy ankieterów. Tym samym wszystkie przypadki osób zmarłych po realizacji wywiadu przyporządkowane zostaną do zbioru respondentów, pomimo iż nie należą one już do populacji. Na szczęście zgony osób dobranych do badania stanowią zazwyczaj tak nieznaczny odsetek całej wylosowanej próby, że traktowanie takich przypadków jako niedostępnych członków populacji, lub też przeciwnie, jako osób nienależących do populacji, nie będzie miało prawie żadnego przełożenia na wartości wskaźników realizacji próby oraz na jakość przeprowadzonego badania[199].

[199] Doświadczenia ESS-u w Polsce wskazują, iż z sytuacją wylosowania osoby zmarłej ankieterzy spotkali się w 15 przypadkach realizacji ESS1 i odpowiednio w 14 przypadkach w ESS2, 14 w ESS3, 8 w ESS4, 22 w ESS5 oraz 21 w ESS6.

* * *

Wydawać by się mogło, że dyskusja nad standardami postterenowej klasyfikacji osób wylosowanych do próby badawczej ma wyłącznie charakter księgowy. Tymczasem ustalenie przyczyn niedostępności jednostek ma przede wszystkim uzasadnienie teoretyczne i merytoryczne (por. Groves i in. 1995: 93–106). Nie chodzi już jednak wyłącznie o to, że zastosowanie różnych standardów kodyfikacji elementów próby uniemożliwia międzysurveyowe porównania wskaźników poziomu jej realizacji; dużo większe znaczenie ma to, że mechanizmy oddziałujące na gotowość jednostek do udziału w badaniu mają zgoła odrębną naturę od procesów warunkujących możliwość dotarcia do wylosowanych osób (por. Groves i in. 1998: 23–42). A zatem trudność w nawiązaniu kontaktu oraz niechęć do kooperacji z ankieterem są dwoma niezależnymi wymiarami porządkującymi w odmienny sposób elementy próby badawczej (por. Lynn i in. 2002: 146). Innymi słowy, traktowanie jednostek niedostępnych jako kategorii homogenicznej nie oddaje prawdziwego obrazu rzeczywistości badawczej. Wystarczy przy tym przeprowadzić proste porównanie rozkładów zmiennych społeczno-demograficznych w warstwach osób niedostępnych, aby wykazać występowanie znacznego zróżnicowania jednostek w obrębie kategorii osób nieprzebadanych. Tabela V.1. zawiera zestawienia rozkładu płci, wieku oraz typu i wielkości miejscowości zamieszkania wylosowanych osób (w warstwie respondentów oraz osób niedostępnych) w badaniach ESS5–PL z 2010 roku[200].

Przyglądając się danym zaprezentowanym w tabeli V.1., można zauważyć, że społeczno-demograficzna kompozycja warstw osób nieprzebadanych jest rzeczywiście dość znacznie zróżnicowana, a zatem jednostki niedostępne nie tworzą zbioru homogenicznego. Jest to szczególnie widoczne w odniesieniu do wieku wylosowanych osób. Respondenci odróżniali się znacząco od niedostępnych oraz – co ważniejsze – rozkład wieku był istotnie zróżnicowany pomiędzy warstwami nieprzebadanych. Wprawdzie w ramach studiów dotyczących drugiej fali ESS-u P.B. Sztabiński i in. (2007: 33) stwierdzili, że wiek nie różnicuje niedostępnych oraz respondentów, jednakże analizy tych autorów mają tę słabość interpretacyjną, że ograniczają się wyłącznie do części, a nie do wszystkich osób nieprzebadanych[201]. Co więcej, w studiach tych nie dokonano rozgrani-

[200] Zakres merytoryczny takich zestawień ograniczony jest dostępnością danych. Muszą one bowiem obejmować wszystkie, bez wyjątku, elementy wylosowanej próby. W zasadzie w takich zestawieniach wykorzystać można wyłącznie dane pochodzące z operatów doboru próby, dane spisowe oraz – w ograniczonym zakresie – informacje odnotowywane przez ankieterów.

[201] Respondentów porównywano z tymi niedostępnymi, którzy wypełnili i odesłali przekazaną im ankietę pocztową. Takie ankiety wysłane zostały do wszystkich 567 jednostek niedostępnych w ESS2-PL; otrzymano 204 zwroty (por. Sztabiński P.B. 2006: 10). Fakt, że takie osoby nie różnią się wiekiem od respondentów, nie oznacza, że w całym zbiorze osób nieprzebadanych różnice nie będą występować.

Tabela V.1. Rozkłady wybranych cech społeczno-demograficznych w warstwie respondentów oraz jednostek niedostępnych w badaniach ESS5–PL (ed. 2010)

		Zbiór jednostek niedostępnych		
	Zbiór respon-dentów	**Odmowa lub przerwanie wywiadu**	**Brak kontaktu**	**Inne przyczyny**
Płeć				
Mężczyzna	48,1 (847)	44,2 (204)	57,3 (121)	35,0 (14)
Kobieta	51,9 (914)	55,8 (258)	42,7 (90)	65,0 (26)
Test chi-kwadrat (zgodność rozkładów) 1. Respondenci *vs*. zbiór jednostek niedostępnych: $\chi^2 = 0{,}061$, $df = 1$, $p = 0{,}806$ 2. Odmowa lub przerwanie *vs*. brak kontaktu: $\chi^2 = 10{,}092$, $df = 1$, $p = 0{,}001$ 3. Odmowa lub przerwanie vs. inne przyczyny: $\chi^2 = 1{,}256$, $df = 1$, $p = 0{,}262$ 4. Brak kontaktu *vs*. inne przyczyny: $\chi^2 = 6{,}755$, $df=1$, $p = 0{,}009$				
Wiek				
15–24 lat	19,1 (336)	11,4 (53)	16,1 (34)	9,8 (4)
25–34 lat	17,9 (316)	19,0 (88)	24,2 (51)	2,4 (1)
35–49 lat	21,1 (372)	25,9 (120)	28,9 (61)	9,8 (4)
50–64 lat	25,7 (453)	25,9 (120)	22,3 (47)	24,4 (10)
65 lat i więcej	16,2 (285)	17,7 (82)	8,5 (18)	53,7 (22)
Test chi-kwadrat (zgodność rozkładów) 1. Respondenci *vs*. zbiór jednostek niedostępnych: $\chi^2 = 18{,}059$, $df = 4$, $p = 0{,}001$ 2. Odmowa lub przerwanie *vs*. brak kontaktu: $\chi^2 = 13{,}812$, $df = 4$, $p = 0{,}008$ 3. Odmowa lub przerwanie vs. inne przyczyny: $\chi^2 = 35{,}431$, $df = 4$, $p<0{,}001$ 4. Brak kontaktu *vs*. inne przyczyny: $\chi^2 = 59{,}246$, $df = 4$, $p<0{,}001$				
Typ i wielkość miejscowości				
Wieś	41,5 (732)	28,4 (132)	30,6 (64)	31,7 (13)
Miasto do 10 tys.	6,6 (116)	4,1 (19)	4,3 (9)	7,3 (3)
Miasto 10–19 tys.	6,8 (120)	7,3 (34)	4,8 (10)	2,4 (1)
Miasto 20–49 tys.	10,6 (187)	11,6 (54)	10,5 (22)	9,8 (4)
Miasto 50–99 tys.	7,4 (130)	11,0 (51)	7,7 (16)	4,9 (2)
Miasto 100–199 tys.	7,5 (133)	9,7 (45)	9,6 (20)	19,5 (8)
Miasto 200–499 tys.	8,9 (157)	11,2 (52)	13,9 (29)	12,2 (5)
Miasto od 500–999 tys.	6,2 (109)	9,9 (46)	14,4 (30)	4,9 (2)
Warszawa	4,4 (78)	6,7 (31)	4,3 (9)	7,3 (3)
Test chi-kwadrat (zgodność rozkładów) 1. Respondenci *vs*. zbiór jednostek niedostępnych: $\chi^2 = 55{,}736$, $df = 8$, $p<0{,}001$ 2. Odmowa lub przerwanie *vs*. brak kontaktu: $\chi^2 = 8{,}181$, $df=8$, $p = 0{,}416$ 3. Odmowa lub przerwanie vs. inne przyczyny: $\chi^2 = 8{,}311$, $df = 8$, $p = 0{,}404$ Brak kontaktu *vs*. inne przyczyny: $\chi^2 = 7{,}624$, $df = 8$, $p = 0{,}471$				

Źródło: Obliczenia własne na podstawie repozytorium danych ESS5-SDDF-PL. W kolejnych kolumnach tabeli przedstawione zostały rozkłady procentowe oraz – w nawiasach – rozkłady liczebności

czenia przyczyn niedostępności, zestawiając respondentów ze zbiorem jednostek niedostępnych w ogóle. Powracając do analizy danych ESS5-PL, można zauważyć, że chociaż w zbiorze respondentów oraz jednostek niedostępnych rozkład płci nie był istotnie zróżnicowany, to jednak w warstwach wyodrębnionych z uwagi na przyczynę nieprzeprowadzenia wywiadów rozkłady takie były już odmienne. Jedynie charakterystyki typu i wielkości miejscowości zamieszkania nie wyróżniały się szczególnie znaczącymi dysproporcjami międzywarstwowymi.

Niejednorodny charakter jednostek niedostępnych staje się jeszcze bardziej widoczny po przeprowadzeniu analiz identyfikujących czynniki warunkujące prawdopodobieństwo wystąpienia określonego rodzaju niedostępności. Wyniki takich studiów – oparte na modelu regresji logistycznej – przedstawione zostały w tabeli V.2.[202].

Przeprowadzone analizy ukazują niezwykle intersujące układy zależności pomiędzy zmiennymi społeczno-demograficznymi a szansami wystąpienia określonego typu niedostępności. Zacząć warto od ukazania tego, że płeć oddziałuje w sposób istotny na prawdopodobieństwo nawiązania kontaktu, a także niezrealizowania wywiadów z innych przyczyn, nie ma natomiast statystycznie istotnego przełożenia na szanse uzyskania odmowy. Ważne jest przede wszystkim to, że charakter oddziaływania płci w warstwach nierespondentów jest odmienny. Większe prawdopodobieństwo niedostępności spowodowanej brakiem kontaktu mają mężczyźni (iloraz szans wyniósł 1:1,478), podczas gdy kobiety odznaczają się już większą szansą na to, że przyczyną niezrealizowania wywiadu będzie odmowa (stosunek szans mężczyzn do szans kobiet wyniósł jak 1:0,867) lub też, że źródłem ich niedostępności będzie jakaś inna przyczyna (szanse mężczyzn w porównaniu do szans kobiet wyniosły jak 1:0,656). Wnioski te pozostają zgodne z ustaleniami innych badaczy. Wprawdzie J. Goyder (1987: 84) wątpił, czy płeć ma w ogóle jakieś znaczące przełożenie na szanse nawiązania kontaktu oraz na gotowość do udziału w badaniu, jednak w wielu innych opracowaniach wskazuje się już, iż znacznie większa trudność dotarcia do mężczyzn jest konsekwencją ich częstszego przebywania poza domem (por. Domański 2006: 38; Groves i in. 1998: 136) oraz większej aktywności na rynku pracy (por. Stoop i in. 2010: 119). Z drugiej strony potwierdza się jednak scep-

[202] Dla każdej charakterystyki społeczno-demograficznej (tj. dla zmiennych niezależnych), ustalone zostały tak zwane kategorie referencyjne, czyli takie wartości płci, wieku oraz typu i wielkości miejscowości, które stanowią punkt odniesienia w ocenie oraz interpretacji charakteru oddziaływania pozostałych kategorii zmiennych społeczno-demograficznych na szanse niedostępności jednostek. Dla przykładu, jeżeli oszacowanie współczynnika regresji dla jakiejś konkretnej wartości zmiennej społeczno-demograficznej (np. kategorii mężczyzn) jest większe od jedności, to taka kategoria ma relatywnie większe szanse na bycie jednostką niedostępną niż osoby ze zbioru referencyjnego (tzn. kobiety). Analogicznie, wartości mniejsze od jedności świadczą o relatywnie niższych szansach niedostępność osób z określonej kategorii jednostek próby.

Tabela V.2. Regresja logistyczna ilorazu szans (1) niedostępności z uwagi na brak kontaktu (*n* = 220) relatywnie do szans nawiązania kontaktu (*n* = 2271), (2) niedostępności z uwagi na odmowę lub przerwanie wywiadu (*n* = 463) oraz (3) niedostępności z innych przyczyn (*n* = 40) relatywnie do szans realizacji wywiadu (*n* = 1768) – analizy w oparciu o repozytorium ESS5-PL (ed. 2010)

	Brak kontaktu (1)	**Odmowy lub przerwane wywiady (2)**	**Inny powód niedostępności (3)**
Płeć			
Mężczyzna	***1,478*** **	0,867	***0,656*** *
Kobieta (ref.)	–	–	–
Wiek			
15–24 lat	***1,797*** *	***0,559*** **	***0,190*** ***
25–34 lat	***2,518*** ***	0,979	***0,042*** **
35–49 lat	***2,600*** ***	1,165	***0,107*** ***
50–64 lat	***1,661*** *	0,921	***0,301*** **
65 lat i więcej (ref.)	–	–	–
Typ i wielkość miejscowości			
Wieś	0,851	***0,450*** **	0,432
Miasto do 10 tys.	0,772	***0,411*** **	0,609
Miasto 10–19 tys.	0,768	0,703	0,193
Miasto 20–49 tys.	1,068	0,720	0,528
Miasto 50–99 tys.	1,070	0,988	0,363
Miasto 100–199 tys.	1,282	0,853	1,452
Miasto 200–499 tys.	1,637	0,823	0,863
Miasto 500–999 tys.	***2,360*** **	1,033	0,362
Warszawa (ref.)	–	–	–
Test Hosmera–Lemeshowa	$\chi^2 = 3{,}306$, *df = 8*	$\chi^2 = 8{,}158$, *df = 8*	$\chi^2 = 6{,}569$, *df = 8*

Źródło: obliczenia własne na podstawie repozytorium danych ESS5
*** Istotność na poziomie p<0,001; ** Istotność na poziomie p<0,05; * Istotność na poziomie p<0,1.

tycyzm Goydera odnośnie nieznacznego wpływu płci na szanse kooperacji (por. Voogt i in. 2002: 331; Groves. i in. 1998: 136). A zatem odnotowywane w literaturze metodologicznej wyższe odsetki realizacji próby w warstwie kobiet nie wynikają z ich znacząco większej gotowości do udziału w badaniu, ale z tego, że charakteryzują się one istotnie większą dostępnością (por. Stoop 2005: 70).

Największe prawdopodobieństwo nawiązania kontaktu charakteryzuje również respondentów z kategorii wiekowej osób powyżej 65. roku życia[203]. Analizy ukazują jednak, że osoby najstarsze mają większą skłonność do odmawiania udziału w badaniu. Na wprost proporcjonalne przełożenie wieku wy-

[203] W pozostałych grupach wiekowych iloraz szans nieskontaktowania relatywnie do takich szans w warstwie osób 65+ jest większy od jedności.

losowanych jednostek na szanse dotarcia, a także odwrotnie proporcjonalne oddziaływane na gotowość do udziału w badaniu, zwracano uwagę już wielokrotnie (por. Goyder 1987: 84; Stoop i in. 2010: 119). Szczególnie utrudniona – z uwagi na większe wskaźniki odmów – okazuje się przy tym realizacja badań ze starszymi kobietami zamieszkującymi w jednoosobowych gospodarstwach domowych (por. Stoop 2004: 35). Jest to zapewne w jakimś stopniu pochodną większego poziomu obaw związanych z wpuszczeniem kogoś obcego do domu, bowiem, jak ukazuje Groves i in. (1998: 134), efekt wieku zanika, gdy – jako zmienną kontrolną – uwzględni się wielkość gospodarstwa domowego. Z kolei zaobserwowana w studiach metodologicznych mniejsza dostępność osób z najmłodszych kategorii wiekowych (por. Domański 2006: 42; Voogt i in. 2002: 331) wynika przede wszystkim z ich częstszego przebywania poza domem (por. Stoop 2005: 69). Analizy danych ESS5-PL ukazują też, że w najstarszej kategorii wiekowej znacznie większe jest prawdopodobieństwo niedostępności wynikającej z innych powodów niż niemożliwość nawiązania kontaktu lub odmowa udziału w badaniu. Prawidłowość ta wydaje się jednak przede wszystkim konsekwencją sposobu definiowania zbioru jednostek niedostępnych z innych przyczyn. Otóż zdecydowaną większość sytuacji kryjących się za takim typem niedostępności stanowią przypadki niezrealizowania wywiadów z powodu choroby oraz fizycznej lub psychicznej niezdolności wylosowanej osoby do udziału w badaniu, a zatem, w sposób zupełnie oczywisty, wiek musi być tutaj czynnikiem różnicującym szanse wystąpienia tego źródła niedostępności. Doskonałym tego potwierdzeniem są analizy zaprezentowane przez R. Grovesa oraz L. Lyberga w monografii *Nonresponse in Household Interview Surveys* z których wynika, że „osoby starsze […] są nieproporcjonalnie częściej klasyfikowane do grupy 'innych' przypadków niezrealizowania wywiadów z uwagi na występowanie problemów zdrowotnych uniemożliwiających ich udział w badaniu" (por. Groves i in. 1998: 133). Do podobnych wniosków prowadzą także studia z holenderskiej edycji Generalnego Sondażu Wyborczego (ed. 1998), z których wynika, że osoby niedostępne z powodu niezdolność do udziału w badaniu różnicuje tylko wiek oraz płeć, przy czym prawie połowę wszystkich takich jednostek stanowią osoby 65+ oraz kobiety (por. Voogt i in. 2002: 331).

Można wreszcie wskazać, że najmniejsze szanse na nawiązanie kontaktu odnotowano w miastach o liczbie mieszkańców od 500 do 999 tys. osób[204]. Z kolei najmniejsze prawdopodobieństwo wystąpienia odmowy udziału w badaniu charakteryzuje osoby z obszarów wiejskich oraz z małych miasteczek

[204] Szansa na to, że z mieszkańcami dużych miast nie uda się w ogóle nawiązać kontaktu, jest ponad dwukrotnie wyższa niż szansa przypisana mieszkańcom Warszawy. Sytuacja ta jest trudna do wytłumaczenia. Zgodnie z ustaleniami literaturowymi należałoby bowiem oczekiwać, że szanse nawiązania kontaktu będą najmniejsze w największych miastach. Być może odnotowana w Warszawie mniejsza liczba przypadków niedostępności spowodowanej brakiem kontaktu wynika z koncentracji jakichś działań badawczych zwiększających szanse dotarcia do respondentów z Warszawy. Nie ma jednak dostępu do danych potwierdzających to przypuszczenie.

(w obu przypadkach szansa odmowy stanowi nieco ponad 0,4 szansy przypisanej kategorii referencyjnej, tj. mieszkańcom Warszawy). Niezwykle pouczającą egzemplifikacją problemów, jakie sprawia prowadzenie badań w dużych miastach, są wnioski sformułowane przez K. Grzeszkiewicz-Radulską (2009). Wykorzystując dane z ogólnopolskich badań opinii publicznej zrealizowanych przez CBOS w latach 1993–2003, autorka ta dokonała porównania czterech charakterystyk terenowej fazy badań (wskaźnika realizacji próby, nawiązania kontaktu, odmów oraz kooperacji[205]), zestawiając je z typem oraz wielkością miejscowości. Z porównań tych wynika, że w miastach powyżej 500 tys. mieszkańców (inaczej niż na wsiach oraz w małych miastach) wartości współczynników realizacji próby oraz kooperacji kształtowały się na najniższych poziomach, z kolei odsetki odmów udziału w badaniu były największe. Nieco mniejszym poziomem międzywarstwowego zróżnicowania charakteryzował się współczynnik odmów, choć i tak wartości uzyskiwane dla miast pow. 500 tys. mieszkańców kształtowały się na poziomie najniższym, a dla wsi oraz małych miasteczek na najwyższym (por. Grzeszkiewicz-Radulska 2009: 166–180). Opisane zależności wydają się mieć jednak charakter uniwersalny i wskazywane były w wielu przywoływanych wcześniej opracowaniach metodologicznych, zarówno w literaturze polskiej (por. Lutyńska 1989: 222; Domański 2006: 38, 42), jak też zagranicznej (por. na przykład Goyder 1987: 40; Goyder i in. 1992: 39–48; Groves i in. 1998: 87, 176; Stoop 2004: 43; Stoop 2005: 69; Johnston 2006: 293, 300).

Podsumowując tę część rozważań, należy przypomnieć, że błędy systematyczne wynikające z niezrealizowania części wywiadów są cechą indywidualną każdej zmiennej z osobna. Z faktu, że nie zaobserwowano różnic w rozkładach cech społeczno-demograficznych w zbiorze respondentów oraz w warstwie (lub warstwach) osób nieprzebadanych, nie można jeszcze wnioskować, że inne zmienne nie będą w ogóle obarczone błędami braku danych. Podobnie zresztą, nawet istotne dysproporcje pomiędzy osobami przebadanymi a niedostępnymi na przykład w charakterystykach płci, wieku, typu i wielkości miejscowości, czy też innych zmiennych „metryczkowych", nie muszą wcale świadczyć o systematycznym błędzie oszacowania statystyk substantywnych. Doskonałym potwierdzeniem komplikacji, na jakie narażone jest wnioskowanie o błędzie niepełnej realizacji próby poprzez analizę podobieństwa rozkładów cech społeczno-demograficznych w postbadawczych warstwach jednostek próby sondażowej, są rozważania R.J.J. Vogta oraz H. van Kempin z przywołanego już wcześniej artykułu poświęconego wynikom holenderskich badań nad postawami wyborczymi. W części wstępnej tego opracowania przeczytać można bowiem, że:

> ważne jest, aby uświadomić sobie, iż różnice pomiędzy respondentami i jednostkami niedostępnymi w zakresie zmiennych społeczno-demograficznych nie oznaczają automatycznie, iż te dwie grupy różnią się wartościami innych

[205] Wskaźnik kooperacji podaje informację o odsetku jednostek, z którymi po nawiązaniu kontaktu udało się przeprowadzić wywiad.

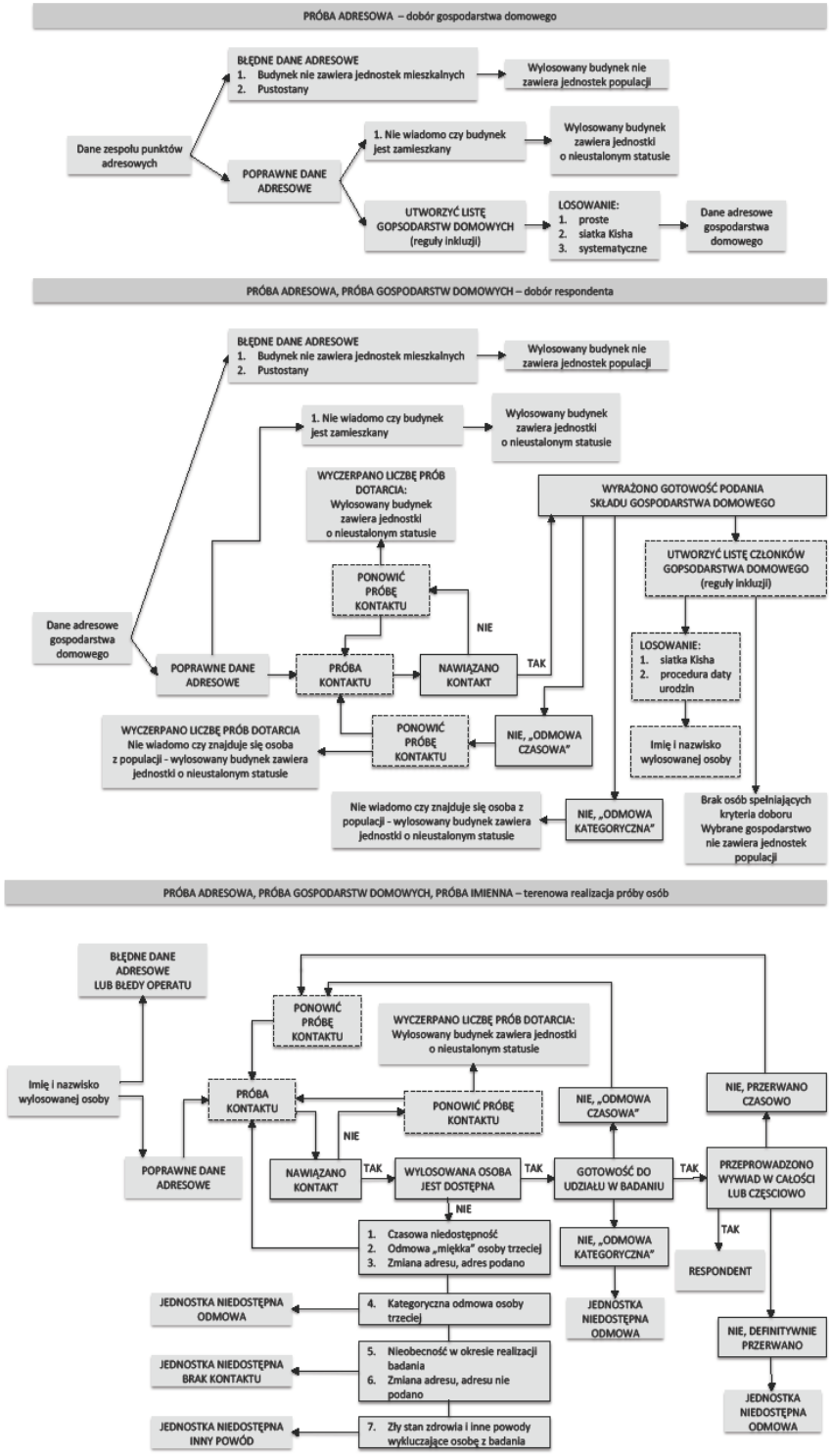
> zmiennych. Takich różnic można spodziewać się tylko wówczas, gdy istnieje zależność pomiędzy wartościami tych zmiennych a charakterystykami społeczno-demograficznymi. Jednak, tak jak obecność błędu w odniesieniu do zmiennych społeczno-demograficznych nie oznacza od razu wypaczenia kluczowych zmiennych, tak brak błędu nie oznacza też, iż kluczowe zmienne będą wolne od błędu systematycznego. (Voogt i in. 2002: 325–326)

Innymi słowy, studia nad konsekwencjami niepełnej realizacji próby badawczej wymagają przede wszystkim rozpoznania mechanizmów niedostępności jednostek, a także określenia czynników kształtujących zarówno szanse na dotarcie i nawiązanie kontaktu z wylosowanymi osobami, jak również prawdopodobieństwo ich udziału w badaniu. Zrozumienie takich mechanizmów ma znaczenie zupełnie fundamentalne dla badań sondażowych, umożliwia bowiem podjęcie skutecznych działań zmierzających do zminimalizowania ryzyka błędu braku danych oraz do ograniczenia innych negatywnych skutków niepełnej realizacji próby. Problematyka ta będzie przedmiotem analiz w dalszej części rozdziału. Zanim jednak to nastąpi, rozważone będą kwestie o nieco innym charakterze, a mianowicie: na ile decyzja o wyborze określonego typu operatu doboru próby przekłada się na wzorce jej terenowej realizacji. Na niektóre aspekty tych zagadnień zwracano uwagę w rozdziale III, rozważając błędy pokrycia populacji wynikające z wewnątrzzespołowej selekcji jednostek indywidualnych w ramach adresowych operatów gospodarstw domowych oraz budynków mieszkalnych (por. sekcja III.5.). Na przykładzie danych Europejskiego Sondażu Społecznego sprawdzone będzie teraz, czy wykorzystanie operatów adresowych oraz imiennych różnicuje postbadawczą strukturę zbioru respondentów oraz jednostek niedostępnych. Wydaje się bowiem, że typ operatu nie powinien pozostawać bez wpływu na wzorzec realizacji próby.

V.2. Schematy terenowej realizacji prób adresowych oraz imiennych

W literaturze metodologicznej niewiele jest studiów, w których relacje zachodzące pomiędzy typem operatu oraz wzorcem terenowej realizacji sondażowej próby badawczej byłyby przedmiotem systematycznego namysłu. W większości opracowań uwaga skupia się raczej na przełożeniu – charakterystycznego dla adresowych prób osób – procesu wewnątrzzespołowej selekcji jednostek na błędy niepełnego (lub nadmiarowego) pokrycia populacji docelowej, mniej natomiast koncentruje się na tym, czy wykorzystanie rejestrów budynków mieszkalnych lub wykazu gospodarstw domowych przekłada się w jakiś odmienny sposób na wskaźniki realizacji próby oraz postbadawczą strukturę zbioru respondentów i osób nieprzebadanych.

Na niektóre właściwości operatów doboru próby różnicujące dwa główne wymiary niedostępności jednostek (tj. brak kontaktu oraz niechęć do udziału w badaniu) zwrócili ostatnio uwagę Stoop i in. (2010) w monografii referującej



Ryc. V.2. Schemat realizacji prób adresowych oraz imiennych.

Źródło: opracowanie własne

doświadczenia z realizacji trzech pierwszych odsłon Europejskiego Sondażu Społecznego. Zanim przeanalizowane zostaną pewne problemy sygnalizowane przez autorów tej pracy, należy w pierwszej kolejności wskazać, że tym, co decyduje tak naprawdę o odmiennym charakterze terenowej realizacji prób adresowych oraz imiennych, jest poziom zagregowania jednostek populacji (inny w operatach adresowych, inny zaś w rejestrach imiennych). W próbach dobieranych z operatów imiennych ankieterzy otrzymują bowiem wykazy konkretnych osób, z którymi usiłują następnie przeprowadzić wywiady (jednostki próby są zatem z góry określone), podczas gdy w próbach adresowych przekazywana jest im wyłącznie informacja o gospodarstwach domowych (z których następnie dobierają osoby do wywiadów) lub też o budynkach mieszkalnych (w obrębie których przeprowadzają najpierw selekcję odpowiedniej liczby gospodarstw domowych, a następnie – w ramach tych gospodarstw – wybierają osoby do badania). Wydaje się, że to właśnie te etapy losowania wielostopniowego – będące immanentną cechą adresowych prób osób – decydują o odmiennej (w porównaniu z próbami imiennymi) strukturze zbioru respondentów oraz jednostek nieprzebadanych (por. Stoop i in. 2010: 56).

W schemacie terenowej realizacji sondażowej próby badawczej można wyodrębnić trzy główne etapy działań. Nie wszystkie muszą się pojawić, zależy to od charakteru danych zawartych w operacie doboru próby. Da przykładu, pierwszy z takich etapów występuje wyłącznie w adresowych próbach budynków mieszkalnych (ulic, kodów pocztowych itd.) i obejmuje wszystkie przedsięwzięcia zmierzające w pierwszej kolejności do spisania gospodarstw domowych znajdujących się pod wylosowanym adresem, a następnie do selekcji odpowiedniej liczby gospodarstw. Drugi etap – występujący w adresowych próbach budynków mieszkalnych oraz w próbach gospodarstw domowych – obejmuje z kolei działania zmierzające do spisania osób zamieszkujących wylosowane gospodarstwa, a następnie – w każdym mieszkaniu – do selekcji jednej osoby. Z kolei trzeci etap związany jest już z właściwą fazą badań terenowych i pojawia się niezależnie od typu operatu. Składają się na nie te wszystkie działania, których celem jest realizacja wywiadów z jednostkami wylosowanymi do próby wprost z operatu imiennego, bądź też pośrednio z operatu adresowego.

Oczywiście etap pierwszy nie musi się wcale zakończyć doborem gospodarstwa domowego, podobnie jak etap drugi niekoniecznie skutkuje wewnątrzzespołową selekcją respondentów, a etap trzeci nie zawsze kończy się realizacją wywiadów. Warto zatem mieć na względzie, że w adresowych próbach budynków mieszkalnych niedostępność jednostek może się pojawić, zanim jeszcze ustalone zostaną gospodarstwa, w których należałoby przeprowadzić selekcję respondentów. Sytuacja taka może wynikać z przekazania ankieterom (przeprowadzającym spis mieszkań znajdujących się pod wylosowanym adresem) nieprecyzyjnych danych uniemożliwiających im lokalizację niektórych budynków, może też być konsekwencją wylosowania budynków niezamieszkanych. Oba przypadki przekładają się w odmienny sposób na wskaźniki realizacji

Tabela V.3. Oczekiwane wzorce terenowej realizacji próby z uwagi na typ operatu

Kategoria niedostępności	Próba imienna	Próba gospodarstw domowych	Próba adresowa budynków
(1) Brak kontaktu	1. Wyższe niż w próbach gospodarstw oraz budynków wskaźniki braku kontaktu – możliwe jest wylosowanie osób niezamieszkujących już pod wskazanym w operacie adresem (takich, z którymi nie uda się następnie nawiązać kontaktu w nowym miejscu zamieszkania).	1. Oczekiwać należy niższych niż w próbach imiennych wskaźników braku kontaktu – wewnątrzzespołowa selekcja obejmuje wyłącznie osoby zamieszkałe w gospodarstwie.	1. Nieco niższe niż w próbach gospodarstw domowych wskaźniki braku kontaktu – niezamieszkane gospodarstwa pomija się zazwyczaj w procesie selekcji mieszkań.
(2) Odmowa udziału w badaniu	1. Oczekiwać można wyższych wskaźników kooperacji niż w próbach adresowych. 2. Nie występują charakterystyczne dla prób adresowych przypadki odmów podania składu gospodarstwa domowego. 2. Pozytywny wpływ na wskaźnik kooperacji może mieć personalizacja listów zapowiednich.	1. Prawdopodobnie wyższe wskaźniki niezrealizowania wywiadu z powodu odmowy podania składu gospodarstwa domowego. 2. Możliwe negatywne przełożenie bezosobowego charakteru listu zapowiedniego na gotowość do udziału w badaniu.	1. Porównywalne z próbami gospodarstw domowych odsetki odmów podania składu gospodarstwa domowego. 2. Z uwagi na utrudnioną możliwości przesłania listów zapowiednich w ogóle oczekiwać można wyższych, niż w próbach imiennych oraz próbach gospodarstw domowych, wskaźników odmów.
(3) Inne powody niezrealizowania wywiadów	1. Oczekiwać można większej liczby jednostek o nieustalonym statusie przynależności do populacji – jednostką obserwacji jest konkretna osoba, niezamieszkane przez nikogo gospodarstwo domowe nie oznacza, że przypisany do niej element próby jest jednostką nienależącą do populacji.	1. Mieszkania, w których nikt nie przebywa, traktuje się często jako obiekty niezawierające jednostek populacji. 2. Prawdopodobnie mniejsza, niż w próbach imiennych, liczba jednostek niedostępnych z uwagi na chorobę lub niezdolność psychofizyczną do udziału w badaniu – może pojawić się tendencja do pominięcia takich osób w spisie domowników.	1. Budynki niezamieszkane traktuje się jako obiekty niezawierające jednostek populacji. 2. Porównywalne z próbami gospodarstw domowych odsetki iedostępności z innych powodów.

Źródło: opracowanie własne

próby. Otóż te wszystkie osoby, które prawdopodobnie dobrałoby się do próby, gdyby tylko (istniejący) adres udało się zlokalizować, należałoby przypisać do zbioru jednostek o nieustalonym statusie przynależności do populacji docelowej (przypadki takie obniżają więc wartość wskaźnika *response rate*), podczas gdy wylosowanie adresów niezamieszkanych powinno się już traktować taksamo jak sytuację doboru jednostek nienależących do badanej populacji (nie ma to wpływu na wskaźnik realizacji próby). Z punktu widzenia oceny dostępności jednostek oraz poziomu ich gotowości do kooperacji ważne jest to, że na

tym wstępnym etapie realizacji badania nie podejmuje się jeszcze próby nawiązania kontaktu z przedstawicielami wylosowanych gospodarstw (spisywanie oraz selekcję mieszkań oddziela się przeważnie od fazy wewnątrzzespołowego doboru respondentów – por. Stoop i in. 2010: 56), co w praktyce oznacza, iż ten – charakterystyczny dla adresowych prób budynków mieszkalnych – etap realizacji próby nie powinien mieć szczególnego przełożenia na mierniki dostępności oraz wskaźniki gotowości jednostek próby do udziału w badaniu[206].

W próbach imiennych, inaczej niż w próbach adresowych, oczekiwać należałoby nieco większych wskaźników niedostępności jednostek spowodowanej brakiem kontaktu[207]. Jest to prostą konsekwencją tego, że wszystkie przypadki niezrealizowania wywiadów na skutek zmiany adresu miejsca zamieszkania kwalifikować będą wylosowane osoby do zbioru jednostek niedostępnych z uwagi na brak kontaktu (o ile oczywiście nie uda się do nich dotrzeć w nowym miejscu zamieszkania), podczas gdy w adresowych próbach budynków mieszkalnych oraz w próbach gospodarstw domowych sytuacje takie nie będą w ogóle miały miejsca. W operatach tego typu wewnątrzzespołowa selekcja obejmuje bowiem wyłącznie te osoby, które zamieszkują pod wskazanym adresem (por. Stoop i in. 2010: 131). Z drugiej strony, wyłączając ten szczególny przypadek niedostępności spowodowanej zmianą miejsca zamieszkania, należy przypuszczać, że typ operatu nie będzie już różnicował prawdopodobieństwa tego, iż w trakcie podejmowanej przez ankietera próby nawiązania kontaktu wylosowana osoba będzie obecna w mieszkaniu. Wprawdzie w rejestrach adresowych dobór jednostki zależy od tego, czy wcześniej udało się w ogóle nawiązać kontakt z jakimś przedstawicielem zamieszkiwanego przez nią gospodarstwa, nie oznacza to jednak, że osoba wylosowana do próby przebywała w tym czasie w mieszkaniu.

O ile typ operatu nie powinien mieć rzeczywiście – poza nielicznymi wyjątkami – znaczącego przełożenia na prawdopodobieństwo nawiązania kontaktu, o tyle należy już się spodziewać istotnego zróżnicowania wskaźników niedostępności spowodowanej odmową udziału w badaniu. Po pierwsze, można wskazać, że głównym czynnikiem decydującym o gotowości jednostek próby do udziału w badaniu jest sam proces wewnątrzzespołowej selekcji respondentów. Wyłączając przypadki doboru jednoosobowych gospodarstw domowych oraz sytuacje, w których osoba udzielająca informacji o członkach gospodarstwa jest też tą, z którą następnie prowadzony jest wywiad, w próbach adresowych ankieterzy muszą zazwyczaj przekonać jedną osobę do przekazania informacji

---

[206] Innymi słowy, wzorce terenowej realizacji prób budynków mieszkalnych oraz gospodarstw domowych powinny być do siebie bardzo podobne.

[207] Poprzez nawiązanie kontaktu rozumieć należy dotarcie do tej konkretnej osoby, która została wylosowana do próby, nie zaś – tak jak przyjmuje się często w analizach dotyczących prób adresowych (por. Stoop i in. 2010: 323; Groves 1989: 137) – dotarcie do którejkolwiek z osób zamieszkujących wylosowane gospodarstwo.

potrzebnych do przeprowadzenia selekcji respondenta, a następnie skłonić inną osobę do udziału w badaniu. Stanowi to – rzecz jasna – dodatkową barierę w realizacji prób adresowych oraz skutkuje większym odsetkiem odmów udziału w badaniu (por. Stoop i in. 2010: 15, 131). Po drugie, dość istotną niedogodnością związaną z próbami budynków oraz gospodarstw domowych jest też utrudniona, lub wręcz niemożliwa, możliwość skierowania – bezpośrednio do osób, z którymi zamierza się zrealizować wywiady – listów zapowiednich informujących o celach badania oraz spodziewanej wizycie ankietera (por. Stoop i in. 2010: 21). Co prawda niektórzy badacze powątpiewają w większą skuteczność spersonalizowanego charakteru takiego listu w próbach imiennych (por. Luiten i in. 2011: 11–20), wskazując nawet na potencjalnie negatywne oddziaływanie, jakie może to mieć dla decyzji o kooperacji z ankieterem (por. Sztabiński P.B. 2011: 122–123), jednakże, mimo tych zastrzeżeń, w większości opracowań podkreśla się wysoką przydatność listów zapowiednich w osiąganiu wyższych wskaźników realizacji próby (por. von der Lippe i in. 2011: 103–116; Biemer i in. 2003: 109–110; Groves i in. 1998: 276–281). Ponieważ w próbach budynków mieszkalnych wysyłanie listów informujących o badaniu jest w ogóle utrudnione (z operatu losuje się przecież zbiór wielu mieszkań, a nie konkretne gospodarstwa lub osoby), to w konsekwencji wskaźniki kooperacji w próbach tego typu mogą być niższe niż te, które uzyskuje się w próbach gospodarstw domowych oraz w próbach imiennych (por. Stoop i in. 2010: 131).

Naturalnie nie tylko typy operatów, ale również będące ich częścią procedury wewnątrzzespołowej selekcji jednostek mogą w odmienny sposób wpływać na wzorce terenowej realizacji próby. Wystarczy przypomnieć ustalenia z rozdziału III, w którym wskazywano, iż probabilistyczne metody takiego doboru prowadzą znacznie częściej – niż ma to miejsce w procedurach quasi-losowych – do odmów udziału w badaniu jeszcze przed tym, zanim przeprowadzona zostanie selekcja respondenta docelowego. Większość z metod probabilistycznych, w tym także doskonale znana siatka Kisha, wymaga wcześniejszego sporządzenia wykazu osób zamieszkujących wylosowane gospodarstwa, co może być uznane za zbyt daleko posuniętą ingerencję w prywatność domowników i w konsekwencji prowadzić do niższych wskaźników kooperacji. Ponieważ w metodach quasi-losowych (w tym w procedurze daty urodzin) rezygnuje się ze spisywania osób zamieszkujących dobrane gospodarstwa, to w efekcie uzyskuje się też zazwyczaj większe wskaźniki kooperacji (por. Gaziano 2005: 124–157).

Dla zapewnienia reprezentatywnego charakteru sondażowej próby badawczej najważniejsze jest jednak to, aby losowanie jednostek przebiegało całkowicie niezależne od (1) badacza, (2) ankieterów przeprowadzających selekcję, (3) domowników udzielających informacji potrzebnych do losowania jednostek, (4) jak też od dobranych osób. O ile w próbach losowanych z operatów imiennych taka selekcja pozostaje zazwyczaj zobiektywizowana (badacz, ankieter,

jak i też wylosowana osoba nie mają wpływu na to, która z jednostek populacji zostanie do niej wybrana), o tyle w próbach adresowych trzeba już postępować w taki sposób, aby dobór konkretnych jednostek był zobiektywizowany, to znaczy by badacz, ankieter, domownik wylosowanego gospodarstwa oraz potencjalny respondent, nie mieli żadnego wpływu na to, która z jednostek populacji zostanie wylosowana, która zaś zostanie pominięta. A zatem to, czy w losowaniu wielostopniowym prowadzonym z operatów adresowych taki zobiektywizowany charakter doboru uda się zachować, czy też nie, nie jest wcale pewne. Wystarczy przypomnieć przywoływane w rozdziale III wyniki studiów metodologicznych poświęconych błędom popełnianym w trakcie wewnątrzzespołowej selekcji jednostek próby z operatów adresowych, by ukazać, że w wielu przypadkach proces takiego doboru przebiegał w sposób nieprawidłowy i to zarówno z uwagi na nieintencjonalne, jak i celowe działania ankieterów, domowników wylosowanych gospodarstw, czy też wreszcie potencjalnych respondentów (por. Lavrakas i in. 2000: 890–895).

W realizacji adresowych prób budynków mieszkalnych niewłaściwe są zatem takie strategie doboru, które dają ankieterom swobodę w decydowaniu o wyborze pewnych gospodarstw domowych oraz pomijaniu innych. Zupełnie niedopuszczalne jest, aby losowane były tylko takie gospodarstwa, w których ktoś jest obecny, pomijane zaś te, których mieszkańcy przebywają poza domem. Podobnie rzecz ma się z wewnątrzzespołową selekcją jednostek próby. Wybór respondenta nie może być uwarunkowany tym, który z domowników przebywa w tym czasie w mieszkaniu. Nie może być też tak, że losuje się innego członka gospodarstwa domowego niż ten, który zgodnie z prawidłowo przeprowadzoną selekcją powinien znaleźć się w próbie. Takie odstępstwa od ustalonych reguł losowego lub quasi-losowego doboru gospodarstw oraz osób w ramach tych gospodarstw są w gruncie rzeczy tym samym, czym w próbach imiennych byłaby zamiana wylosowanej osoby na jakąś inną[208]. W skrajnych przypadkach może się bowiem zdarzyć tak, że ankieterzy będą starali się najpierw skłonić kogoś (niekoniecznie tego, który powinien się w próbie znaleźć) do udziału w badaniu, a następnie – wtórnie wobec tych działań – wskazywać będą, jakoby proces selekcji (gospodarstw oraz osób) prowadził właśnie do tej konkretnej osoby, którą wcześniej wskazali jako respondenta. Oczywiście znaczącą rolę w przeciwdziałaniu takim praktykom będzie miała kontrola jakości pracy ankieterów, o ile jednak w próbach imiennych dość łatwo jest ustalić, czy wywiad

[208] Co oczywiste, dobrane osoby mogą nie chcieć uczestniczyć w badaniu (w tym sensie mają wpływ na skład próby badawczej), jednak sam fakt wylosowania jakiejś jednostki pozostaje (ma pozostawać) poza jej wolą. Z podobną sytuacją będzie się miało do czynienia w losowaniu gospodarstw domowych. Nawet jeśli któreś z osób zamieszkujących dobrane gospodarstwa nie będą chciały następnie udzielić informacji potrzebnych ankieterowi do przeprowadzenia wewnątrzzespołowej selekcji respondentów, to wylosowanie takiego gospodarstwa powinno być od tego zupełnie niezależne.

został przeprowadzony z wylosowaną osobą, o tyle w próbach adresowych taka kontrola – choć możliwa – jest już niezwykle utrudniona[209] (por. Sawiński i in. 2005: 365).

Należy wobec tego zadać pytanie: w jakim stopniu – w próbach adresowych – postbadawcza struktura zbioru respondentów oraz jednostek niedostępnych wynika ze specyficznych charakterystyk operatów tego typu, a na ile uzyskiwane wzorce terenowej realizacji próby odzwierciedlają zniekształcenie procesu selekcji respondentów? Jeżeli bowiem w próbach adresowych miałoby rzeczywiście dochodzić do częstych i systematycznych nieprawidłowości w losowaniu jednostek (np. poprzez dobór osób częściej przebywających w domu oraz bardziej skłonnych do udziału w badaniu), to wskaźniki kontaktu powinny być – wbrew oczekiwaniom – dużo większe, natomiast odsetki odmów udziału w badaniu mniejsze niż te, które otrzymuje się w próbach imiennych. Byłoby to jednak dość niepokojące zjawisko i wskazywałoby na ograniczoną możliwość porównywania wyników badań z realizacji prób losowanych z operatów adresowych oraz imiennych. Przypuszczenia te warto zweryfikować empirycznie, odnosząc się do praktyki badawczej.

Na rycinie V.3. zamieszczone zostały wybrane charakterystyki realizacji prób badawczych z czwartej rundy Europejskiego Sondażu Społecznego. Każdą z krajowych prób przyporządkowano do jednej z trzech kategorii (z uwagi na typ operatu)[210], a następnie, w ramach takich grup, wyznaczono średnią arytmetyczną z wartości pewnych wskaźników w kolejnych krajach wraz z odpowiadającą jej wielkością błędu standardowego oraz z 90-procentowym przedziałem ufności dla oszacowania średniej wartości wskaźnika w grupie. Do analiz wybrane zostały takie mierniki terenowej realizacji próby, które – zgodnie z przedstawionymi wcześniej ustaleniami – powinny w największym stopniu różnicować postbadawcze wzorce terenowej realizacji prób adresowych oraz imiennych. Zestaw wskaźników obejmuje zatem mierniki: (a) wysokiej dostępności jednostek (tj. odsetek osób, do których udało się dotrzeć w trakcie

---

[209] Nie chodzi wyłącznie o to aby sprawdzić, czy wywiad został przeprowadzony z osobą dobraną przez ankietera do próby, ale czy wybrano osobę, która w próbie powinna się znaleźć. Jedną ze strategii przeciwdziałania niepożądanym praktykom ankieterskim jest oddzielenie fazy selekcji jednostek od etapu realizacji wywiadów. Co prawda, działanie takie jest dość powszechnie stosowane w losowaniu gospodarstw domowych z adresowych operatów budynków mieszkalnych (por. Stoop i in. 2010: 56), nie ma jednak szczególnego zastosowania w wewnątrzzespołowej selekcji jednostek próby z grona osób zamieszkujących dobrane gospodarstwa. A zatem, losowanie takie prowadzone jest już najczęściej przez tego samego ankietera, który następnie stara się zrealizować wywiad z dobraną przez siebie osobą. Wydaje się, że w tym przypadku oddzielenie etapu selekcji jednostek od fazy realizacji wywiadów wiązałoby się z poważnymi komplikacjami organizacyjnymi oraz negatywnie przekładałoby się na wskaźniki realizacji próby, choć z całą pewnością miałoby pozytywny wpływ na jakość procesu selekcji.

[210] Dla badań ESS-u przyporządkowanie krajowych prób badawczych do odpowiednich typów operatów przedstawione zostało w tabeli III.2. w trzecim rozdziale pracy.

pierwszej lub podczas drugiej próby nawiązania kontaktu)[211], (b) wysokiej gotowość do kooperacji (tzn. odsetek respondentów, z którymi wywiad udało się zrealizować podczas tej samej wizyty, w której nawiązano kontakt), (c) wysokiej gotowości do kooperacji przy jednocześnie łatwej dostępności (tzn. odsetek respondentów zaklasyfikowanych do kategorii powstałej po skrzyżowaniu (a) oraz (b))[212], a także (d) niskiej dostępność oraz niskiej gotowość do udziału w badaniu (tj. odsetek respondentów, z którymi nawiązanie kontaktu wymagało przeprowadzenia trzech lub większej liczby wizyt, natomiast do zrealizowania wywiadu konieczne było podjęcie jeszcze co najmniej jednej dodatkowej wizyty, licząc od momentu nawiązania kontaktu)[213]. Pozostałe dwa mierniki charakteryzują zbiór jednostek niedostępnych z uwagi na poziom (e) odmów udziału w badaniu wyrażonych bezpośrednio przez respondenta oraz (f) odmów pojawiających się na etapie wewnątrzzespołowej selekcji jednostek próby[214].

Konstrukcja tych wskaźników – być może poza (e) oraz (f) – wymaga krótkiego komentarza. Warto zatem wskazać, że przyjęty sposób definiowania dostępności jednostek pozostaje zgodny z propozycją tych badaczy, którzy za najlepszą miarę dostępności uznają liczbę wizyt potrzebnych ankieterowi do nawiązania pierwszego kontaktu z wylosowaną osobą[215] (por. Stoop i in. 2010: 116; Domański 2006: 51; Stoop 2005: 54). Oczywiście zrealizowanie wywiadów wymagać może jeszcze wielu wizyt (wysiłki te nie muszą wcale zakończyć

---

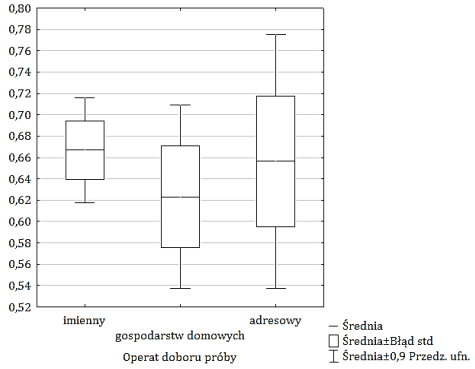
[211] Kategoria jednostek o wysokiej dostępności pozostaje przeciwieństwem warstwy osób „trudno dostępnych" (oryg. *hard-to-reach*), którą I. Stoop (2005: 39) zdefiniowała jako te jednostki próby, do których dotarcie wymagało podjęcia co najmniej trzech wizyt.

[212] Kategoria ta pozostaje równoważna warstwie tak zwanych „łatwych respondentów" (oryg. *easy-respondents*) (por. Stoop 2004: 39) czy też „ochoczych respondentów" (oryg. *willingness respondents*) (por. Lynn i in. 2002: 139).

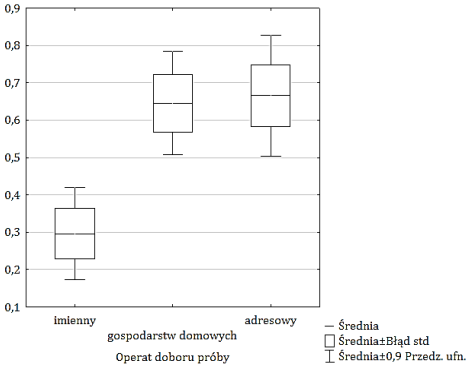
[213] Do tej warstwy zaliczone zostają również te wszystkie osoby, które przynajmniej raz odmówiły udziału w badaniu, a następnie zostały przekonane do udzielenia odpowiedzi i ostatecznie wzięły udział w wywiadzie. Takich respondentów określa się w literaturze mianem „nawróconych odmawiających" (oryg. *converted refusals*) lub „niechętnych respondentów" (oryg. *reluctant respondents*) (por. Jäckle i in. 2013: 1–15; Sakshaug i in. 2012: 113–122; Sztabiński i in. 2012: 95; Kaminska i in. 2010: 956–984; Billiet i in. 2009: 3–43; Schouten i in. 2009: 110; Sztabiński i in. 2009: 67–95; Billiet i in. 2007: 135–162; Stoop 2004: 39; Lynn i in. 2002: 139; Smith 1983: 391; Robins 1963: 276–286).

[214] Wartości tego ostatniego miernika ustalane są wyłącznie dla prób adresowych.

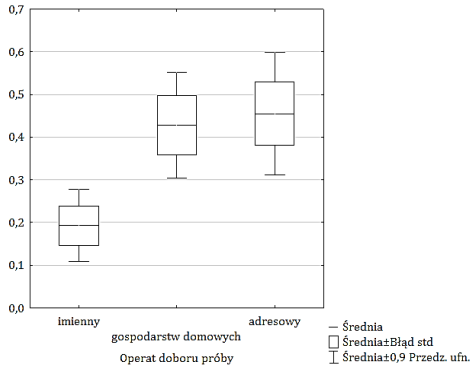
[215] Pomimo że miara ta uznawana jest za najtrafniejszy wskaźnik dostępności jednostek, to nie jest ona pozbawiona pewnych oczywistych wad. Dla przykładu: R. Groves i in. (1998: 81–82) ukazują, że dostępność można traktować jako prawdopodobieństwo tego, iż z wylosowanymi do próby osobami uda się nawiązać kontakt w okresie przewidzianym na terenową realizację próby. Tak definiowana dostępność jest stosunkiem czasu przebywania wylosowanej osoby w domu do czasu przeznaczonego na realizację badania. Można ją zatem szacować tylko przy założeniu losowości terminu, w którym podejmuje się próbę nawiązania kontaktu. Tymczasem, jak ukazuje I. Stoop (2005: 54), ankieterzy nie podejmują kontaktów w losowych terminach, a raczej wprowadzają własne strategie postępowania. W konsekwencji „trudna dostępność" może być przede wszystkim pochodną strategii działania ankieterów, a nie wynikać z wylosowania do próby osób „trudno osiągalnych".



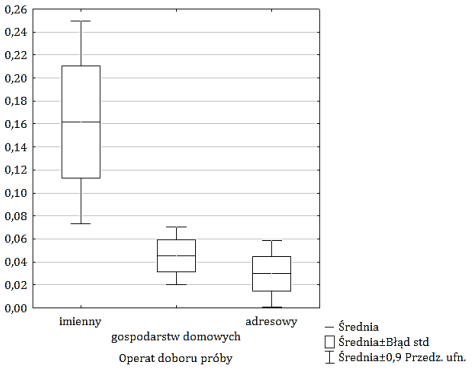
(a) Wysoka dostępność



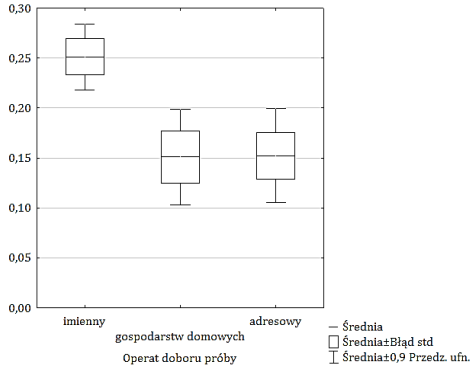
(b) Wysoka gotowość do udziału w badaniu



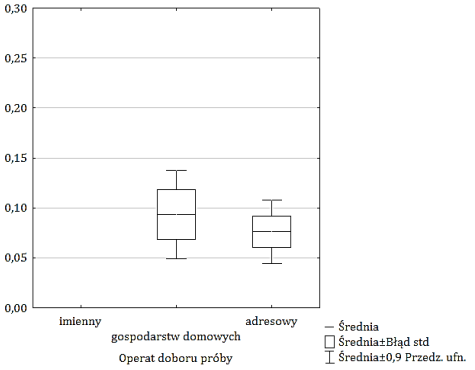
(c) Wysoka dostępność oraz wysoka gotowość do udziału w badaniu



(d) Niska dostępność oraz niska gotowość do udziału w badaniu



(e) Odmowa udziału wyrażona przez respondenta



(f) Odmowa na etapie selekcji respondenta

Ryc. V.3. Empiryczne zróżnicowanie wskaźników dostępności oraz gotowości do udziału w badaniu wg typów operatu doboru próby.

Źródło: obliczenia własne na podstawie repozytorium danych ESS4–2008

się sukcesem), jednakże te dodatkowe działania identyfikują już bardziej niechęć jednostek próby do kooperacji niż ich trudną dostępność (por. Domański 2006: 51). A zatem miernikiem gotowości do udziału w badaniu może być liczba wizyt, jaka była potrzebna do realizacji wywiadu, licząc od tej, w której udało się po raz pierwszy nawiązać kontakt z jednostką wylosowaną do badania. Tym samym wskaźniki dostępności charakteryzują zbiór tych wszystkich osób, z którymi udało się nawiązać kontakt (niezależnie od tego, czy następnie wywiad został zrealizowany, czy też nie), natomiast miary gotowości do kooperacji odnoszą się już wyłącznie do zbioru respondentów.

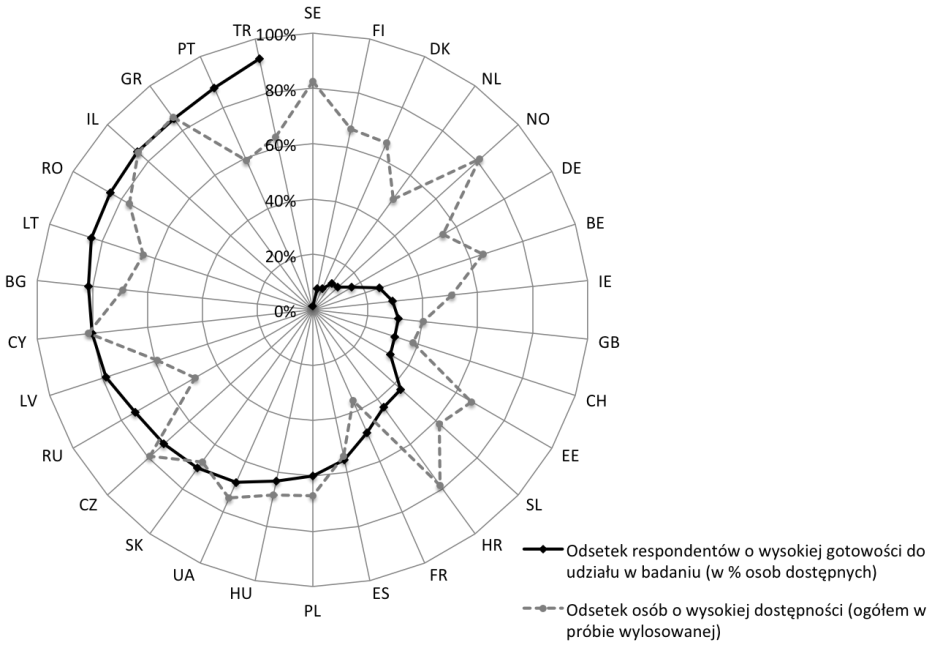
Przyglądając się wykresom zaprezentowanym na ryc. V.3., można zauważyć, że o ile typ operatu nie wpływał znacząco na szanse nawiązania kontaktu (wykres a), to miał już istotne przełożenie na poziom odmów (wykresy e oraz f), a także na wartości wskaźników gotowości jednostek próby do udziału w badaniu (wykresy b, c, d). Istotne jest przede wszystkim to, że wzorce realizacji obu typów prób adresowych były do siebie podobne, pozostając jednak odmienne od schematów realizacji prób imiennych. Potwierdza to wyrażone już wcześniej przypuszczenie, że wewnątrzzespołowa selekcja jednostek będzie głównym źródłem zróżnicowania wzorców terenowej realizacji prób imiennych oraz adresowych i to niezależnie od tego, czy miałaby ona sprzyjać systematycznemu zniekształcaniu procesu doboru jednostek, czy też skutkować większymi trudnościami w realizacji adresowych prób osób. Oczywiście w zależności od tego, z którym przypadkiem będzie się miało do czynienia, otrzyma się też inne wzorce terenowej realizacji prób badawczych. Tym samym zestawienie wskaźników gotowości do kooperacji oraz poziomu odmów udziału w badaniu powinno umożliwić stwierdzenie, na ile postbadawcze wzorce realizacji prób adresowych wynikają z właściwości przypisywanych operatom tego typu (porównywalna do prób imiennych dostępność oraz większy odsetek odmów), na ile zaś są konsekwencją zniekształcenia procesu selekcji wewnątrzzespołowej (większe niż w próbach imiennych wskaźniki dostępności oraz mniejsze wskaźniki odmów).

Zaprezentowane wyniki studiów empirycznych ukazują, że w próbach adresowych znaczną liczbę przypadków niezrealizowania wywiadów stanowią odmowy pojawiające się na etapie wewnątrzzespołowej selekcji respondentów (wykres f). Znaczące jest jednak to, że odsetek odmów udziału w badaniu wyrażonych bezpośrednio przez wybrane jednostki okazał się już istotnie mniejszy właśnie w próbach adresowych (wykres e). Z jednej strony potwierdza się więc intuicyjne przypuszczenie, że wielostopniowa selekcja stanowi barierę w realizacji prób adresowych. Z drugiej zaś, mniejsze wskaźniki odmów w próbach tego typu świadczyć mogą o zniekształceniu procesu selekcji jednostek poprzez wewnątrzzespołowy dobór osób bardziej skłonnych do kooperacji z ankieterem. Do podobnie niepokojących wniosków prowadzą również analizy wartości

pozostałych wskaźników. Przyglądając się ich zróżnicowaniu w grupach wyodrębnionych z uwagi na typ operatu, można wskazać, że w próbach imiennych warstwa respondentów składała się w znacznie mniejszym stopniu z jednostek o wysokiej dostępności i wysokiej gotowości do udziału w badaniu (wykresy b oraz c), więcej było natomiast osób o niskiej dostępności oraz niskiej gotowości do kooperacji z ankieterem (wykres d). Wystarczy przy tym wskazać, że w próbach adresowych udział „łatwych respondentów" stanowił mniej więcej 65%, podczas gdy w próbach imiennych osiągnął poziom 30%. Co więcej, około 16% wszystkich respondentów w próbach imiennych stanowili tak zwani „trudni respondenci", podczas gdy w próbach gospodarstw domowych udział tej kategorii wyniósł jedynie 5%, natomiast w próbach budynków mieszkalnych 3%. Tym samym otrzymane wyniki są sprzeczne z tym, czego należałoby oczekiwać, gdyby tylko proces wewnątrzzespołowej selekcji w operatach adresowych pozostał całkowicie wolny od systematycznych zniekształceń.

Trzeba być jednak niezwykle ostrożnym w formułowaniu jednoznacznych wniosków w tym zakresie. Wykorzystywanie danych z badań międzynarodowych ma bowiem swoje wyraźne ograniczenia. Warto wskazać, że każda z krajowych prób badawczych traktowana jest w takich zestawieniach jako reprezentacja pewnego typu operatu, co oznacza – *de facto* – przyjęcie założenia, iż to właśnie typ operatu pozostaje głównym źródłem międzykrajowych zróżnicowań terenowej realizacji prób badawczych. Tymczasem istnieć może wiele innych czynników, które w znacznie większym stopniu decydują o odmienności wzorców realizacji prób sondażowych. Najbardziej interesujące wydają się oczywiście te wszystkie czynniki, które pozostają pod kontrolą badacza (do nich zalicza się rodzaj operatu dobru próby, ale także technikę gromadzenia danych, liczbę prób przeznaczonych na nawiązanie kontaktu i realizację wywiadów, reguły konwersji odmów, występowanie gratyfikacji za udział w badaniu, wyszkolenie ankieterów oraz kontrola ich pracy itd. (por. de Heer 1999: 136–139), niemniej jednak na zróżnicowanie wzorców terenowej realizacji próby wpływać mogą także czynniki pozostające poza kontrolą badacza (por. Smith 2007: 45; Groves i in 1998: 29–31). Należy więc przyjrzeć się nieco bliżej realizacji prób badawczych w poszczególnych krajach uczestniczących w ESS4-2008, by zobaczyć, czy rzeczywiście wykorzystanie operatów adresowych prowadzi częściej, niż ma to miejsce w próbach imiennych, do zniekształcenia procesu selekcji jednostek. Na rycinie V.4. uszeregowano kraje biorące udział w badaniach czwartej rundy ESS pod względem wartości wskaźnika (wysokiej) gotowości respondentów do udziału w badaniu, podając również krajowe wartości miernika (wysokiej) dostępności wylosowanych jednostek.

Wartości tych wskaźników dostarczają kolejnych dowodów na to, że łatwość nawiązania kontaktu oraz gotowość do udziału w badaniu stanowią niezależne wymiary porządkujące w odmienny sposób jednostki próby badawczej. A zatem osoby, z którymi trudno jest nawiązać kontakt, mogą być rzeczywiście



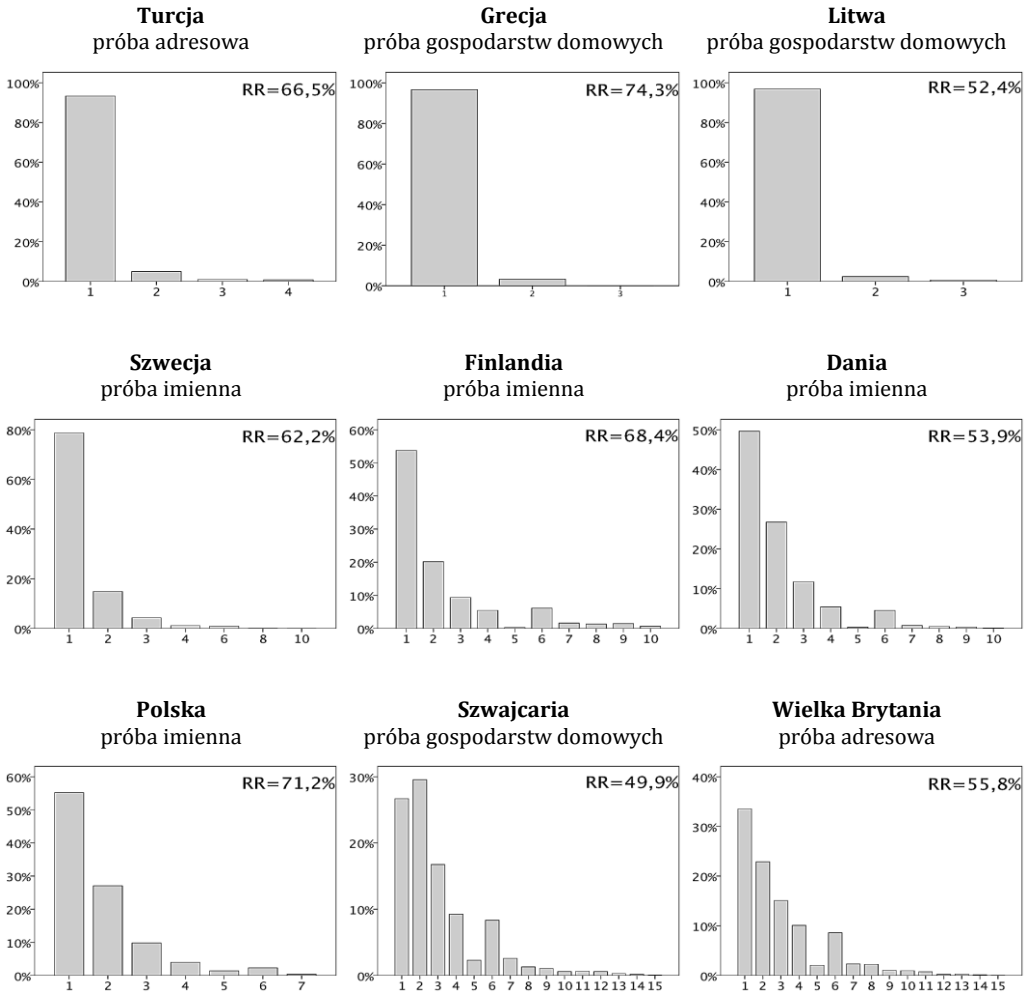
Ryc. V.4. Odsetek respondentów o wysokiej dostępności oraz gotowości do udziału w badaniu – zróżnicowania międzykrajowe na przykładzie ESS4–2008

Źródło: obliczenia własne na podstawie repozytorium danych ESS4-2008

w takim samym stopniu skłonne (lub w takim samym stopniu niechętne) do kooperacji z ankieterem jak osoby łatwo dostępne (por. Stoop i in. 2010: 116). Przyglądając się uszeregowaniu krajów, zauważyć można również niezwykle symptomatyczny fakt, iż w grupie o największych wartościach wskaźnika wysokiej gotowości do udziału w badaniu znalazły się wyłącznie te państwa, w których dobór próby był prowadzony z operatów adresowych. Dopiero na Węgrzech oraz w Polsce (a więc w dwóch krajach zajmujących odpowiednio czternastą oraz piętnastą pozycję w rankingu krajów o wysokiej kooperatywności jednostek) do losowania próby wykorzystano operaty imienne. Co więcej, w grupie o najniższych wartościach mierników gotowości do udziału w badaniu dominują już te kraje, w których próby dobierano z operatów jednostkowych. Zajmują one odpowiednio siedem ostatnich pozycji, dopiero na ósmym (od końca) miejscu uplasowała się Irlandia, w której losowanie próby prowadzone było z rejestru gospodarstw domowych[216]. A zatem, nawet jeśli typ operatu nie

[216] Irlandia stanowi zresztą przypadek dość specyficzny. Badacze dysponują operatem imiennym, który z uwagi na niepełne pokrycie badanej populacji wykorzystuje się jedynie do losowania gospodarstw domowych. Zagadnienia te zostały omówione w trzecim rozdziale pracy.

miałby być czynnikiem determinującym określony wzorzec terenowej realizacji próby, to bezdyskusyjne jest stwierdzenie, że w krajach o najwyższych wskaźnikach kooperacji wykorzystywane były właśnie rejestry adresowe. Jednakże dopiero szczegółowa analiza przypadków niektórych państw pozwoli ukazać, że wzorce terenowej realizacji prób adresowych dają w wielu przypadkach mało wiarygodne rezultaty.

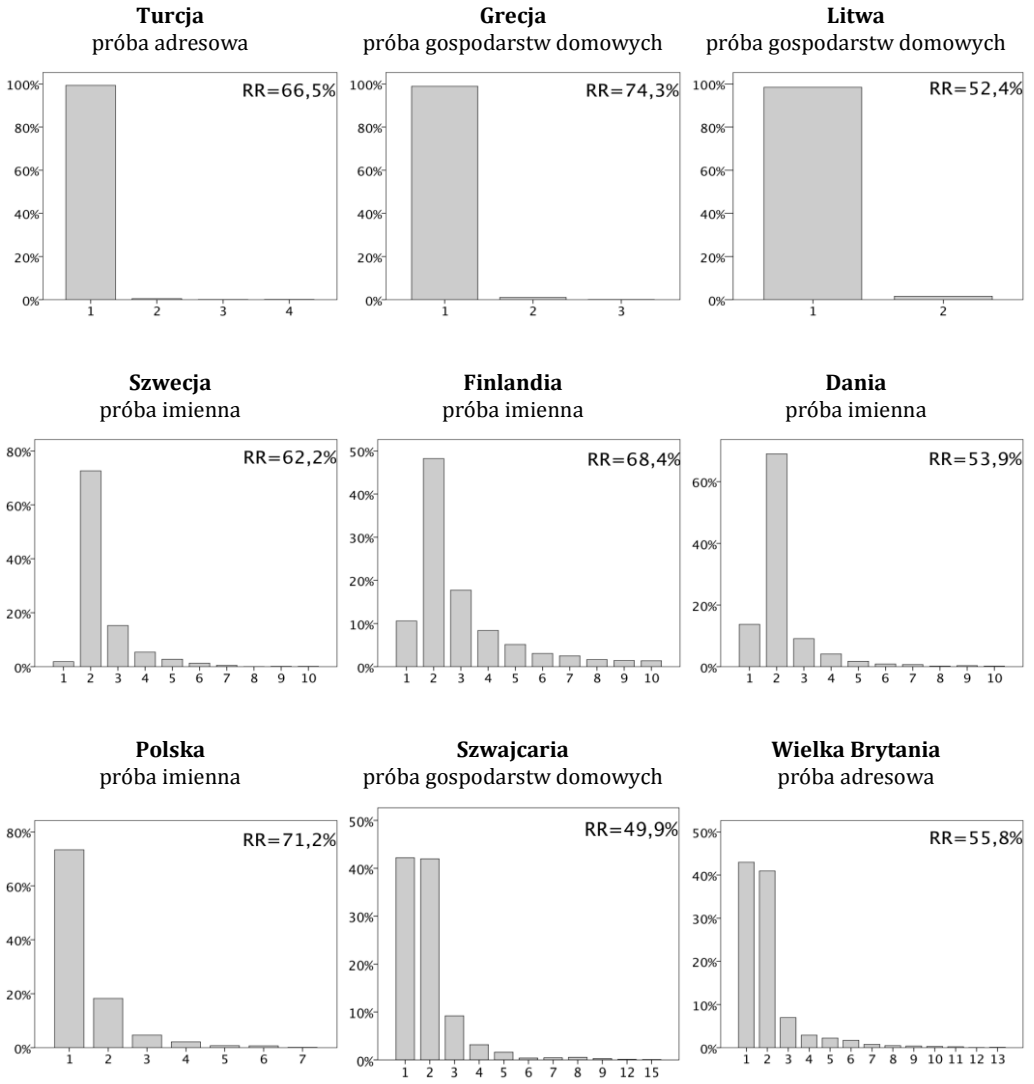


Ryc. V.5. Rozkład liczby wizyt potrzebnych do nawiązania kontaktu z respondentem w wybranych krajach ESS4–2008

Źródło: obliczenia własne na podstawie repozytorium danych ESS4–2008

Do dalszych studiów wybranych zostało dziewięć państw reprezentujących pewne charakterystyczne typy wzorców terenowej realizacji prób adresowych oraz imiennych. Turcja, Grecja oraz Litwa są chyba najbardziej wyrazistymi przykładami krajów wykorzystujących operaty adresowe (w skład tej grupy zaliczyć można wszystkie trzynaście państw o najwyższych wskaźnikach wysokiej gotowości do udziału w badaniu (zob. ryc. V.4.), dla których w postbadawczych wzorcach terenowej realizacji próby sondażowej doszukać można się dowodów na mniej lub bardziej znaczące zniekształcenie procesu selekcji jednostek. Szwecja, Finlandia oraz Dania reprezentują zbiór krajów wykorzystujących operaty imienne, w których poziom dostępności jednostek nie odbiega istotnie od przeciętnej wartości w próbach tego typu, ale charakteryzują się one już najmniejszymi wartościami wskaźników wysokiej gotowości do kooperacji. Z kolei schematy realizacji prób badawczych w Polsce (operat imienny), Szwajcarii (operat gospodarstw domowych) oraz Wielkiej Brytanii (operat adresowy) reprezentują zbiór tych wszystkich krajów biorących udział w ESS4–2008, w których postbadawcze wzorce są najbardziej zgodne z tym, czego należałoby się spodziewać z uwagi na charakterystyki przypisywane operatom imiennym i adresowym.

Przyglądając się rozkładowi liczby wizyt potrzebnych do nawiązania pierwszego kontaktu z osobami wylosowanymi do próby (ryc. V.5.), nietrudno zauważyć, że w Turcji, Grecji oraz na Litwie prawie wszystkie przypadki nawiązania kontaktu miały miejsce już w trakcie pierwszej wizyty ankietera pod wskazanym adresem. Dla kontrastu, w Szwajcarii oraz w Wielkiej Brytanii w trakcie pierwszej wizyty udawało się dotrzeć jedynie do mniej więcej 30% wszystkich osób. Trudno wyjaśnić te zróżnicowania inaczej niż tym, że wynikają one z jakiejś specyfiki działań podejmowanych w trakcie terenowej realizacji próby. Co więcej, Turcja oraz Grecja zaliczają się do grupy krajów o wysokich wskaźnikach realizacji próby, z kolei na Litwie wartość *response rate* była już bardzo niska. Analizując ten problem nieco dokładniej, można wskazać, że w Turcji odsetek odmów pojawiających się na etapie wewnątrzzespołowej selekcji wyniósł niecałe 10 pp., w Grecji niespełna 5 pp., natomiast na Litwie osiągnął poziom aż 25,9 pp. Z kolei odsetki odmów wyrażanych bezpośrednio przez osoby wylosowane do próby były już w tych krajach na porównywalnych poziomach (10,2 pp. w Turcji, 14,4 pp. w Grecji oraz 13,7 pp. na Litwie). Niezwykle symptomatyczne jest to, że pomimo dość znaczącej liczby przypadków odmów (bezpośrednich lub na etapie wewnątrzzespołowej selekcji) nie podejmowano w tych krajach w ogóle próby ich konwersji. Przeciwieństwem takiej strategii postępowania są schematy realizacji próby w Szwajcarii oraz w Wielkiej Brytanii. W obu państwach otrzymano równie znaczne odsetki odmów jak w Turcji, Grecji oraz na Litwie, jednak działaniami zmierzającymi do ich konwersji objęto już odpowiednio 40 oraz 25 procent wszystkich osób niechętnych udziałowi w badaniu.



Ryc. V.6. Rozkład liczby wizyt potrzebnych do przeprowadzenia wywiadu (po nawiązaniu kontaktu z respondentem) w wybranych krajach ESS4-2008

Źródło: obliczenia własne na podstawie repozytorium danych ESS4-2008

Działania takie mają przełożenie nie tylko na rozkład liczby wizyt potrzebnych do nawiązania kontaktu, ale także na rozkład liczby wizyt, jakie okazały się konieczne do przeprowadzenia wywiadu (ryc. V.6.).

W wielu krajach uczestniczących w projekcie ESS rozkład liczby wizyt koniecznych do realizacji wywiadu (licząc od momentu nawiązania kontaktu),

przyjmował postać, którą chyba najlepiej reprezentuje przypadek Polski. Chociaż większość wywiadów udawało się zrealizować podczas tej samej wizyty, w trakcie której nawiązano kontakt, to jednak niemałą część respondentów stanowiły jednostki niechętne udziałowi w badaniu. Na tym tle wyróżniają się rozkłady otrzymane w Szwajcarii oraz Wielkiej Brytanii[217], przede wszystkim jednak te ze Szwecji, Finlandii oraz Danii (podobne rozkłady otrzymano również w Norwegii oraz Holandii). W krajach tych zdecydowana mniejszość wywiadów przeprowadzana była podczas tej samej wizyty, w której nawiązano kontakt, najczęściej potrzebne było podjęcie przynajmniej jeszcze jednej takiej próby dotarcia w celu realizacji wywiadu. Objaśnienie takiego stanu rzeczy odnaleźć można w informacjach zawartych w repozytoriach ESS-u, a także w przywoływanej wcześniej monografii Stoop i in. (2010) poświęconej analizie metodologicznej pierwszych trzech odsłon tego projektu. Autorzy wspomnianej książki zwracają uwagę, by przy interpretacji międzykrajowych zróżnicowań wskaźników kooperacji pamiętać, iż w Norwegii, Finlandii, Szwecji oraz Danii zdecydowana większość pierwszych prób nawiązania kontaktu podejmowana była przez telefon. Efektem takiej rozmowy mogło być co najwyżej nawiązanie kontaktu oraz umówienie się z wylosowaną osobą na termin wizyty ankietera, w badaniach ESS nie przewiduje się bowiem możliwości realizacji wywiadów w inny sposób niż tylko bezpośrednio (por. Stoop 2010: 149). Ponieważ uwaga ta odnosi się także do czwartej rundy badań ESS-u, to znaczna liczba wywiadów zrealizowanych w tych krajach dopiero przy drugim kontakcie wynika z przyjętego schematu realizacji badania i ani nie jest konsekwencją losowania próby z operatu imiennego, ani też specyfiki populacji krajów skandynawskich[218]. Gdyby wyłączyć z prowadzonej analizy pierwsze (realizowane przez telefon) próby nawiązania kontaktu, to rozkład liczby wizyt potrzebnych do realizacji wywiadu byłby już w tych krajach bardzo podobny do – uznanego za wzorcowy/typowy – rozkładu w polskim komponencie projektu ESS4.

Nietypowe są również rozkłady otrzymane w Turcji, Grecji oraz na Litwie (a także w innych krajach o najwyższych wskaźnikach kooperacji). Można bowiem zauważyć, że niemal wszystkie wywiady udawało się zrealizować w tych

---

[217] A także innych krajów, na przykład Irlandii oraz Francji, w których zastosowanie operatów adresowych prowadzi do – zgodnej z charakterystykami operatów tego typu – mniejszej liczby jednostek „łatwo dostępnych".

[218] Konsekwencją przyjęcia takiego schematu realizacji próby mogłoby być również to, że odnotowana w tych krajach (por. ryc. V.5.) znaczna liczba jednostek „łatwo dostępnych" nie wynika rzeczywiście z ich łatwej dostępności, ale jest pochodną tego, że za nawiązanie kontaktu uznawano wszystkie przypadki podjęcia rozmowy telefonicznej przez kogoś zamieszkującego gospodarstwo domowe wylosowanej osoby. W sposób zupełnie oczywisty takim rozmówcą nie musi być ta konkretna osoba dobrana z operatu imiennego. W prowadzonych tutaj analizach sytuacja taka została wyeliminowana. Rozkład liczby wizyt potrzebnych do nawiązania kontaktu obejmuje bowiem wyłącznie przypadki skontaktowania się z osobą wylosowaną do próby.

państwach podczas tej samej (najczęściej pierwszej) wizyty, w której nawiązano kontakt z wylosowaną osobą. Innymi słowy, warstwa respondentów składała się prawie wyłącznie z osób o wysokiej dostępności oraz wysokiej gotowości do kooperacji; tym samym respondenci trudno dostępni oraz niechętni udziałowi w badaniu stanowili nieliczne wyjątki. Zresztą takie same wzorce realizacji próby odnotowywane były w tych krajach również we wcześniejszych odsłonach Europejskiego Sondażu Społecznego (por. Stoop i in. 2010: 146, 150). Wydaje się zatem dość oczywiste, że znaczne wartości wskaźników wysokiej dostępności oraz wysokiej gotowości do udziału w badaniu w tych krajach wynikają głównie ze schematu realizacji wywiadów, a nie z jakiejś specyfiki tych krajów, czy też wreszcie z tego, że ma się do czynienia z operatem adresowym. W krajach tych etap wewnątrzzespołowej selekcji traktowany był w dość specyficzny sposób, jako swego rodzaju skrining mający wskazać te jednostki populacji, które charakteryzują się wysoką dostępnością oraz skłonnością do udziału w badaniu. Wywiady prowadzone były następnie wyłącznie z takimi osobami, z kolei te, które były choć trochę niechętne udziałowi w badaniu, nie stanowiły w tych krajach przedmiotu dalszych działań badawczych. Stawia to pod znakiem zapytania zasadność międzykrajowych porównań wyników badań. W gruncie rzeczy, jaki mają one sens, skoro kompozycja zbioru respondentów pod względem dostępności oraz gotowości do kooperacji podlega tak silnym zniekształceniom? Oczywiście taki skrining mógłby być również powodem wypaczenia prób imiennych. Warto jednak zauważyć, że w żadnym z krajów, w których wykorzystywano operaty imienne, nie uzyskano wzorców terenowej realizacji próby odstających w nieprzewidywalny sposób od tego, czego można by oczekiwać po terenowej realizacji takich prób. Tym samym, nawet jeśli to nie specyfika operatów adresowych, a raczej ustalony przez badaczy schemat terenowej realizacji próby odpowiada za zniekształcenie wyników badania, to bezdyskusyjne jest stwierdzenie, że wyłącznie w realizacji prób adresowych napotykano na mało wiarygodne wzorce jej terenowej realizacji. Innymi słowy, nawet jeśli operaty adresowe nie skutkują w sposób nieuchronny zniekształceniem próby, to dają – znacznie większą niż w próbach imiennych – możliwość przeprowadzenia procesu selekcji wielostopniowej w sposób prowadzący do takiego zniekształcenia.

V.3. Od braku danych do błędu braku danych – deterministyczny oraz probabilistyczny paradygmat błędu niepełnej realizacji próby

Warto odejść już od problemów związanych z postbadawczą klasyfikacją respondentów oraz jednostek niedostępnych, a także od wzorców realizacji prób sondażowych, i przyjrzeć się, w jaki sposób niepełna realizacja próby

przekłada się na błędy braków danych. W większości współczesnych analiz poświęconych konsekwencjom wynikającym z niedostępności pewnych jednostek wylosowanych do próby badawczej przeciwstawia się dwa sposoby definiowania błędu niepełnej realizacji próby. Jeden z nich wywodzi się z założeń przyjmowanych w tak zwanym paradygmacie deterministycznym, drugi zaś w probabilistycznym (por. na przykład Stoop i in. 2010: 31; Brick i in. 2009: 170–171; Montaquila i in. 2008: 561–586; Groves 2006: 647–649; Stoop 2005: 32–33; Särndal i in. 2005: 49–50; Groves i in. 2004: 182). Warto zresztą przypomnieć, iż w drugim rozdziale pracy zdefiniowany został błąd braku danych w duchu założeń paradygmatu deterministycznego. Pokazano wówczas, że jego wielkość daje się określić jako iloczyn (1) frakcji jednostek niedostępnych oraz (2) różnicy wartości estymatorów w zbiorze respondentów i warstwie osób niedostępnych[219]. Podane zostały także definicje błędów odpowiednie dla oszacowań parametrów wskaźnika struktury (por. wzory II.14 i II.14') oraz średniej arytmetycznej (por. wzory II.15 i II.15'). Ze wzorów tych wynikało, że wielkość

---

[219] Zdecydowana większość procedur służących oszacowaniu wielkości błędu niepełnej realizacji sondażowej próby badawczej opiera się bezpośrednio na założeniach przyjmowanych w paradygmacie deterministycznym. Wielkość błędu wyznaczana jest poprzez określenie różnicy między warstwą respondentów oraz zbiorem osób niedostępnych. Oczywiście, znacznym utrudnieniem w ocenie wielkości tak definiowanego błędu pozostaje ograniczony dostęp do danych charakteryzujących jednostki nieprzebadane. Źródłem takich dodatkowych informacji są jednak: (1) dane zawarte w operatach doboru próby oraz w repozytoriach statystyki publicznej (przykłady ciekawych oszacowań błędu w oparciu o zewnętrzne dane odnaleźć można w opracowaniach Sakshaug i in. 2012: 113–122, Kreuter i in. 2010: 880–906, Johnson i in. 2006: 704–719 oraz Kohler 2007: 55–67), (2) tak zwane badania *follow-up* przeprowadzane po badaniach właściwych na grupie wszystkich lub losowo dobranych jednostek niedostępnych (por. wczesne zastosowania tej metody w Hansen i in., 1946: 517–529 oraz przykłady jej współczesnych implementacji przedstawione w opracowaniach Peytchev i in. 2009: 785–806, Sztabiński i in. 2008: 39–84, Sztabiński i in. 2007: 25–54, Olson 2006: 737–758, Stoop 2004: 31–38, Domański 1999: 67–92), (c) *paradane* (por. Lyberg 2012: 107–130, Olson 2011a: 21–26, Smith 2011: 389–402, West 2011: 1–8, Kreuter i in. 2010: 389–407, Kreuter i in. 2009: 203–226, Bates i in. 2008: 591–612, Beaumont 2005: 227–231) charakteryzujące fazę kontaktu ankietera z respondentem, (d) charakterystyki jednostek niedostępnych pozyskiwane metodą *Basiq Question Procedure* (por. Bethlehem i in. 1985: 287–300, Kersten i in. 1984: 369–380) lub jej zmienioną wersję w postaci metodologii *PEDAKSI* (por. Lynn 2003: 239–269), (e) zestawienia specyficznych podzbiorów respondentów (np. osób trudno osiągalnych, niechętnych pomiarowi itp.) oraz traktowanie takich warstw jednostek przebadanych jako substytutu osób niedostępnych (do tej klasy procedur zaliczyć można ważenie *Politza-Simmonsa* (por. Politz i in. 1949: 9–31), jak również znane w literaturze metody szacowania wielkości błędu w oparciu o tak zwane *kontinuum oporu* oraz *model klas* niedostępności (por. Lin i in. 1995: 236–258, Smith 1983: 386–404). Dokładne usystematyzowanie oraz omówienie tych (oraz innych) procedur odnaleźć można w monografii R. Grovesa i in. (1998: 48–51) oraz artykule R. Grovesa (2006: 654–657). Do opracowań tych odnosi się też wielu badaczy (por. na przykład Matsuo i in. 2010: 165–166, Stoop i in. 2010: 207–281, Billiet i in. 2009: 8–9, Stoop 2005: 105–112, Stoop 2004: 25–26). W polskiej literaturze wyczerpujące omówienie głównych metod służących empirycznej ocenie wielkości błędu braku danych w duchu założeń deterministycznych wraz z analizą ograniczeń związanych z ich praktycznym zastosowaniem odnaleźć można w pracy K. Grzeszkiewicz-Radulskiej (2009: 45–68).

błędu systematycznego uwarunkowana była w znacznej mierze stopniem podobieństwa rozkładów badanej zmiennej w warstwie respondentów oraz w zbiorze osób nieprzebadanych; miał na nią także wpływ wskaźnik realizowalności próby. Jeżeli zatem różnice w rozkładach analizowanych zmiennych w tych warstwach były niewielkie (lub nie występowały w ogóle), to wielkość błędu systematycznego była bliska (lub równa) wartości zerowej. Z drugiej strony, im różnice były większe, tym większa była też wartość błędu. W tym pierwszym przypadku błąd miał charakter losowy (zmniejszał precyzję estymacji), ale już nie systematyczny (to znaczy nie prowadził do zniekształcenia wyników pomiaru) i to niezależnie od tego, jak duża część próby pozostała nieprzebadana. Z kolei w drugim przypadku niepełna realizacja próby obniżała zarówno precyzję, jak i dokładność prowadzonego badania. Należy jednak pamiętać, że sytuacje, w których błąd niepełnej realizacji próby ma charakter całkowicie losowy, są wyjątkiem, trzeba raczej przyjąć, że uchybienia wynikające z braków danych będą przyjmować postać błędów systematycznych.

Zanim przedstawione zostaną założenia – zyskującej coraz większe znaczenie – probabilistycznej koncepcji błędu braku danych oraz wynikające z niej konsekwencje praktyczne i metodologiczne, przeanalizowane będą wpierw podstawy modelu deterministycznego[220], tak, aby wyraźnie widoczny był kontrast między oboma sposobami podejściami do niepełnej realizacji próby badawczej. Typowym przykładem reprezentującym taki deterministyczny sposób myślenia o błędach braków danych są analizy zaprezentowane przez W. Cochrana w 1977 roku w monografii *Sampling Techniques*. W rozdziale trzynastym tej pracy, w części poświęconej jednostkom niedostępnym, autor ten zamieszcza następującą konkluzję:

> w badaniach nad niedostępnością jednostek 'wygodnie' jest myśleć o populacji tak, jakby można było ją podzielić na dwie warstwy, pierwsza składałaby się ze wszystkich jednostek, dla których udałoby się przeprowadzić pomiar, gdyby znalazły się one w próbie, druga natomiast z jednostek, dla których pomiar nie zostałby przeprowadzony. Skład tych dwóch warstw zależy ściśle od procedur wykorzystanych do nawiązania kontaktu z wylosowanymi jednostkami oraz realizacji z nimi wywiadu. (Cochran 1977: 359–360)

Takie deterministyczne ujmowanie błędu braku danych pozostaje zbieżne z przyjmowaną w literaturze badań reprezentatywnych definicją błędu niepeł-

---

[220] Powołując się na ustalenia Särndala i in. (2005: 50), należy dodać, że takie deterministyczne spojrzenie na błąd braku danych dominowało w literaturze metodologicznej mniej więcej do początku lat 80. XX wieku. Wystarczy przeprowadzić jednak pobieżną kwerendę biblioteczną, by ukazać, że duch deterministyczny pojawia się również we współczesnych studiach nad niedostępnością jednostek wylosowanych do próby (por. Jabkowski 2011: 29–30; Sztabiński F. 2011: 50–51; Stoop i in. 2010: 205–206; Billiet i in. 2009: 6; Billiet i in. 2007: 137; Jabkowski 2007: 72–73; Biemer 2001: 300).

nego pokrycia jednostek populacji przez operat doboru próby[221]. Istotnie, w obu przypadkach populację dzieli się na dwie rozłączne warstwy pokrywające w stopniu zupełnym całą populację, tj. odpowiednio na zbiory jednostek znajdujących się w operacie i jednostek pominiętych przez operat lub respondentów i jednostek niedostępnych. Co więcej, przyjmuje się też, iż szanse realizacji wywiadu z osobą dobraną do próby z warstwy osób niedostępnych będą równe zeru, podobnie jak zerowe są też szanse wylosowania jednostki znajdującej się poza operatem doboru próby (por. Särndal i in. 2005: 50).

Choć takie myślenie o populacji jako o sumie jej dwóch rozłącznych warstw wydaje się właściwe w analizie błędu operatu, to jednak w studiach nad niepełną realizacją próby badawczej jest to już zbyt dużym uproszczeniem. O ile rzeczywiście osoby znajdujące się poza operatem nie mają żadnych szans na wylosowanie, o tyle każda jednostka pokryta przez operat oraz dobrana do próby charakteryzuje się już pewną – nieznaną, ale większą od zera – skłonnością do udziału w badaniu. Tym samym, ponieważ szanse realizacji wywiadu zależą od podejmowanych działań badawczych (na przykład od techniki pomiarowej, procedur dotarcia do respondenta, nawiązania z nim kontaktu czy też działań mających na celu skłonienie wylosowanych osób do udziału w badaniu), to nie da się z góry ustalić, która z dobranych do próby jednostek zalicza się do populacyjnej warstwy respondentów, a która do zbioru niedostępnych jednostek populacji[222]. Zresztą na ograniczenia wynikające z deterministycznej koncepcji błędu zwrócił uwagę W. Cochran w przywołanej już monografii *Sampling Techniques*. Stwierdził on, iż:

> taki podział na dwie rozłączne warstwy jest oczywiście uproszczeniem. Szanse odgrywają rolę w tym, czy jednostkę uda się odnaleźć i zrealizować z nią pomiar […]. Celem pełniejszego opisu tego problemu moglibyśmy przypisać każdej jednostce prawdopodobieństwo odpowiadające szansie tego, że gdyby tylko została ona wylosowana do próby, to udałoby się z nią przeprowadzić pomiar, przy zastosowaniu określonych procedur terenowych. (Cochran 1977: 360)

Podobne zastrzeżenia zgłaszali również P. Biemer oraz L. Lyberg w monografii *Introduction to Survey Quality*, mówiąc, że:

---

[221] Zwrócono na to uwagę we wprowadzeniu do rozdziału III, przywołując ustalenia P. Biemera oraz L. Lyberga (2003) którzy wskazali, że „błąd operatu doboru próby oraz błąd braku danych oddziałują w bardzo podobny sposób na błąd średniokwadratowy. Niektórzy mogliby nawet uznać, że jednostki pominięte w operatach doboru próby są jednym z typów jednostek niedostępnych, gdyż w obu przypadkach informacje o jednostkach nie są znane" (Biemer i in. 2003: 63). W tym samym tonie wypowiadał się R. Groves, mówiąc, że „podobnie jak w niepełnym pokryciu populacji przez operat doboru, niepełna realizacja próby jest błędem braku obserwacji" (Groves 1989: 133).

[222] W przypadku błędu niepełnego pokrycia populacji przez operat ma się przynajmniej pewność, że dobiera się jednostki z jednej tylko warstwy, tj. elementy populacji docelowej zawarte w operacie.

> jeśli dobieramy próbę, to proces losowania nie uwzględnia podziału na jednostki dostępne i niedostępne. […] Ilekroć losujemy osobę z warstwy jednostek niedostępnych, oznaczymy ją jako niedostępną, o ile [nie uda się z nią przeprowadzić badania – P.J.]. […] Taki teoretyczny podział populacji na dwie kategorie jest prostym deterministycznym modelem bardzo złożonej rzeczywistości. […] Tak naprawdę, podział na jednostki dostępne i niedostępne nie jest taki prosty. […] Bardziej złożony model powinien uwzględniać założenie, iż jednostki charakteryzują się pewnymi prawdopodobieństwami udziału w badaniu. […] Model deterministyczny zakłada natomiast, że te prawdopodobieństwa są równe 0 lub 1. (Biemer i in. 2003: 82)

Należy przy tym wskazać, że ów deterministyczny (a tym samym uproszczony) sposób patrzenia na błędy braków danych odnaleźć można w wielu znakomitych opracowaniach poświęconych problemom reprezentatywności sondażowych prób badawczych. Doskonałym tego przykładem są rozważania R. Grovesa (1989) z jego monumentalnego dzieła *Survey Errors and Survey Costs*, w którym wspomina on co prawda, że:

> kategoria jednostek niedostępnych może zostać zdefiniowana jako stała grupa osób w populacji pokrytej operatem losowania. Alternatywnie kategorię respondentów można traktować jako zmienną w kolejnych replikacjach próby, ponieważ każda osoba posiada pewne prawdopodobieństwo udziału w badaniu. (Groves 1989: 134),

jednak studia empiryczne oraz analizy teoretyczne na temat błędu braku danych autor ten prowadzi już wyłącznie w duchu paradygmatu deterministycznego[223]. Zresztą w podobnym tonie wypowiadają się również R. Groves oraz M. Couper w monografii *Nonresponse in Household Interview Surveys*. We wstępie do rozdziału poświęconego statystycznym konsekwencjom błędów braków danych autorzy tego znanego dzieła wskazują na zakorzenienie przyjmowanej przez nich teorii błędu w założeniach paradygmatu deterministycznego. Wskazują bowiem, że:

> każda jednostka populacji docelowej jest – trwale i na zawsze – respondentem lub jednostką niedostępną […]. W trakcie doboru próby nie wiemy, do której warstwy należy jednostka, stąd losujemy osoby z warstwy respondentów oraz jednostek niedostępnych. Możemy przypuszczać, że frakcja respondentów

[223] Ponieważ R. Groves rozważał błędy braku danych w paradygmacie całkowitego błędu pomiaru, to uchybienia wynikające z niepełnej realizacji próby definiował jako wartość oczekiwaną z błędów każdej możliwej replikacji próby badawczej. Co prawda w takim ujęciu frakcja jednostek niedostępnych oraz różnica wartości estymatorów są zmiennymi losowymi, jednak w odniesieniu do konkretnej próby badawczej model przyjęty przez Grovesa jest całkowicie zgodny z podejściem deterministycznym (por. Groves 1989: 40–41).

Tabela V.4. Deterministyczny oraz probabilistyczny model błędu braku danych – różnice w obu sposobach definiowania błędu oraz wynikające z tego konsekwencje

Wymiary porównań	Model deterministyczny	Model probabilistyczny
(1) Podstawowe założenia	1. Badaną populację można podzielić na warstwę respondentów oraz osób niedostępnych. 2. Niepełna realizacja próby jest efektem tego, że prawdopodobieństwo realizacji wywiadów z osobami wylosowanymi z warstwy jednostek niedostępnych wynoszą 0.	Każdą jednostkę wylosowaną do próby badawczej charakteryzuje niezerowa skłonność do udziału w badaniu, a zatem – w sensie liczbowym – skłonność taka mieści się w przedziale (0;1].
(2) Definiowanie błędu niepełnej realizacji próby	Wielkość systematycznego błędu braku danych zależy od: a) wskaźnika realizacji próby, b) różnicy pomiędzy wartością estymatora w warstwie respondentów oraz warstwie jednostek niedostępnych.	Wielkość systematycznego błędu braku danych zależy od: a) przeciętnej skłonności wylosowanych osób do udziału w badaniu, b) kowariancji pomiędzy wartościami analizowanej zmiennej oraz wartościami jednostkowych skłonności do udziału w badaniu.
(3) Wskaźnik realizacji próby	1. Wskaźnik realizacji próby badawczej jest oszacowaniem wielkości warstwy osób dostępnych w badanej populacji. 2. Wielkość tej warstwy nie jest stała, ale zależy od wielu czynników pozostających pod kontrolą oraz poza kontrolą badacza.	1. Wskaźnik realizacji próby badawczej jest wartością oczekiwana przeciętnej skłonności jednostek populacji do udziału w badaniu. 2. Na skłonność jednostek do udziału w badaniu wpływa wiele czynników pozostających pod kontrolą oraz poza kontrolą badacza.
(4) Szacowanie wielkości błędu niepełnej realizacji próby	W szacowaniu wielkości błędu braku danych dąży się do porównania wartości zmiennych w zbiorze respondentów oraz jednostek niedostępnych lub do znalezienia (w zbiorze respondentów) jednostek będących substytutami osób niedostępnych.	W szacowaniu wielkości błędu braku danych podstawowym wyzwaniem jest estymacja jednostkowych skłonności do udziału w badaniu oraz identyfikacja mechanizmu kształtującego charakter niedostępności jednostek.
(5) Uzasadnienie dla działań zmierzających do maksymalizacji wskaźnika realizacji próby	Maksymalizacja wskaźnika realizacji próby jest celem samym w sobie. Wszystkie działania terenowe służące zmniejszeniu frakcji jednostek niedostępnych pozwalają ograniczyć maksymalną wielkość błędu systematycznego.	Działania zmierzające do maksymalizacji wskaźnika realizacji próby mają uzasadnienie tylko i wyłącznie wtedy, gdy ich konsekwencją jest zmniejszenie zróżnicowania w zbiorze jednostkowych skłonności do udziału w badaniu lub wyeliminowanie kowariancji pomiędzy wartościami analizowanej zmiennej oraz jednostkowymi skłonnościami do udziału w badaniu.

Źródło: opracowanie własne

> w próbie powinna być równa frakcji tej warstwy w populacji, choć wielkość ta będzie zróżnicowana z uwagi na proces próbkowania. (Groves i in. 1998: 2–3)

W kolejnym paragrafie tej książki odnaleźć można wzmiankę o stochastycznej naturze błędu niepełnej realizacji próby badawczej – autorzy mówią bowiem, że „alternatywne spojrzenie na jednostki niedostępne zakłada, iż każda osoba z próby ma pewne prawdopodobieństwo bycia respondentem lub jednostką niedostępną" (Groves in. 1998: 12) – by chwilę później rozważania te

skonkludować stwierdzeniem, iż „następstwa, jakie ta [probabilistyczna – P.J.] perspektywa wywiera na wielkość błędów wypaczenia estymatorów średnich z próby, parametrów wielkości całkowitych, różnic w średnich oraz estymatorów współczynników regresji, są niewielkie" (Groves in. 1998: 12). Ten krótki fragment tekstu ukazuje zatem, że R. Groves oraz M. Couper rozważali konsekwencje metodologiczne błędów braku danych w świetle koncepcji deterministycznej głównie z tego powodu, że byli przeświadczeni o niewielkiej wartości dodanej wynikającej z analizy tego typu błędu w perspektywie stochastycznej. Jak dalece stanowisko takie było błędne, wykażą analizy w dalszej części tego rozdziału. Zacząć należy od charakterystyki założeń stojących u podstaw probabilistycznego modelu błędu braku danych.

W paradygmacie probabilistycznym zakłada się, że uczestnictwo w badaniu osób wylosowanych do próby jest procesem dającym się opisać w świetle teorii (quasi)randomizacji (por. Oh i in. 1983). C-E. Särndal oraz B. Svenson (1987: 279–294) uznali nawet, że na niedostępność pewnych jednostek próby można patrzeć jak na losowanie dwustopniowe. Podobnie bowiem, jak każda jednostka ma określone (i niezerowe) szanse wylosowania do próby z dobranego wcześniej zespołu osób (na przykład z gospodarstwa domowego), tak każda wylosowana jednostka charakteryzuje się pewną skłonnością do tego, by wziąć udział w badaniu, tj. by być respondentem. Podstawowym problemem w patrzeniu na niepełną realizację próby jak na losowanie dwustopniowe jest to, że o ile w schemacie doboru z operatu zespołowego prawdopodobieństwa selekcji są badaczowi znane na każdym etapie losowania próby, o tyle skłonność (szansa) wylosowanych jednostek do udziału w badaniu jest nieznana i można ją co najwyżej oszacować na podstawie tak zwanych zmiennych pomocniczych, tj. danych zewnętrznych (na przykład pochodzących z operatu) lub informacji o respondentach i osobach nieprzebadanych zebranych w trakcie terenowej realizacji badań. W modelu probabilistycznym przyjmuje się zatem, że każda jednostka ma pewną określoną – choć nieznaną – skłonność (oznaczaną jako ϕ) do bycia respondentem lub jednostką niedostępną[224]. Ponadto zakłada się nie-

[224] W zasadzie należałoby tu również przyjąć, że skłonność jednostek do udziału w badaniu składa się z komponentu związanego z szansą nawiązania kontaktu (dostępnością jednostek próby) oraz z prawdopodobieństwem przeprowadzenia wywiadu (gotowością wylosowanych osób do kooperacji). Zresztą wskazano już wcześniej, że mechanizm oddziałujący na dostępność jednostek próby ma odrębny charakter od tego, który kształtuje gotowość wylosowanych osób do współpracy z ankieterem (por. Goyder 1987: 80; Groves i in. 1998: 47; Lynn i in. 2002: 146). Jeżeli zatem jest tak, że te dwa wymiary są od siebie niezależne, to skłonność do udziału w badaniu (na której opiera się probabilistyczny paradygmat błędu braku danych) jest *de facto* wypadkową prawdopodobieństwa nawiązania kontaktu oraz przeprowadzenia wywiadu. Dla przejrzystości prowadzonych tu rozważań w dalszej części tego rozdziału pojęcie skłonności wykorzystywane będzie w jego wersji ogólnej, a rozróżnienie na komponent dostępności oraz gotowości do kooperacji przywołane będzie ponownie w części empirycznej tego rozdziału.

zerowość szans udziału w badaniu[225], tj. że dla dowolnej *i*-tej jednostki populacji *N*-elementowej spełnione są nierówności $0 < \phi_i \leq 1$.[226]

Przyglądając się studiom teoretycznym i empirycznym poświęconym zagadnieniom niepełnej realizacji próby badawczej rozważanym w duchu założeń paradygmatu probabilistycznego, można zauważyć, iż w wielu takich opracowaniach (por. na przykład Stoop i in. 2009: 31–32; Groves 2006: 648–649; Groves i in. 2006: 722; Stoop 2005: 32–33) przywołuje się formułę błędu braku danych podaną przez Jelke G. Bethlehema (2002: 275–287) w osiemnastym rozdziale pracy zbiorowej *Survey Nonresponse* (por. Groves i in. 2002). Autor ten definiuje systematyczny błąd braku danych (dla estymatora średniej arytmetycznej) w postaci wyrażenia:

(V.1.) $\mathrm{B}(\bar{y}^*_{HT}) \approx \bar{\phi}^{-1}\mathrm{Cov}(\phi, Y)$ (por. Bethlehem 2002: 276),

gdzie $\bar{\phi}$ jest przeciętnym prawdopodobieństwem przeprowadzenia pomiaru z jednostkami populacji wylosowanymi do próby, natomiast $\mathrm{Cov}(\phi, Y)$ oznacza kowariancję pomiędzy skłonnością do udziału w badaniu oraz rozpatrywaną zmienną Y.

Formuła ta uwypukla różnice w deterministycznym oraz probabilistycznym sposobie patrzenia na błąd niepełnej realizacji próby badawczej. O ile bowiem w paradygmacie deterministycznym wielkość błędu zależy przede wszystkim od różnicy wartości estymatorów w warstwie respondentów oraz jednostek niedostępnych, o tyle w modelu stochastycznym uwarunkowana jest już poziomem kowariancji pomiędzy szansą jednostek na udział w badaniu oraz wartościami rozpatrywanej zmiennej. Przybliżenie (V.1.) ukazuje więc, że niepełna realizacja próby badawczej przełoży się w sposób znaczący na błąd braku danych, jeśli tylko prawdopodobieństwa realizacji wywiadu będą skorelowane

---

[225] Widać tutaj wyraźną odmienność w założeniach modelu probabilistycznego oraz deterministycznego. W modelu deterministycznym przyjmowano bowiem, że błąd braku danych pozostaje konsekwencją doboru do próby jednostki z populacyjnej warstwy osób niedostępnych, tj. takich, dla których ϕ_i=0. Na określenie takich jednostek używa się czasami w literaturze metodologicznej zwrotów: „stale odmawiający" (oryg. *permanent refusers*) (por. Brick 2013: 339; Stec i in. 1999: 923), „uporczywie odmawiający" (oryg. *persistant refusers*) (por. Stoop 2004: 30–38), „stale spoza grupy respondentów" (oryg. *permanent nonrespondents*) (por. Mabli 2012: 187–213; Inkmann 2010: 384–402; Atrostic i in. 2001: 224), czy też „uporczywi nie-respondenci" (oryg. *persistent nonrespondents*) (por. Brick i in 2009: 173–174; Schouten i in 2009: 109–110; Kaldjian 2004: 500). Brick i in. (2009: 173) proponują przy tym, aby zjawisko niedostępności rozpatrywać w oparciu o model hybrydowy, to znaczy dla tych wszystkich jednostek, dla których ϕ_i = 0, stosować paradygmat deterministyczny (traktować je tak, jak jednostki niepokryte przez operat doboru próby), z kolei dla tych osób, dla których ϕ_i>0, wykorzystać model probabilistyczny.

[226] Przyjęcie tego ostatniego założenia daje – przynajmniej teoretycznie – szanse na przeprowadzenie estymacji w sposób nieobciążony. Wystarczy zastosować estymatory postaci Horwitza-Thomsona (a w zasadzie pewne ich modyfikacje), w których wymagana jest znajomość prawdopodobieństw udziału jednostek w badaniu. Będzie o tym mowa w dalszej części pracy.

(liniowo) z wartościami obserwowanej zmiennej[227]. Z drugiej strony, jeśli $\mathrm{Cov}(\phi, Y) \approx 0$, to niezależnie od przeciętnych szans udziału w badaniu jednostek dobranych do próby badawczej wielkość błędu też będzie bliska zeru. Z analogiczną sytuacją miało się zresztą do czynienia w deterministycznym modelu błędu braku danych. Istotnie, jeśli różnica wielkości estymatorów w warstwie respondentów i osób niedostępnych była niewielka, to – bez względu na to, jaką część wylosowanej próby badawczej udało się zrealizować – wartość błędu systematycznego pozostawała bliska zeru. Nietrudno zresztą wykazać, iż nieobciążonym estymatorem populacyjnej wielkości $\bar{\phi}$ jest właśnie – znany z modelu deterministycznego – ważony lub nieważony współczynnik realizacji próby badawczej.

W przywołanym artykule Bethlehema (2002: 276) odnaleźć można także informację o tym, że wzór (V.1.) określa wielkość systematycznego błędu niepełnej realizacji próby badawczej dla zwykłego nieważonego estymatora parametru średniej arytmetycznej. Stwierdzenie takie powtarzają następnie w swoich pracach I. Stoop (2005: 32) oraz R. Groves i in. (2006: 722). W rzeczywistości jednak konstatacja taka jest nieprawdziwa, zaś podany błąd systematyczny odpowiada pewnej szczególnej modyfikacji klasycznego estymatora Horvitza-Thomsona. Wczytując się dokładniej w opracowanie Bethlehema (2002: 276), można zauważyć, iż autor ten odwołuje się w tej pracy do swojego wcześniejszego artykułu z 1988 roku, w którym wyprowadził wielkość (V.1.). Zasadne wydaje się odniesienie do tego tekstu celem doprecyzowania tego, jaka w rzeczywistości jest postać estymatora, dla którego błąd braku danych wyraża przybliżenie (V.1.).

Warto wpierw przypomnieć ustalenia poczynione w III oraz IV rozdziale tej pracy, z których wynika, iż w sytuacji losowania jednostek z nierównymi prawdopodobieństwami selekcji (oznaczanymi jako $\pi_1, \pi_2, \ldots, \pi_N$), estymator szacowanego parametru (np. średniej) będzie nieobciążony, jeśli tylko przyjmie postać zaproponowaną przez Horvitza-Thomsona (1952), tj. gdy będzie równy wyrażeniu:

(V.2.) $$\bar{y}_{HT} = \frac{1}{N}\sum_{i=1}^{N}\frac{y_i s_i}{\pi_i},$$

[227] Podobnie jak w modelu deterministycznym, tak i tutaj można wskazać, że znak (kierunek) kowariancji pomiędzy szansami nawiązania kontaktu z respondentem i wartościami badanej zmiennej, może być odmienny od znaku (kierunku) kowariancji pomiędzy szansą realizacji wywiadów oraz wartościami pomiaru. W takiej sytuacji wypaczenie będącego konsekwencją niemożliwości nawiązania kontaktu z wylosowaną do próby osobą może być odwrotne od wypaczenia wynikającego z braku gotowości respondenta do kooperacji. Taki przypadek przełoży się na niewielką wartość błędu niepełnej realizacji próby. W niezwykle pouczającym opracowaniu J.M. Brick oraz J.M. Montaquila (2009: 173) zwracają uwagę na fakt, iż najczęściej ma się jednak do czynienia z sytuacją, w której kierunek kowariancji pomiędzy prawdopodobieństwem dostępności oraz wartościami pomiaru będzie taki sam, jak kierunek kowariancji pomiędzy skłonnością do kooperacji i wartościami zmiennej, co skutkować będzie znacznym błędem systematycznym.

gdzie dla dowolnego $i \in \{1,2,\dots,N\}$, $s_i = 1$, jeśli tylko i-ta jednostka populacji została wylosowana do próby oraz odpowiednio, $s_i = 0$, jeżeli i-ta jednostka w próbie się nie znalazła, przy czym $\mathrm{E}(s_i = 1) = \pi_i$.

Jedną z konsekwencji niepełnej realizacji próby badawczej będzie to, że estymator Horvitza-Thomsona postaci (V.2.) przestanie być (w większości przypadków) nieobciążony. Bierze się w nim, co prawda, pod uwagę prawdopodobieństwo selekcji, ale już nie szansę realizacji wywiadu. Takim nieobciążonym oszacowaniem parametru średniej byłby, mimo wszystko, estymator postaci[228]:

(V.3.) $$\bar{y}_{HT} = \frac{1}{N}\sum_{i=1}^{N}\frac{y_i r_i}{\pi_i \phi_i}$$ (por. Bethlehem 1988: 253).

Ponieważ jednak szanse udziału w badaniu (czyli wielkości ϕ_i) pozostają niewiadome, to w praktyce nie da się takiego nieobciążonego oszacowania przeprowadzić. J.G. Bethlehem (1988: 254) zaproponował więc, aby każde ϕ_i zastąpić (nieobciążonym) estymatorem parametru $\bar{\phi}$ – a zatem – by każdej jednostce przypisać wartość oszacowania przeciętnej skłonności wszystkich jednostek populacji do udziału w badaniu. Przyjmuje ono postać $\bar{r}_{HT} = \frac{1}{N}\sum_{i=1}^{N} r_i \pi_i^{-1}$ (przy czym $\mathrm{E}(\bar{r}_{HT}) = \bar{\phi}$) i jest – doskonale znanym – ważonym wskaźnikiem realizacji próby badawczej (por. *AAPOR: Standard Definitions…* 2011: 50). Estymator (V.3.) można więc zastąpić jego zmodyfikowaną wersją[229]:

(V.3'.) $$\bar{y}^{*}_{HT} = \frac{1}{N}\sum_{i=1}^{N}\frac{y_i r_i}{\pi_i \bar{r}_{HT}}$$ (Bethlehem 1988: 254),

której obciążenie wyraża właśnie formuła (V.1.). Innymi słowy, przywoływana za Bethlehemem (2002: 276) definicja błędu systematycznego w modelu probabilistycznym nie jest (wbrew sugestiom tego autora wyrażonym w artykule z 2002 roku), wypaczeniem nieważonego estymatora średniej, lecz obciążeniem zmodyfikowanej wersji estymatora Horvitza-Thomsona parametru średniej.

Co oczywiste, sposób wyznaczania wielkości błędu zależy nie tylko od wykorzystanego estymatora, ale też od szacowanego parametru. Przykłady definicji błędu braku danych dla innych parametrów odnaleźć można w opracowaniu

---

[228] Ponieważ w modelu probabilistycznym przyjmuje się założenie, iż każdą jednostkę populacji charakteryzuje pewna niezerowa skłonność do udziału w badaniu (tj. że quasi-prawdopodobieństwo realizacji wywiadu mieścić się będzie w przedziale $0 < \phi_i \leq 1$), to szansa na to, że daną jednostkę populacji wylosuje się do próby badawczej oraz uda się z nią przeprowadzić pomiar (zdarzenia te są niezależne), można zapisać w postaci iloczynu π_i oraz ϕ_i. Ujmując to w sposób formalny $\mathrm{E}(r_i = 1 | s_i = 1) = \pi_i \phi_i$, przy czym dla dowolnej i-tej jednostki populacji, $r_i = 1$ jeśli jednostka została wylosowana do próby (tzn. $s_i = 1$) oraz uczestniczyła w badaniu, z kolei dla pozostałych osób w próbie $r_i = 0$. Z uwagi na niepełną realizację próby badawczej spełniona jest nierówność $\sum_{i=1}^{N} r_i \leq \sum_{i=1}^{N} s_i$, która oznacza, iż liczebność próby zrealizowanej jest zawsze mniejsza lub – w idealnym przypadku – równa liczności próby wylosowanej.

[229] Przy pełnej realizacji próby będzie ona równoważna (V.2.).

Bethlehema (1988: 257), w którym podana została formuła błędu dla estymatora średniej arytmetycznej w schemacie losowania warstwowego[230], pracy R. Grovesa i in. (2004: 27), w której scharakteryzowano błąd dla oszacowania populacyjnej różnicy dwóch średnich arytmetycznych[231], artykule J.M. Bricka oraz M.E. Jonesa (2008: 51–73), w którym podano definicje błędów braku danych dla estymatorów populacyjnych wielkości całkowitych, różnic średnich oraz innych parametrów, czy też tekście J.M. Montaquila i in. (2008: 561–586), w którym sformułowano błąd braku danych dla estymatora proporcji. Wszystkich pojawiających się w literaturze metodologicznej definicji błędu nie będę tu omawiać, jednak dla zachowania porządku w strukturze pracy (w II rozdziale definiowano wielkości błędów dla dwóch parametrów, tj. średniej oraz frakcji w populacji) przywołana zostanie również formuła błędu dla zmodyfikowanego estymatora Horvitza-Thomsona służącego oszacowaniu wielkości populacyjnej proporcji. Estymator taki – oparty na pomyśle Bethlehema (1988) odnoszącym się oryginalnie do estymatora parametru średniej – przyjmuje postać:

(V.4.) $$\hat{p}^{*}_{HT} = \frac{1}{N}\sum_{i=1}^{N}\frac{y_i r_i}{\pi_i \bar{r}_{HT}}$$, gdzie:

- $y_i = 1$, jeżeli *i*–ta jednostka populacji posiada estymowaną cechę (w pozostałych przypadkach $y_i = 0$),

---

[230] W schemacie doboru próby badawczej z populacji rozwarstwionej obciążenie estymatora średniej populacyjnej (przy założeniu, że oszacowania średnich warstwowych przyjmują postać zmodyfikowanych estymatorów Horwitza–Thomsona; por. wzór (5.4.) w Bethlehem (1988: 257)), można zapisać w postaci $\mathrm{B}(\bar{y}^{*}_{PS}) \approx \sum_{j=1}^{H} h_j \bar{\phi}_j^{-1}\mathrm{Cov}(\phi_j, Y_j)$, gdzie H jest liczbą warstw w populacji, h_j jest frakcją *j*–tej warstwy w populacji, natomiast $\bar{\phi}_j^{-1}\mathrm{Cov}(\phi_j, Y_j)$ jest oszacowaniem wielkości błędu niepełnej realizacji próby badawczej dla każdej z wewnątrzwarstwowych modyfikacji estymatora H–T. Innymi słowy, w obrębie każdej warstwy błąd estymatora średniej wyznaczany jest zgodnie z formułą (V.1.). W interesującym artykule *Propensity to Response and Nonresponse Bias*, Brick i in. (2008: 51–73) uszczegóławiają problematykę błędów niepełnej realizacji próby dla estymatorów w losowaniu warstwowym. Analizy tych autorów oparte są na definicji błędu estymatora warstwowego podanej przez Bethlehema (1988: 257) i prowadzą do bardzo ważnych wniosków praktycznych. Autorzy ci pokazują bowiem, że jeżeli procedury poststratyfikacyjne wykorzystywane są przez badaczy w celu ograniczenia błędu systematycznego, to zmienną, która najlepiej nadaje się do poststratyfikacji, jest charakterystyka silnie skorelowana ze skłonnością do udziału w badaniu. Dla przykładu, gdyby udało się zidentyfikować taką zmienną, że w ramach każdej kategorii wartości tej zmiennej prawdopodobieństwa udzielenia odpowiedzi byłyby jednakowe, to dzięki poststratyfikacji udałoby się wyeliminować systematyczny błąd niepełnej realizacji próby (por. Brick i in. 2008: 55). W nieco innej postaci wniosek o podobnym wydźwięku sformułował też R. Groves (2006: 650–651) przedstawiając tak zwany *model wspólnej przyczyny*, wiążący skłonność do udziału w badaniu oraz wartości badanej zmiennej. Ustalenia te przywołane będą w dalszej części rozdziału. W tym momencie wystarczy wskazać, że jeżeli badacz jest w stanie zidentyfikować czynnik mający wpływ zarówno na prawdopodobieństwo udzielenia odpowiedzi, jak i na uzyskiwane wartości zmiennych, to kontrolując wpływ takiego czynnika (np. w poststratyfikacji), można wyeliminować błąd systematyczny wynikający z niepełnej realizacji próby badawczej.

[231] Autorzy tego artykułu wychodzą od definicji błędu niepełnej realizacji próby dla estymatora średniej podanego przez Lessler i in. (1987: 134–137).

- $r_i = 1$ jeśli jednostka została wylosowana do próby oraz uczestniczyła w badaniu (w przeciwnym razie $r_i = 0$),
- $\bar{r}_{HT}$ jest – tak jak w (V.3'.) – ważonym wskaźnikiem realizacji próby badawczej,

a jego obciążenie wynikające z niepełnej realizacji próby wyraża formuła:

$$\text{(V.5.)} \qquad \mathrm{B}(\hat{p}^*_{HT}) \approx p(1-p)(1-\lambda)(p+(1-p)\lambda)^{-1},$$

gdzie $\lambda = \bar{\phi}_2\bar{\phi}_1^{-1}$, przy czym $\bar{\phi}_1$ jest przeciętnym prawdopodobieństwem udziału w badaniu jednostek populacji ze zbioru osób posiadających cechę estymowaną w proporcji, a $\bar{\phi}_2$ jest przeciętną skłonnością do udziału w badaniu jednostek z warstwy osób nieposiadających szacowanej cechy (por. Montaquila i in. 2008: 564).

Wzór (V.5.) uwidacznia ciekawą właściwość błędu niepełnej realizacji próby badawczej dla estymatora proporcji, nietrudno bowiem zauważyć, iż dla pewnej określonej wielkości parametru p wartość błędu systematycznego uzależniona będzie wyłącznie od ilorazu przeciętnych szans udziału w badaniu jednostek należących do warstwy osób nieposiadających oraz warstwy osób posiadających estymowaną cechę. Jeśli zatem przeciętne szanse udziału w badaniu jednostek z obu warstw będą takie same (to znaczy $\lambda = 1$), to oszacowanie proporcji pozostanie nieobciążone, niezależnie od tego, jakie będą wskaźniki realizacji próby w obu warstwach. Z drugiej strony, błąd systematyczny przyjmie wartość dodatnią, jeżeli tylko $0 < \lambda < 1$ (tzn. wtedy, gdy $\bar{\phi}_1 > \bar{\phi}_2$), a wartość ujemną, jeśli $\lambda > 1$ (to znaczy, gdy $\bar{\phi}_1 < \bar{\phi}_2$)[232]. Pozostaje to całkowicie zgodne nie tyle z intuicjami, co z założeniami probabilistycznego paradygmatu błędu braku danych. Istotnie, wielkość błędu będzie zerowa (dane nie będą zniekształcone w skutek niepełnej realizacji próby), jeżeli tylko skłonność do udziału w badaniu wśród osób posiadających oraz nieposiadających właściwości estymowanej cechy będzie taka sama w obu zbiorach jednostek. W takiej sytuacji szansa realizacji wywiadu nie będzie powiązana z wartościami badanej zmiennej, stąd błąd systematyczny będący konsekwencją niepełnej realizacji próby pozostanie zerowy. Jeśli jednak zależność taka wystąpi, to znaczy osoby posiadające szacowaną charakterystykę będą miały większą (lub mniejszą)

---

[232] Taki sposób definiowania błędu pociąga za sobą niezwykle ciekawe implikacje praktyczne. J.M. Montaquila i in. (2008: 565–566) ukazują, że działania podejmowane w celu zwiększenia wskaźnika realizacji próby (frakcji respondentów w próbie wylosowanej) nie muszą wcale skutkować obniżeniem wartości błędu systematycznego. Otóż, takie dodatkowe przedsięwzięcia pozwolą zmniejszyć wielkość błędu systematycznego estymatora parametru proporcji, jeśli tylko prowadzić będą do zrównania $\bar{\phi}_1$ oraz $\bar{\phi}_2$. Innymi słowy, jeżeli w wyniku zabiegów podejmowanych przez badacza na rzecz maksymalizacji wskaźnika realizacji próby dysproporcje pomiędzy charakterystykami $\bar{\phi}_1$ i $\bar{\phi}_2$ będą się zwiększać, to wraz z większą frakcją respondentów zwiększy się też wielkość błędu. Problematyka zależności pomiędzy wskaźnikiem realizacji próby a wielkością systematycznego błędu braku danych podjęta będzie w dalszej części tego rozdziału.

skłonność do bycia respondentem od osób, które badanej charakterystyki nie posiadają, to wielkość błędu będzie znacząca i przyjmie odpowiednio wartość dodatnią (lub ujemną)[233]. W tym pierwszym przypadku wielkość populacyjnej proporcji ustalona na podstawie pomiaru będzie przeszacowana w stosunku do jej wartości „prawdziwej", w drugim natomiast, pozostanie niedoszacowana.

V.3.1. Modele błędu braków danych w (probabilistycznej) koncepcji R. Grovesa

Analiza błędu niepełnej realizacji próby badawczej, prowadzona w duchu założeń paradygmatu probabilistycznego, pozwala spojrzeć na błędy braków danych oraz na działania podejmowane w trakcie realizacji próby z zupełnie innej perspektywy, niż pozwalał na to model deterministyczny. I. Stoop i in. (2009: 33) wyróżnili przy tym trzy główne typy strategii działania na rzecz ograniczania negatywnych skutków niepełnej realizacji sondażowej próby badawczej:

> Pierwsza strategia polega na dążeniu do jak największych wskaźników realizacji próby. […] [B]ardziej właściwa może być druga strategia, […] która polega na dążeniu do tego, aby w kluczowych warstwach próby […] wylosowane jednostki miały jednakowe szanse udziału w badaniu […]. Strategia taka […] polega na dążeniu do wysokiego wskaźnika realizacji próby w sposób przemyślany […]. Trzecia strategia […] polega na minimalizowaniu błędu braku danych poprzez zamianę nielosowego charakteru niepełnej realizacji próby na wzorzec losowy względem pewnych zmiennych, to znaczy poprzez 'odnajdywanie' zmiennych powiązanych z mechanizmem niedostępności jednostek. (Stoop i in. 2010: 33)

Ważne jest to, że przyjęcie perspektywy probabilistycznej ma przełożenie na praktykę badawczą i na ocenę efektywności pewnych przedsięwzięć podejmowanych na rzecz zmniejszania frakcji jednostek niedostępnych oraz ograniczania wielkości błędów systematycznych. Priorytetem przestaje być bezwzględne dążenie do maksymalizacji wskaźnika realizacji próby[234], a na

---

[233] Przykładem estymatora, dla którego wielkość błędu systematycznego wynikającego z niepełnej realizacji próby będzie dodatnia, może być oszacowanie frekwencji wyborczej (por. Jabkowski 2011: 46). Nadreprezentacja (w próbie) osób uczestniczących w wyborach wynika z faktu, że skłonność do kooperacji respondenta z ankieterem jest uwarunkowana zainteresowaniem sprawami publicznymi (por. Billiet i in. 2009: 17) oraz poczuciem obywatelskiego obowiązku (por. Domański 2006: 34), co jednocześnie warunkuje też gotowość do udział w wyborach (por. CBOS 2010: 9). Mówiąc inaczej, ponieważ większą skłonnością do udziału w badaniu charakteryzują się osoby bardziej zainteresowane sprawami polityki, a te częściej uczestniczą w wyborach, to niepełna realizacja próby skutkować będzie przeszacowaniem rzeczywistej wielkości parametru populacyjnego (w sensie statystycznym jest to konsekwencja tego, że $\bar{\phi}_1 > \bar{\phi}_2$).

[234] W literaturze badań reprezentatywnych przedstawiono wiele metod pozwalających na zwiększenie odsetka realizacji sondażowej próby badawczej (por. na przykład Stoop i in. 2010:

znaczeniu zyskują działania dające większe szanse przeprowadzenia wywiadu z jednostkami lub kategoriami osób, których skłonność do udziału w badaniu jest mniejsza niż innych, lub ujmując to inaczej, takie działania, które pozwolą na uzyskanie bardziej „zbalansowanego" zbioru odpowiedzi (por. Schouten i in. 2009: 112). O ile zatem maksymalizacja poziomu realizacji próby uzasadniona jest w koncepcji deterministycznej możliwością ograniczenia maksymalnej wielkość błędu systematycznego (por. Stoop i in. 2010: 33), o tyle w perspektywie probabilistycznej ma uzasadnienie, jeśli tylko pozwala zmniejszyć wariancję w zbiorze jednostkowych skłonności do udziału w badaniu lub ograniczyć poziom kowariancji pomiędzy wartością analizowanej zmiennej a skłonnością jednostek wylosowanej próby do udziału w badaniu. W niezwykle trafny sposób wyraził to R. Groves w artykule *Nonresponse Rates and Nonresponse Bias in Household Surveys*, stwierdzając, że:

> gdyby schemat realizacji badania mógł w jakiś sposób prowadzić do tego, że dla wszystkich zmiennych kowariancja [pomiędzy zmienną a szansami przeprowadzenia pomiaru – P.J.] byłaby równa zeru, to błąd wypaczenia danych mógłby zostać wyeliminowany. (Groves 2006: 650)

Ciekawym fragmentem artykułu R. Grovesa są analizy poświęcone weryfikacji hipotezy o braku zależności pomiędzy wielkością błędu braku danych oraz poziomem realizacji próby. W części wstępnej tego opracowania odnaleźć można odniesienie do trzech znanych studiów – Curtin i in. (2000: 413–428), Keeter i in. (2000: 125–148) oraz Merkle i in. (2002: 243–258) – których autorzy

---

142–203; Brick i in. 2009: 166–167; Weisberg 2005: 176–187; Groves i in. 2004: 189–195; Groves i in. 1998: 271–293; Groves 1989: 208–218). Polegają one najczęściej na: (1) wydłużaniu okresu przewidzianego na terenową fazę badania, (2) podejmowaniu przez ankieterów wielokrotnych prób dotarcia do wylosowanych osób, (3) przesyłaniu listów informujących o badaniu, (4) wynagradzaniu respondentów oraz (5) przeprowadzaniu konwersji odmów udziału w badaniu. Część z tych działań ma bezpośrednie przełożenie na wskaźniki kontaktu, inne z kolei oddziałują na poziom kooperacji. Dla przykładu, wydłużenie czasu przewidzianego na realizację wywiadów oraz podejmowanie wielokrotnych prób dotarcia do wylosowanych osób jest prostą metodą ograniczania frakcji jednostek niedostępnych z powodu braku kontaktu (por. Jabkowski 2011: 27–58; Sztabiński i in. 2009: 67–95). Wprawdzie działania takie pozwalają redukować wielkość systematycznego błędu braku danych (por. Kreuter i in. 2010: 880–906), to jednak mogą mieć też negatywny wpływ na trafność oraz rzetelność prowadzonego pomiaru (por. Domański 2006: 43–46). Z kolei trzy pozostałe procedury, to znaczy „nawracanie" odmawiających, przesyłanie listów zapowiednich oraz wynagradzanie respondentów za udział w badaniu, pozwalają ograniczyć frakcję jednostek niedostępnych z powodu odmowy udziału w badaniu. Wyczerpujący przegląd analiz empirycznych oraz studiów teoretycznych poświęconych skuteczności listów zapowiednich odnaleźć można w artykule P.B. Sztabińskiego (2011: 107–148). Główne ustalenia metodologów w zakresie efektywności gratyfikacji zawiera z kolei opracowanie P.B. Sztabińskiego i in. (2012: 87–122), natomiast analizy literaturowe oraz studia empiryczne dedykowane konwersji odmów przedstawione zostały w monografii Stoop i in. (2010: 162–203).

wykazali, iż działania podejmowane na rzecz zwiększenia poziomu realizacji próby nie muszą przekładać się na wielkości błędów[235]. W tym samym akapicie, w którym R. Groves odwołuje się do tych ustaleń, podaje też – niejako dla kontrastu – przykłady zgoła odmiennych rekomendacji wyrażonych *explicite* w trzech podręcznikach akademickich z zakresu metodologii badań sondażowych. Sformułowano w nich wniosek o priorytetowym znaczeniu dążenia do zminimalizowania frakcji jednostek niedostępnych, jako działania umożliwiającego osiągnięcie „zadowalającego" poziomu reprezentatywności próby (por. Alreck i in. 1995: 184; Singleton i in. 2005: 145; Babbie 2007: 262). Trzeba jednak wskazać, że takie mylne sugestie, jakoby wysokie wskaźniki realizacji próby pozwalały na eliminację błędu braku danych oraz zapewniały reprezentatywność próby na odpowiednim poziomie, pojawiają się nie tylko w podręcznikach dla studentów, ale także w metodologicznych opracowaniach naukowych. Wystarczy przy tym przywołać niedawne studium F. Sztabińskiego poświęcone jakości danych w badaniach sondażowych, w którym czytamy, że „minimalizacja błędu 'braku odpowiedzi' jest bardzo ważna, ponieważ wysoki odsetek realizacji próby zapewnia reprezentatywność badania" (Sztabiński F. 2011: 51), czy też podsumowanie artykułu opublikowanego w 2007 roku przez autora tej monografii, w którym można przeczytać, iż w tekście autor „starał się wykazać, że jedyną efektywną metodą ograniczania negatywnych skutków występowania jednostek niedostępnych jest zwiększenie odsetka realizacji założonej na wstępie próby" (Jabkowski 2007: 84). W świetle paradygmatu probabilistycznego widać, w jak niewielkim stopniu sugestie takie osadzone są w rzeczywistości empirycznej oraz jak niewielkie mają przełożenie na praktykę badawczą.

Powracając jednak do przeprowadzonej przez R. Grovesa weryfikacji hipotezy o braku empirycznych dowodów na powiązanie wartości wskaźnika realizacji próby z wielkością błędu braku danych, należy wskazać, iż autor ten wykorzystał dane o 235 różnych (nieważonych) estymatorach parametrów średnich oraz wskaźników struktury, które zamieszczone zostały w trzydziestu artykułach naukowych opublikowanych w renomowanych czasopismach metodologicznych (por. Groves 2006: 670–672)[236]. Dla potrzeb swoich analiz

---

[235] Wnioski o podobnym charakterze autor tej monografii sformułował w artykule Jabkowski (2011: 27–58). Przedmiotem tych studiów była ocena strategii ograniczania błędu niepełnej realizacji próby poprzez umożliwienie podjęcia przez ankieterów wielokrotnych prób dotarcia do wylosowanych osób. Wykazano, że choć działania takie przekładają się na wyższe wskaźniki realizacji próby, to w przypadku niektórych estymatorów nie mają bezpośredniego przełożenia na wielkość błędu niepełnej realizacji próby.

[236] W opublikowanym dwa lata później artykule R. Groves i E. Peytcheva (2008: 167–189) rozszerzyli zakres metaanalizy, skupiając uwagę na 59 artykułach oraz 959 różnych estymatorach. Analizy te dały takie same rezultaty, jakie uzyskał Groves (2006), ukazując – w najlepszym przypadku – słabą zależność pomiędzy jakością danych a wskaźnikiem realizacji próby.

R. Groves zdefiniował trzy mierniki błędów braku danych[237] oraz naniósł ich wartości na płaszczyznę euklidesową wraz z odpowiadającymi im wielkościami poziomu realizacji próby (por. ryc. 2, 3 oraz 4 w omawianym artykule). Wnioski z przeprowadzonych analiz okazały się niezwykle pouczające. Z jednej strony R. Groves wykazał, że wiele estymatorów było obciążonych znacznymi błędami systematycznymi, z drugiej natomiast, że wskaźnik realizacji próby badawczej był słabym predykatorem poziomu błędu[238]. Innymi słowy, wielkość frakcji jednostek niedostępnych oraz poziom błędu braku danych okazały się być ze sobą powiązane w stopniu co najwyżej niewielkim (por. Groves 2006: 662). Pamiętać należy jednak, że przywołane studia mają poważne ograniczenia interpretacyjne. Z faktu tego, iż w zestawieniu wyników z różnych projektów badawczych (dotyczących odmiennej tematyki, realizowanych różnymi technikami badawczymi itd.) wskaźnik realizacji próby nie przekłada się jakoś szczególnie na wielkość błędu systematycznego, nie wynika jeszcze, że w konkretnym badaniu działania zmierzające do zmniejszenia frakcji jednostek niedostępnych nie będą miały w ogóle przełożenia na wielkość błędu (por. na przykład Jabkowski 2011; Lynn i in. 2002). Wspominano już wcześniej, że procedury takie mogą mieć zarówno pozytywny, jak i negatywny wpływ na jakość badania; wszystko zależy od tego, czy prowadzą one do osłabienia, czy też do wzmocnienia poziomu kowariancji pomiędzy skłonnością do udziału w badaniu oraz uzyskiwanymi rozkładami zmiennych. W opinii R. Grovesa kluczowym wyzwaniem staje się więc:

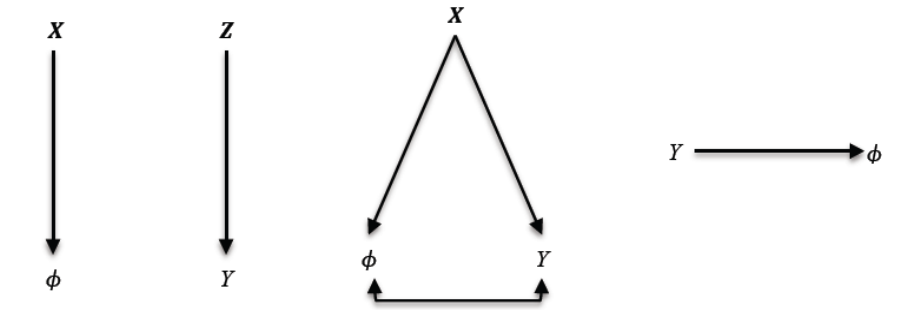
> wypracowanie modeli […] opisujących mechanizmy wiążące skłonność do udziału w badaniu z błędem braku danych. […] Muszą one [tj. owe mechanizmy – P.J.] zostać określone dla pojedynczych estymatorów, nie zaś dla całego pomiaru. (Groves 2006: 663).

Podążając za tym postulatem, R. Groves przedstawia propozycję pięciu modeli wiążących skłonność jednostek do udziału w badaniu z wartościami pomiaru oraz z błędami niepełnej realizacji próby (por. Groves 2006: 650–652).

---

[237] Umożliwiały one porównanie wielkości błędów dla estymatorów różnych parametrów oraz pomiaru przeprowadzonego odmiennymi technikami badawczymi. Pierwszy z tych współczynników zdefiniowany został jako bezwzględna relatywna wielkości błędu (tj. poziom błędu w stosunku do poziomu wartości estymatora), drugi jako różnica standaryzowanej wartości estymatora w warstwie respondentów oraz w całej próbie, z kolei definicja trzeciego współczynnika odpowiadała formule błędu w paradygmacie deterministycznym (w zasadzie była to wartość bezwzględna z tak określonego błędu) (por. Groves 2006: 659–661).

[238] Dla przykładu, wartość współczynnika korelacji pomiędzy frakcją jednostek niedostępnych oraz wielkością bezwzględnej relatywnej wielkości błędu wyniosła 0,33 jednostki. Współczynnik determinacji (tj. kwadrat miary korelacji) wyznaczony z tych danych ukazał więc, że frakcja jednostek niedostępnych pozwalała wyjaśnić niespełna 11% zmienności poziomu błędu systematycznego w analizowanym zbiorze danych (por. Groves 2006: 658–659).



Ryc. V.7. Modele przyczynowe skłonności do udziału w badaniu

Źródło: R. Groves (2006: 651)

Trzy z tych modeli odnoszą się wprost do probabilistycznej koncepcji błędu[239]. W każdym z nich autor przyjmuje odmienne założenie o związku przyczynowym pomiędzy zbiorem czynników warunkujących szanse jednostek do udziału w badaniu (ϕ), jak i zbiorem czynników mających przełożenie na uzyskiwane rozkłady zmiennych (Y). Charakter tych relacji wpływa na wielkość błędu oraz na jego specyfikę (losowość lub systematyczność), ma też przełożenie na efektywność pewnych działań badawczych oraz postbadawczych podejmowanych na rzecz wyeliminowania błędu systematycznego.

Pierwszy z przedstawionych w tej pracy układów relacji, tak zwany model niezależnej przyczyny (oryg. *seperate causes model*), opisuje pomiar, w którym zbiór czynników ***X*** warunkujących skłonność jednostek do udziału w badaniu

---

[239] W dwóch pozostałych modelach Groves (2006: 651–652) analizuje, w jaki sposób błędy pomiarowe oddziałują na błąd braku danych. Studia te wychodzą poza zakres analiz błędu prowadzonego w paradygmacie deterministycznym oraz probabilistycznym. Pierwszy z tych dodatkowych modeli, nazwany przez R. Grovesa jako *nonresponse–measurement error model*, opisuje sytuację pomiaru, w którym zachodzi związek pomiędzy obiema wielkościami, to znaczy „skłonność jednostek do udziału w badaniu determinuje wielkość błędu pomiarowego rozpatrywanej zmiennej. […] W tym przypadku wartość oczekiwana błędu pomiarowego jest niezerowa […]. Ponieważ błąd uzależniony jest od szans realizacji wywiadów, to zachodzi związek pomiędzy ϕ oraz Y" (Groves 2006: 652). R. Groves podaje przykład ciekawych badań (por. Cannel i in. 1963) obrazujących taki właśnie charakter oddziaływania błędu pomiarowego na błąd braku danych. W studiach tych pokazano, że respondenci rekrutowani we wczesnej fazie realizacji badań byli bardziej zmotywowani i przekazywali informacje dokładniej, niż osoby, wobec których musiano podjąć znacznie więcej działań, by skłonić je do udziału w projekcie. Estymatory były wypaczone z uwagi na zróżnicowane błędy pomiarowe w grupie „łatwych" oraz „trudnych" do zrekrutowania respondentów. Groves wskazuje jednak, że gdyby wielkości błędów pomiarowych były znane, wówczas błąd mógłby zostać wyeliminowany. Drugi z tych modeli (nazwany oryg. *nonresponse error attenuation model*) opisuje pomiar, w którym zachodzi bezpośrednia relacja pomiędzy prawdziwymi wartościami pomiaru a skłonnością jednostek do udziału w badaniu, jednak z uwagi na niską rzetelność przeprowadzonego pomiaru, zależność ta jest „tłumiona" wariancją błędów pomiarowych (por. Groves 2006: 652).

(ϕ) jest niezależny od czynników $\boldsymbol{Z}$ mających przełożenie na rozkłady analizowanej zmiennej Y[240]. Ponieważ w takiej sytuacji wielkość $\rho_{\phi,Y}$ jest równa zeru (obie wielkości nie są ze sobą powiązane nawet w sposób pośredni[241]), to poziom błędu systematycznego będzie równy zeru niezależnie od przeciętnej szansy realizacji wywiadu ($\bar{\phi}$) w zbiorze jednostek wylosowanych do próby badawczej. W takim wypadku niepełna realizacja próby skutkuje obniżeniem precyzji pomiaru (umniejsza wielkość próby), nie ma jednak wpływu na jego dokładność. W świetle paradygmatu deterministycznego z sytuacją opisaną w modelu niezależnej przyczyny będzie się miało do czynienia wówczas, gdy szacunki pewnych parametrów w warstwie respondentów będą zbieżne z nieznanymi oszacowaniami tych samych parametrów w warstwie jednostek niedostępnych. Wielkość błędu byłaby wówczas równa zeru i pozostała niezależna od wielkości frakcji jednostek niedostępnych w próbie badawczej.

W drugim modelu – określanym mianem *modelu wspólnej przyczyny* (oryg. *common cause model*) – miałoby się do czynienia z sytuacją, w której wystąpiłby już błąd systematyczny. Istotnie, ponieważ skłonność jednostek do udziału w badaniu (ϕ) oraz rozkłady wartości zmiennej Y posiadałyby wspólne źródło $\boldsymbol{X}$, to pomiędzy ϕ oraz Y zachodziłby związek, czyli $\rho_{\phi,Y} \neq 0$. Można jednak zauważyć, że błąd systematyczny mógłby zostać w prosty sposób wyeliminowany poprzez korygowanie danych w oparciu o wagi powiązane z $\boldsymbol{X}$ lub poprzez wykorzystanie $\boldsymbol{X}$ jako zmiennej kontrolnej. Jest to prostą konsekwencją tego, że korelacja cząstkowa uwzględniająca wpływ $\boldsymbol{X}$ na ϕ oraz Y byłaby wówczas równa zeru (tj. $\rho_{\phi,Y\cdot\boldsymbol{X}} = 0$). W świetle koncepcji deterministycznej sytuacja opisana w modelu wspólnej przyczyny odpowiada pomiarowi obciążonemu błędem systematycznym z uwagi na fakt, iż w warstwach pewnej zmiennej lub zmiennych pozostających w korelacji z charakterystyką poddawaną pomiarowi zróżnicowane są wskaźniki realizacji próby. Pomimo że w obrębie każdej warstwy błąd miałby charakter losowy, to jednak estymator szacowanego parametru zostałby wypaczony. Wielkość błędu zależałaby od międzywarstwowego zróżnicowania poziomu realizacji próby, to znaczy im byłoby ono większe, tym większe byłoby też wypaczenie pomiaru.

---

[240] R. Groves (2006) stosuje w swoim artykule odmienne oznaczenia na zbiór czynników warunkujących ϕ oraz Y. W rzeczywistości przez $\boldsymbol{Z}$ oznacza zbiór czynników warunkujących skłonność jednostek do udziału w badaniu, a przez $\boldsymbol{X}$ zbiór czynników mających przełożenie na Y. Zamiana oznaczeń wynika tutaj głównie z tego, że w literaturze metodologicznej bardziej rozpowszechnione jest oznaczanie przez $\boldsymbol{X}$, a nie przez $\boldsymbol{Z}$, czynników oddziałujących na skłonność jednostek próby do udzielenia odpowiedzi.

[241] W rzeczywistości sytuacja taka jest trudna do wyobrażenia. Wystąpiłaby ona, jeśli tylko wszystkie ϕ_i byłyby sobie równe. Co więcej, im mniejsze zróżnicowanie ϕ_i, tym mniejsza maksymalna wielkość, jaką osiągnąć może błąd niepełnej realizacji próby badawczej (por. Schouten i in. 2009: 107). Kwestie to poruszone będą w dalszej części pracy w ramach charakterystyki tzw. wskaźnika reprezentatywności zbioru odpowiedzi.

Wreszcie w modelu trzecim (oryg. *survey variable cause model*), w którym analizowana zmienna sama w sobie jest czynnikiem warunkującym skłonność jednostek do udziału w badaniu, korelacja pomiędzy ϕ i Y zachodzi poprzez prostą relację przyczynową obu charakterystyk. Co więcej, w modelu bezpośredniej przyczyny wielkość współczynnika korelacji (a zarazem wielkość błędu systematycznego) okazuje się zazwyczaj większa niż w pomiarze scharakteryzowanym przez schemat wspólnej przyczyny. Podobnie jak w poprzednim modelu, tak i tutaj udałoby się w wyeliminować błąd systematyczny, o ile tylko znane byłyby szanse realizacji wywiadów z jednostkami wylosowanymi do próby[242]. Można też wskazać, że w świetle koncepcji deterministycznej sytuacja opisana w modelu bezpośredniej przyczyny przełożyłaby się na znaczną wielkość błędu systematycznego, co byłoby efektem różnic w rozkładach zmiennej w warstwie respondentów oraz jednostek niedostępnych. Co więcej, im wskaźnik niezrealizowania próby badawczej byłby większy, tym większy byłby też błąd systematyczny.

V.3.2. Mechanizmy niedostępności jednostek w (probabilistycznej) koncepcji J.A. Little'a oraz D.B. Rubina

Scharakteryzowane przez R. Grovesa (2006) probabilistyczne modele błędu braku danych – oparte na relacjach zachodzących pomiędzy skłonnością jednostek populacji do udziału w badaniu oraz wartościami zmiennej poddawanej pomiarowi – pozostają zbieżne z niezwykle popularną w metodologicznych opracowaniach naukowych[243] koncepcją tak zwanych *mechanizmów braków danych* prowadzących do trzech typologicznych wzorców niepełnej realizacji próby badawczej, to znaczy ubytków pomiaru, które mogą być: (a) całkowicie losowe (*MCAR*), (b) losowe (*MAR*) lub (c) nielosowe (*NMAR*)[244]. Chociaż metodologowie odnoszący się do tej koncepcji przywołują ją najczęściej za Roderic-

---

[242] Wskazywano już, że w takiej sytuacji estymator szacowanego parametru musiałby przyjąć postać podobną do estymatora Horvitza-Thomsona. Każdej jednostce należałoby przypisać odwrotność jej skłonności do udziału w badaniu (wówczas korelacja cząstkowa $\rho_{\phi,Y\cdot\phi} = 0$). Niemałym wyzwaniem pozostaje oczywiście to, że szanse realizacji wywiadu pozostają nieznane (inaczej niż prawdopodobieństwa selekcji jednostek do próby) i można co najwyżej próbować je oszacować na podstawie danych empirycznych powiązanych z $\boldsymbol{X}$. Wykorzystanie oszacowań ϕ w procedurach ważenia danych opisane zostało w ostatniej części tego rozdziału.

[243] Por. na przykład Rousseau i in. (2012: 1394–1395), Pokropek (2011: 82–86), Stoop i in. (2010: 31–33), Durrant (2009: 295), Schouten i in. (2009: 102), Durrant i in. (2006: 25–36), Brick i in. (2008: 66–67), Särndal i in. (2005: 102–103), Spiess i in. (2005: 63–64), Bethlehem (1999: 110–142), Lohr (1999: 264–265), czy też Troxel i in. (1997: 857).

[244] Pomimo iż koncepcja ta odnosi się do błędów braków danych na poziomie pojedynczych zmiennych / pytań (*item nonresponse*) (por. Little i in. 1987: 3), to jednak można dokonać jej uogólnienia na błędy będące konsekwencją niezrealizowania pomiaru w całości (*unit nonresponse*).

kiem J.A. Little oraz Donaldem B. Rubinem (1987; 2002) i ich monografią *Statistical Analysis with Missing Data*, to jednak w rzeczywistości podstawowe założenia oraz ramy teoretyczne owych mechanizmów niedostępności przedstawione zostały w połowie lat 70. XX wieku przez D.B. Rubina (1976) w znanym artykule *Inference and Missing Data*.

Głównym celem analiz teoretycznych oraz metodologicznych przeprowadzonych przez Rubina (1976) było zidentyfikowanie mechanizmów prowadzących do takich wzorców niepełnej realizacji próby badawczej, które nie będą miały przełożenia na jakość wnioskowania statystycznego, to znaczy obniżą precyzję estymatorów poprzez zmniejszenie liczebności próby, ale nie doprowadzą do błędów systematycznych. Wprawdzie autor ten nie używał w swoich analizach pojęcia „błędu braku danych", to jednak – w sensie formalnym – jego intencją było opisanie takiego wzorca realizacji próby, przy którym pomiar byłby nieobciążony błędem systematycznym, lub takiego, przy którym błąd mógłby zostać wyeliminowany postrealizacyjnym ważeniem danych wynikowych[245]. Zgodnie z założeniami stojącymi u podstaw tej koncepcji ów mechanizm niedostępności traktowany jest jako proces probabilistyczny. To, z którym ze wzorców niepełnej realizacji próby będzie się miało do czynienia, zależy z kolei w równym stopniu od tego, jakimi właściwościami charakteryzuje się zbiór wartości pomiaru w próbie zrealizowanej, jak też od tego, jakie właściwości ma zbiór nieznanych wartości pomiaru (tj. braków danych) w nierealizowanej części próby. D.B. Rubin (1976: 682, 584) zdefiniował przy tym dwa pojęcia opisujące postrealizacyjne zbiory wartości badanej zmiennej. Autor ten mówi zatem o „losowości braków danych" (*MAR*) oraz o „losowości danych zaobserwowanych" (*OAR*), wskazując, że:

> braki danych są pominięte w sposób losowy [*MAR* – P.J.], jeżeli [...] prawdopodobieństwo warunkowe pojawienia się określonego wzorca braku danych względem brakujących wyników oraz wyników zaobserwowanych, jest takie samo dla każdej wartości ze zbioru braków danych. (Rubin 1976: 582),

a także, iż wartości pomiaru są zaobserwowane w sposób losowy (*OAR*):

> jeżeli dla każdej możliwej wartości brakujących danych [...] prawdopodobieństwo warunkowe pojawienia się określonego wzorca braku danych, względem brakujących wyników oraz wyników zaobserwowanych, jest takie samo dla każdej zaobserwowanej wartości. (Rubin 1976: 582)

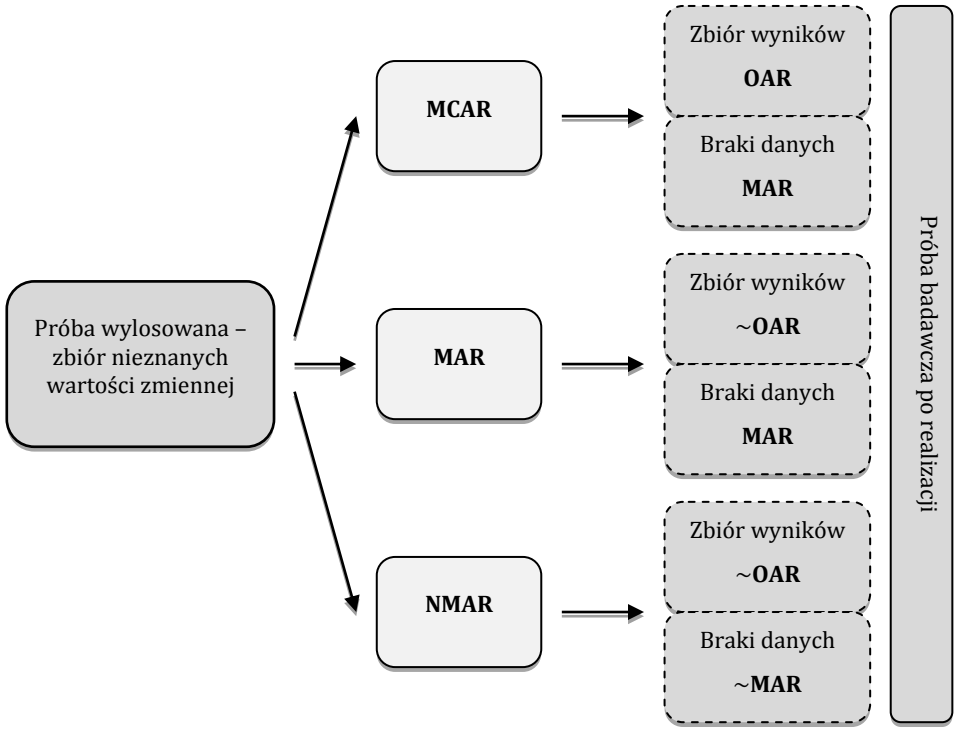
Przekładając te pojęcia na język probabilistycznej koncepcji błędu, można powiedzieć, iż braki danych będą miały charakter (*MAR*), jeśli tylko wszystkie

---

[245] W obu przypadkach D.B. Rubin mówił o mechanizmie braków danych, który jest ignorowalny (por. Rubin 1976: 584–585).

osoby niedostępne będą cechować się taką samą skłonnością do bycia jednostką dostępną (lub niedostępną), z kolei uzyskane wyniki pomiaru będą (*OAR*), jeśli tylko wszystkie osoby dostępne będą miały tę samą skłonność do tego, by być jednostką dostępną (lub niedostępną). Graficzną prezentacją tej koncepcji jest rycina V.8.[246].

Warto przy tym zauważyć, iż na wylosowaną do badania próbę badawczą można spojrzeć jak na zbiór zawierający kompletne informacje o wszystkich war-



Ryc. V.8. Mechanizm niedostępności jednostek oraz wzorce braku danych w koncepcji D.B. Rubina (1976) oraz R. Little'a i D.B. Rubina (1987)

Źródło: opracowanie własne

---

[246] W artykule *Missing by Design: Planned Missing-Data Designs in Social Science* A. Pokropek (2011: 82) przedstawia graficzną prezentację koncepcji Rubina (1976) w postaci schematu podobnego do ryc. V.8. Wykres zamieszczony przez Pokropka nie oddaje jednak w pełni istoty tego modelu. Otóż Pokropek eksponuje wyłącznie właściwości zbioru jednostek niedostępnych (mechanizm braku danych jest utożsamiony z charakterystyką zbioru jednostek niedostępnych), podczas gdy w rzeczywistości to, z jakim wzorcem realizacji próby ma się do czynienia, zależy jednakowo od charakteru danych zaobserwowanych, jak i od właściwości zbioru braku danych. W dalszej części przywołanego artykułu A. Pokropek (2011: 84–86) przedstawia pouczające przykłady charakteryzujące każdy z trzech mechanizmów niedostępności jednostek.

tościach interesujących nas zmiennych. Ważne jest, że zbiór ten obejmuje pełen zakres informacji o każdej jednostce oraz każdej analizowanej zmiennej. Braki danych pojawiają się w trakcie realizacji próby (pomiaru), a w zasadzie – zgodnie z założeniem przyjętym przez Rubina (1976) oraz Little'a i in. (1987) – są one konsekwencją oddziaływania na pomiar pewnego nieobserwowalnego mechanizmu stochastycznego (*MCAR*, *MAR* lub *NMAR*), którego rezultatem jest losowość lub nielosowość zbioru obserwacji oraz zbioru wartości niedostępnych. Innymi słowy, za realizacją próby kryje się pewien statystyczny proces determinujący charakter zrealizowanej próby badawczej (por. Pokropek 2011: 83).

Ów proces (mechanizm niedostępności) jest w rzeczywistości rozkładem zmiennej losowej opisującej szanse realizacji wywiadu z jednostkami wylosowanymi do próby, to znaczy jest prawdopodobieństwem warunkowym zdarzenia polegającego na realizacji pomiaru z jednostką populacji, o ile tylko została ona wylosowana do próby badawczej (por. Haziza 2009: 221). Tym samym w koncepcji Rubina (1976) oraz Little'a i Rubina (1987) daje się dostrzec wyraźne odniesienie do założeń przyjmowanych w probabilistycznym paradygmacie błędu niepełnej realizacji próby. Jeśli bowiem wskaźnik przeprowadzenia pomiaru zdefiniuje się jako dychotomiczną zmienną losową (o wartościach 0–1), to owo prawdopodobieństwo – o którym wspominają obaj autorzy – czyli $\phi_i \stackrel{\text{def}}{=} P(r_i = 1 | s_i = 1)$, gdzie $s_i = 1$, jeśli tylko i-ta jednostka N-elementowej populacji została wylosowana do próby badawczej – jest niczym innym, jak znaną ze stochastycznego modelu błędu (nieznaną, lecz dodatnią) skłonnością jednostek próby do udziału w badaniu[247] (por. Lohr 1999: 264). Zatem w studiach nad niepełną realizacją próby wygodnie jest klasyfikować wzorce braków danych (a w zasadzie mechanizmy prowadzące do tych wzorców) z uwagi na to, czy prawdopodobieństwo realizacji pomiaru zależy (czy też nie zależy) od wartości analizowanej zmiennej (oznaczanej jako Y) i/lub od wartości innych zmiennych uwzględnionych w badaniu (oznaczanych jako wektor zmiennych $\boldsymbol{X}$). Prawdopodobieństwo przeprowadzenia pomiaru może więc zależeć: (1) w bezpośredni sposób od Y i ewentualnie od $\boldsymbol{X}$, (2) od $\boldsymbol{X}$, ale nie od Y (w rzeczywistości chodzi tu o to, że ϕ_i jest niezależne od Y, jeśli tylko uwzględni się wpływ $\boldsymbol{X}$ na tę wielkość), a także (3) nie zależeć ani od Y, ani też od żadnej zmiennej z wektora $\boldsymbol{X}$ (por. Little i in. 1987: 14). Każda z tych sytuacji reprezentuje inny mechanizm braków danych.

J.A. Little i D.B. Rubin (1987) wskazują bowiem, że proces niedostępności jest całkowicie losowy (*MCAR*), jeżeli zajdzie przypadek (3). W sensie formal-

[247] Ponadto przyjmuje się, że prawdopodobieństwo realizacji wywiadu z jedną osobą wylosowaną do próby jest niezależne od prawdopodobieństwa realizacji pomiaru z inną jednostką (por. Haziza 2009: 221; Lohr 1999: 264).

nym oznacza to, że:

(V.5.) $P(r_i = 1|Y, \boldsymbol{X}) = P(r_i = 1)$ (por. Pokropek 2011: 83).

W takiej sytuacji brakujące wartości zmiennej *Y* będą pominięte w sposób losowy (*MAR*), podobnie jak losowo zaobserwowane będą też wartości takiej zmiennej (*OAR*). W konsekwencji – pomimo niepełnej realizacji próby – otrzymany zbiór wyników pomiaru pewnej zmiennej *Y* będzie reprezentatywny względem wartości tej zmiennej w całej, a nie tylko zrealizowanej próbie badawczej (por. Little i in. 1987: 14). Co więcej, estymator $\bar{Y}$ pozostanie nieobciążonym oszacowaniem populacyjnej wielkości średniej (por. Bethlehem 1999: 133), podobnie zresztą tak jak nieobciążone będą estymatory każdego innego parametru w populacji. Można zatem zauważyć, iż mechanizm niedostępności przyjmujący postać *MCAR* skutkuje obniżeniem precyzji estymatorów, nie prowadzi jednak do błędów systematycznych (por. Stoop i in. 2010: 31). Co więcej, w koncepcji R. Grovesa (2006) zdefiniowany przez Little'a i in. (1987) mechanizm braków całkowicie losowych odpowiadałby sytuacji opisanej w *modelu niezależnej przyczyny.* Istotnie, jeżeli skłonność jednostek do udziału w badaniu nie byłaby powiązana z mierzoną zmienną, a także z innymi zmiennymi uwzględnionymi w badaniu, to $\rho_{\phi,Y}$ byłoby równe zeru, podobnie zresztą, jak i zerowa byłaby wielkość błędu nielosowego.

Jeśli przyjrzy się teraz sytuacji (2), czyli takiemu mechanizmowi niedostępności, w którym prawdopodobieństwo realizacji wywiadu jest zależne od wartości zmiennych wektora $\boldsymbol{X}$, ale niezależne od wartości zmiennej *Y* (a w zasadzie niezależne od *Y*, jeśli tylko kontroluje się wpływ $\boldsymbol{X}$ na obserwowane wartości *Y*), to mówi się o losowym – względem wartości pewnych zmiennych z wektora $\boldsymbol{X}$ – mechanizmie braków danych (*MAR*) (por. Stoop i in. 2010: 32). Formalnie oznacza to, że:

(V.6.) $P(r_i = 1|Y, \boldsymbol{X}) = P(r_i = 1|\boldsymbol{X})$ (por. Pokropek 2011: 83).

W takim wypadku zaobserwowane wartości pomiaru nie są już *OAR*, chociaż stanowią one nadal losową podpróbę w obrębie warstw zdefiniowanych przez wartości zmiennych z wektora $\boldsymbol{X}$. Innymi słowy, wartości pomiaru przestają być reprezentacją wyników całej próby badawczej, pomimo tego, że w ramach każdej kategorii wartości $\boldsymbol{X}$, zbiór braków danych w pomiarze zmiennej *Y* jest pominięty w sposób losowy (por. Lohr 1999: 265). Odnosząc się ponownie do opracowania Grovesa (2006), można wskazać podobieństwo pomiędzy modelem wspólnej przyczyny oraz losowym wzorcem braków danych w koncepcji Little'a i in. (1987). Rzeczywiście, jeżeli przyjmie się – zgodnie z mechanizmem *MAR* – że prawdopodobieństwo realizacji wywiadu jest powiązane z wartościami zmiennych z wektora $\boldsymbol{X}$, a jednocześnie z wartościami tych zmiennych powiązane są też wartości pomiaru zmiennej *Y*, to wielkość $\rho_{\phi,Y}$

będzie w tej sytuacji różna od zera[248], co zgodnie z modelem Grovesa (2006) pozostaje konsekwencją tego, iż skłonność do udziału w badaniu oraz wartości badanej zmiennej posiadać będą wspólne źródło w postaci $\boldsymbol{X}$. Oznacza to także, iż zmienne z wektora $\boldsymbol{X}$ mogłyby zostać wykorzystane do wyeliminowania systematycznego błędu braku danych, o ile tylko znane byłyby rozkłady tych zmiennych w całej próbie badawczej (por. Bethlehem 1999: 133; Lohr 1999: 265).

Ostatni z wzorców braków danych rozpatrywany w koncepcji Little'a i in. (1987), tj. *NMAR*, odpowiada pomiarowi w którym prawdopodobieństwo realizacji wywiadu jest powiązane w sposób bezpośredni ze zmienną *Y*, a także, może być też skorelowane z wartościami zmiennych $\boldsymbol{X}$. J.A. Little i in. (1987: 14) mówią wówczas o „nieignorowalnym" mechanizmie braków danych, czyli takiej sytuacji, w której ubytki pomiaru nie są pominięte w sposób losowy, podobnie jak i w losowy sposób nie są zaobserwowane wartości pomiaru. Wzorzec niedostępności nie jest zatem ani *MCAR,* ani też *MAR*, a zatem wartości pomiaru nie stanowią losowego podzbioru całej próby badawczej ani nawet losowego podzbioru w obrębie warstw wyznaczonych przez wartości pomocniczych zmiennych $\boldsymbol{X}$ (por. Bethlehem 1999: 133). Przypadek ten można sformalizować w postaci warunku:

(V.7.) $P(r_i = 1|Y, \boldsymbol{X}) \neq P(r_i = 1|\boldsymbol{X})$ (por. Pokropek 2011: 83),

który ukazuje, iż w nielosowym mechanizmie braków danych prawdopodobieństwo realizacji pomiaru zmiennej *Y* jest bezpośrednio powiązane z wartościami tej zmiennej, nawet jeśli uwzględni się wpływ innych zmiennych $\boldsymbol{X}$. Można zatem zauważyć równoważność modelu bezpośredniej przyczyny R. Grovesa (2006) oraz nieignorowalnego mechanizmu braków danych Little'a i in. (1987). Innymi słowy, jeżeli wzorzec procesu stochastycznego prowadzącego do braków danych jest nielosowy (co w probabilistycznej koncepcji błędu oznacza systematyczne zniekształcenie wyników), to działania polegające na zwiększaniu liczebności próby badawczej lub ważeniu danych w oparciu o obserwowalne w pomiarze zmienne pomocnicze $\boldsymbol{X}$ nie rozwiązują w żaden sposób problemu systematycznego zniekształcenia pomiaru zmiennej *Y* i tym samym nie prowadzą do uzyskania bardziej dokładnych oszacowań wielkości parametrów populacyjnych.

Przedstawione wzorce niedostępności, podobnie zresztą jak i modele R. Grovesa, posłużą w dalszej części pracy do oceny efektywności procedur adjustacji danych wynikowych. Warto przy tej okazji odnieść się do niezwykle obiecującego przedsięwzięcia metodologicznego ostatnich lat, w ramach którego podjęto prace nad zdefiniowaniem – alternatywnego względem wskaźnika

[248] Nawet jeśli ϕ_i oraz *Y* nie będą od siebie zależne w sposób bezpośredni.

realizacji próby badawczej – miernika reprezentatywności sondażowej próby badawczej. Zresztą w ramach wypracowanych mierników reprezentatywności znajdują się też takie, które pozwalają ocenić, w jakim zakresie mechanizm niedostępności jednostek kryjący się za niepełną realizacją jest losowy, w jakim zaś pozostaje on nielosowy względem pewnych zmiennych zewnętrznych wobec procesu zbierania danych.

V.4. Analiza reprezentatywności sondażowej próby badawczej w świetle paradygmatu probabilistycznego

Studia teoretyczne oraz analizy empiryczne osadzone w probabilistycznym paradygmacie błędu braku danych pokazują wyraźnie, że strategia ograniczania negatywnych skutków niepełnej realizacji próby badawczej poprzez maksymalizację poziomu jej realizacji nie musi wcale prowadzić do dokładniejszych oszacowań wielkości parametrów populacyjnych. Skuteczność działań mających na celu redukcję błędu braku danych poprzez dążenie do możliwie największych wskaźników realizacji próby zależy bowiem w równym stopniu od tego, z jakim mechanizmem niedostępności ma się do czynienia, jak też od tego, jakie cechy posiadają osoby, które udało się skłonić do udziału w badaniu. Można zatem wskazać, że jeżeli mechanizm niedostępności pozostaje całkowicie losowy (to znaczy ma charakter *MCAR*), to maksymalizacja *response rate* przyczynia się do wzrostu precyzji estymatorów, nie ma jednak w ogóle – siłą rzeczy – przełożenia na wielkość błędu systematycznego (dodatkowe działania nie poprawią dokładności oszacowań, mogą jednak przyczynić się do pogorszenia jakości pomiaru, to znaczy w ich wyniku mechanizm niedostępności może już nie posiadać charakteru losowego). Po drugie, jeśli mechanizm niedostępności byłby *MAR*, a więc gdyby był on losowy względem pewnej zmiennej, to zwiększanie poziomu realizacji próby przyczyniłoby się do ograniczenia błędu braku danych tylko wtedy, gdyby dodatkowi respondenci rekrutowali się z mniej reprezentowanych warstw populacji. Dużo większe znaczenie w tym przypadku ma jednak zidentyfikowanie zmiennych, względem których mechanizm niedostępności jest *MAR*, co pozwoliłoby – przynajmniej teoretycznie – na wyeliminowanie błędu systematycznego poprzez ważenie wykorzystujące korelaty mechanizmu niedostępności. Po trzecie wreszcie, jeżeli proces niedostępności okazałby się *NMAR*, to błąd braku danych udałoby się ograniczyć tylko wtedy, gdyby respondenci pozyskani w trakcie dodatkowych działań badawczych różnili się od osób już przebadanych. W przeciwnym wypadku wypaczenie estymatorów mogłoby nawet wzrosnąć (por. Stoop i in. 2009: 33). Jak już wskazano wcześniej, skuteczność takiego sposobu postępowania jest nie-

pewna głównie z tego powodu, iż poziom realizacji próby oraz wielkość systematycznego błędu jej niepełnej realizacji pozostają ze sobą skorelowane w stopniu niewielkim. Zresztą w sposób jednoznacznie negatywny o takim sposobie ograniczania błędu braku danych wypowiedział się R. Groves, wskazując, iż „ślepa pogoń za wysokim wskaźnikiem realizacji próby jest nierozsądna; świadome dążenie do wysokiego odsetka realizacji próby jest mądre" (Groves 2006: 668).

Innymi słowy, chociaż działania na rzecz zwiększania poziomu realizacji próby pozwalają ograniczyć maksymalną wielkość systematycznych błędów braków danych, jakie w toku realizacji badań mogą się pojawić, to jednak w praktyce niższy poziom *response rate* nie przekłada się w sposób nieuchronny na większe błędy systematyczne (i odwrotnie)[249]. Zatem o wskaźniku realizacji próby badawczej można z całą stanowczością powiedzieć tyle, że nie jest on dobrym miernikiem jakości realizacji sondażowej próby badawczej[250]. Zresztą to właśnie świadomość ograniczeń kryjących się za tym wskaźnikiem doprowadziła do sformułowania postulatu mówiącego o „potrzebie opracowania […] nowego miernika jakości danych sondażowych, który pozwoliłby na lepsze rozpoznanie ryzyka błędu wypaczenia estymatorów" (Bethlehem i in. 2008: 2). Jednym z najbardziej obiecujących działań wpisujących się w ten postulat są wspomniane już prace nad tak zwanym *wskaźnikiem reprezentatywności zbioru odpowiedzi* (w skrócie *R*-wskaźnikiem) podjęte w ramach projektu

---

[249] Wskaźnik realizacji próby badawczej powiązany jest w sposób odwrotnie proporcjonalny z losowym komponentem całkowitego błędu pomiaru, czyli z wariancją estymatorów (im większy odsetek próby zrealizowanej, tym większa precyzja estymacji). Podstawowy zarzut kierowany w stronę odsetka realizacji próby wynika zatem z jego niezbyt silnego powiązania z wielkością systematycznego błędu braku danych. Co oczywiste, idealną miarą jakości pomiaru – związaną z jego niepełną realizacją – byłaby po prostu wielkość błędu systematycznego. Problemem jest to, iż rzadko kiedy wielkości takie można w ogóle wyznaczyć. Co więcej, błąd braku danych jest cechą przypisaną do konkretnej zmiennej oraz do określonego estymatora. Zatem w oparciu o wielkości błędów systematycznych trudno jest porównywać realizację prób badawczych w różnych projektach, tak jak czyni się to za pomocą wskaźnika realizacji próby.

[250] W jednym z ostatnich opracowań poświęconych wskaźnikowi reprezentatywności próby badawczej Schouten i in. (2012) wskazują na inne ograniczenia wskaźników realizacji próby. Dla przykładu, autorzy ci ukazują, że odsetek próby zrealizowanej nie pozwala na zidentyfikowanie takich kategorii jednostek, które mają decydujący wpływ na poziom reprezentatywności próby. Wprawdzie funkcje takie mogłyby spełniać warstwowe odsetki realizacji wywiadów, jednak nie biorą one pod uwagę liczebności prób badawczych w każdej warstwie. W konsekwencji mało liczne warstwy o niskich wartościach *response rate* sprawiałyby wrażenie tak samo ważnych dla zapewnienia reprezentatywności, jak warstwy bardziej liczne. Co więcej, odsetek próby zrealizowanej nie pozwala w ogóle ocenić tego, które zmienne oraz w jaki sposób oddziałują na proces niedostępności jednostek. A zatem koncentrowanie tak dużej uwagi na wskaźnikach realizacji próby badawczej wynika przede wszystkim z tego, iż wielkości tych mierników daje się udokumentować w niezwykle prosty sposób, niezależnie od tego, jaką zastosowano technikę pomiarową, czego badanie dotyczyło oraz jakie estymatory zostały wykorzystane (por. Schouten i in. 2012: 383–384).

Representativity Indicators for Survey Quality (*RISQ*)[251]. Przedsięwzięcie to jest zresztą dobrze znane w literaturze światowej dzięki serii publikacji w renomowanych periodykach naukowych (por. Schouten i in. 2009: 101–113; Schouten i in. 2011: 1–24; Schouten i in. 2012: 382–399; Shlomo i in. 2012: 201–211; Luiten i in. 2013: 169–189).

Odnosząc się do tych opracowań, należy rozpocząć od kwestii fundamentalnej dla zrozumienia idei wskaźnika reprezentatywności w ogóle, a mianowicie od wskazania, że koncepcja reprezentatywności – na której zasadza się istota *R*-wskaźnika – odnosi się wprost do założeń zgodnych z probabilistycznym paradygmatem błędu niepełnej realizacji próby badawczej. Tak więc do zdefiniowania *reprezentatywności* wykorzystuje się pojęcie *skłonności* jednostek próby do udziału w badaniu. Za punkt wyjścia przyjmuje się zatem, że:

> Podzbiór odpowiedzi jest reprezentatywny […] jeżeli jednostkowe skłonności do udzielenia odpowiedzi są takie same dla wszystkich elementów populacji (tzn. $\phi_i = P(r_i = 1|s_i = 1) = \phi$ dla $\forall_{i\in\{1.2,\ldots,N\}}$) oraz jeśli są one wzajemnie od siebie niezależne. (Schouten i in. 2009: 103)

Chociaż taki sposób ujmowania reprezentatywności napotyka na pewne znaczące ograniczenia praktyczne o charakterze ogólnym, na które uwaga zwrócona będzie nieco później, to jednak przedstawiona definicja jest niezwykle prosta i intuicyjna. Wskazuje ona, że dla zapewnienia reprezentatywności zbioru odpowiedzi nie jest w sumie ważna przeciętna skłonność do udziału w badaniu, lecz to, czy każda wylosowana jednostka ma takie same szanse realizacji z nią wywiadu.

Można to odnieść w prosty sposób do mechanizmów niedostępności jednostek oraz do modeli błędów braków danych. Otóż, jeżeli zbiór danych pomiarowych spełnia przedstawiony warunek reprezentatywności, to mechanizm niedostępności – kryjący się za niepełną realizacją próby badawczej – pozostaje całkowicie losowy w odniesieniu do pomiaru każdej zmiennej uwzględnionej w badaniu (por. Schouten i in. 2009: 103). Oznacza to też, iż szanse realizacji wywiadu nie są skorelowane z wartościami badanych zmiennych, co w ujęciu R. Grovesa (2006) odpowiada modelowi niezależnej przyczyny. Innymi słowy, w zbiorze spełniającym warunek reprezentatywności, pomiar nie jest w ogóle obciążony systematycznym błędem niepełnej realizacji próby badawczej. Z drugiej strony, nie każde odchylenie zbioru odpowiedzi od warunku pełnej reprezentatywności (tj. nie każde – nawet maksymalne – zróżnicowanie wartości ϕ_i) przekłada się w sposób nieunikniony na wypaczenie estymatorów, czy

[251] Szczegółowe informacje o projekcie *RISQ* wraz z licznymi odniesieniami bibliograficznymi odnaleźć można pod adresem internetowym http://www.risq–project.eu/.

też – analogicznie – na systematyczny mechanizm niedostępności jednostek. Co prawda odrzucenie kryterium całkowitej reprezentatywności oznacza już ryzyko pojawienia się błędu systematycznego, jednakże to, czy niepełna realizacja próby zniekształci rzeczywiście proces estymacji, czy też nie, jest uwarunkowane stopniem skorelowania skłonności do udziału w badaniu z wartościami rozpatrywanych zmiennych. Poza tym, chociaż w zbiorze danych niespełniającym warunku reprezentatywności wzorzec niedostępności nie będzie już z całą pewnością *MCAR*, to jednak może mieć on nadal charakter *MAR* lub też – po prostu – może być losowy w odniesieniu do pomiaru jakichś wybranych zmiennych[252]. Tak więc nawet znaczne zróżnicowanie indywidualnych skłonności do udziału w badaniu nie musi przełożyć się w ogóle na błąd systematyczny oraz na nielosowy mechanizm niedostępności jednostek, choć – co trzeba podkreślić – im większe będzie zróżnicowanie w zbiorze ϕ_i, tym większe będzie też ryzyko błędu systematycznego. Zresztą definicja wskaźnika reprezentatywności opiera się właśnie na mierze zróżnicowania (a dokładniej na odchyleniu standardowym) w populacyjnym zbiorze $\phi' = (\phi_1, \phi_2, \dots, \phi_N)'$ i przyjmuje postać wyrażenia:

(V.8.) $$R(\phi') = 1 - 2S(\phi')$$ (por. Schouten i in. 2009: 104).

Co więcej, ponieważ indywidualne skłonności do udziału w badaniu mieszczą się – w sensie liczbowym – w granicach (0;1], to $S(\phi') \leq \sqrt{\bar{\phi}(1-\bar{\phi})} \leq \frac{1}{2}$. Stąd wartość *R*-wskaźnika zawiera się w przedziale [0;1][253].

Wskaźnik reprezentatywności należy traktować jako miernik opisujący stopnień niepodobieństwa pomiędzy empirycznym rozkładem skłonności do udziału w badaniu w zbiorze danych wynikowych oraz teoretycznym rozkładem w zbiorze spełniającym warunek pełnej reprezentatywności. Odmienność tych rozkładów ukazuje skalę potencjalnego ryzyka pojawienia się błędu sys-

---

[252] W przywoływanym opracowaniu mowa jest zresztą o tym, że podzbiór odpowiedzi będzie reprezentatywny dla pewnej zmiennej kategorialnej przyjmującej *H* różnych wartości, jeżeli przeciętna skłonność do udzielenia odpowiedzi w warstwach wyróżnionych z uwagi na wartości takiej zmiennej będzie stała, to znaczy jeżeli dla każdego $k \in \{1,2,\dots,K\}$ spełniony będzie warunek $\bar{\phi}_k = \frac{1}{N_k}\sum_{i=1}^{N_k} \phi_{k,j} = \phi$, gdzie N_k jest liczebnością *k*-tej warstwy, natomiast $\phi_{k;i}$ skłonnością do udziału w badaniu *i*-tej jednostki z *k*-tej warstwy (por. Schouten i in. 2009: 103). Innymi słowy, chociaż wewnątrz warstw takiej zmiennej indywidualne skłonności mogą być zróżnicowane, to ważne jest, aby uśredniona skłonność w każdej warstwie była jednakowa. Wskazywano już zresztą na to uwagę w poprzedniej sekcji tego rozdziału charakteryzując wielkość błędu systematycznego dla oszacowania parametru wskaźnika struktury (por. Montaquila i in. 2008: 564). Taka „osłabiona" wersja reprezentatywności charakteryzuje więc jakość pomiaru konkretnej zmiennej, w odróżnieniu od jej wersji „mocnej" opisującej jakość pomiaru w ogóle.

[253] Przy czym 1 oznacza pełną reprezentatywność zbioru danych, natomiast 0 – maksymalne odchylenie od warunku reprezentatywności.

tematycznego na skutek niepełnej realizacji próby badawczej. Kwestie te wymagają jednak doprecyzowania oraz krótkiego komentarza. Po pierwsze, definiowanie wskaźnika reprezentatywności w oparciu o odchylenie standardowe wartości ϕ_i oznacza – w sensie analitycznym – wyznaczenie odległości euklidesowej pomiędzy jednostkowymi skłonnościami do udziału w badaniu oraz ich przeciętną wartością[254]. Po drugie, minimalna wartość, jaką miernik ten może przyjąć, zależy od średniej wartości $\bar{\phi}$ (por. Schouten i in. 2007: 10; Schouten i in. 2009: 104). Istotnie bowiem, dla $\bar{\phi} = 0{,}5$ minimalna wartość *R*-wskaźnika wynosi 0 (a zatem największe ryzyko niereprezentatywności jest wtedy, gdy przeciętna skłonność do udziału w badaniu osiąga poziom 50 pp.), natomiast im bardziej $\bar{\phi}$ wzrasta z 0,5 do 1 oraz – paradoksalnie – obniża się do 0, tym większa jest też wielkość, poniżej której nie spadnie nigdy wartość wskaźnika reprezentatywności[255]. Istotnie, ponieważ zależność pomiędzy przeciętnym prawdopodobieństwem realizacji wywiadu a wartością *R*-wskaźnika jest paraboliczna, to minimalny poziom reprezentatywności dla na przykład 90-procentowej realizacji próby będzie dokładnie taki sam, jak dla realizacji 10-procentowej[256]. Wynika to między innymi z tego, że przy małych (podobnie, jak i przy dużych) wartościach *response rate* niewiele jest miejsca na znaczne zróżnicowanie indywidualnych skłonności do udziału w badaniu. Po trzecie wreszcie, można wskazać, że *R*-wskaźnik pozwala na ocenę jakości terenowej realizacji próby badawczej niezależnie od tego, jakie są wartości pomiaru, czy też w ogóle bez względu na to, czego badanie dotyczy. Taki był zresztą zamysł autorów tej koncepcji, aby skonstruować miernik charakteryzujący jakość realizacji próby, nie zaś jakość estymacji jakichś konkretnych parametrów (por. Schouten i in 2007: 7). W konsekwencji wskaźnik reprezentatywności pozwala ocenić wyłącznie potencjalne ryzyko zniekształcenia wyników pomiaru na skutek niepełnej realizacji próby (to znaczy daje możliwość wyznaczenia maksymalnych wielkości błędów, jakie mogą się pojawić), nie pozwala natomiast wnioskować, czy przy określonym poziomie ryzyka pomiar rzeczywiście został wypaczony, jak i też, które zmienne okazały się obciążone błędami braków danych, a które pozostały od nich wolne.

---

[254] Jeżeli odległość taka wynosi 0 (wariancja jest zerowa), to zbiór danych jest w pełni reprezentatywny, bowiem z sytuacją taką będzie się miało do czynienia tylko wtedy, gdy skłonność wszystkich jednostek do udziału w badaniu będzie jednakowa.

[255] W opublikowanym niedawno artykule B. Schoutena i in. (2012) zwrócono uwagę na to, iż w praktyce udaje się w stosunkowo prosty sposób „manipulować" realizacją badań tak, aby poprzez obniżanie wskaźnika realizacji próby poprawiać poziom jej reprezentatywności. Autorzy tego opracowania zastrzegają jednak, że dążenie do poprawy wartości R-wskaźnika nie powinno nigdy opierać się na działaniach zmierzających do celowego obniżania odsetka realizacji próby w niektórych warstwach jednostek (por. Schouten i in. 2012: 393).

[256] Przy maksymalnym zróżnicowaniu skłonności do udziału w badaniu wartość *R*-wskaźnika nie będzie nigdy mniejsza niż 0,4.

Powiązanie wartości *R*-wskaźnika z poziomem błędu braku danych jest zresztą jedną z głównych zalet przypisywanych temu miernikowi jako wyznacznikowi jakości pomiaru. W literaturze poświęconej projektowi *RISQ* odnaleźć można formułę, w której wskaźnik reprezentatywności jest jednym z komponentów służących ocenie wielkości maksymalnego błędu niepełnej realizacji próby. Poziom takiego błędu, a w zasadzie jego wartość bezwzględną dla zmodyfikowanej wersji estymatora Horwitza-Thomsona parametru średniej, można wyznaczyć, przyjmując założenie o całkowitym skorelowaniu jednostkowych skłonności do udziału w badaniu z wartościami rozpatrywanych zmiennych. W takiej sytuacji górną granicą wartości bezwzględnej błędu określonego znanym wyrażeniem $\mathrm{B}(\bar{y}_{HT}^{*}) \approx \bar{\phi}^{-1}\mathrm{Cov}(\phi, Y)$ (por. Bethlehem 1988: 254), jest wielkość (por. Cobben i in. 2008: 7–9, Schouten i in. 2009: 107–108):

$$\mathrm{B}_{\max}(\bar{y}_{HT}^{*}) = \frac{(1-R(\phi\prime))S(Y)}{2\bar{\phi}}. \tag{V.9.}$$

Ponieważ przyjęty sposób oszacowania granicy błędu braku danych zakłada przypadek skrajnie niekorzystny (tzn. pełne skorelowanie ϕ oraz Y), to w rzeczywistości wielkości błędów będą znacznie mniejsze niż wskazują na to wartości maksymalne. Co więcej, choć z przyjętej w projekcie *RISQ* definicji reprezentatywności zbioru odpowiedzi wynika, iż dla zapewnienia odpowiedniego jej poziomu nie jest ważna przeciętna skłonność do udziału w badaniu (zbiór pomiaru może być w pełni reprezentatywny bez względu na to jakie jest $\bar{\phi}$)[257], to jednak $\bar{\phi}$ wpływa już w sposób znaczący na wielkość $\mathrm{B}_{\max}$. Innymi słowy, nawet wysoka wartość *R*-wskaźnika przy niskiej przeciętnej skłonności do udziału w badaniu prowadzić może do znacznych błędów systematycznych. Wobec tych faktów należy nieco zmodyfikować wyrażaną wielokrotnie w tej pracy nieufność wobec procedur terenowych mających na celu maksymalizację poziomu realizacji próby. Można bowiem wskazać, iż każde przedsięwzięcie mające na celu zwiększenie wskaźnika *response rate* wymaga rozpoznania tego, czy prowadzi ono jednocześnie do uzyskania bardziej zbalansowanej (tj. mniej zróżnicowanej z uwagi na rozkład skłonności do udziału w badaniu) realizacji próby badawczej. Jeśli tak, to działania takie są w pełni uzasadnione, jeśli zaś nie, to trzeba rozważyć, czy zysk wynikający z większej realizacji próby nie jest czasami niwelowany utratą poziomu reprezentatywności zbioru odpowiedzi. Należy zresztą dodać, że empiryczne implementacje *R*-wskaźnika wskazały przeważnie na brak zależności występującej pomiędzy odsetkiem realizacji próby a poziomem jej reprezentatywności (por. Cobben i in. 2008: 24).

Nie są to, rzecz jasna, jedyne komplikacje związane z empiryczną implementacją *R*-wskaźników. Warto zaznaczyć, że sytuacje, w których znane są popula-

[257] A zatem *response rate* – będąc oszacowaniem $\bar{\phi}$ – nie jest dobrą charakterystyką reprezentatywności próby.

cyjne wielkości indywidualnych skłonności do udziału w badaniu, stanowią raczej nieliczne wyjątki. Przyjąć należy, że szanse realizacji wywiadów należy oszacować na podstawie zestawu zmiennych zewnętrznych wobec procesu pomiaru, to jest takich danych pomocniczych, których wielkości znane są dla wszystkich jednostek w wylosowanej próbie badawczej[258]. Ma to poważne konsekwencje metodologiczne, na które warto teraz zwrócić nieco więcej uwagi. O ile bowiem w analizie reprezentatywności zbioru odpowiedzi (czy też w ogóle w studiach nad błędem braku danych definiowanym w duchu założeń probabilistycznych) interesujące są *de facto* populacyjne wielkości $\phi_i \stackrel{\text{def}}{=} P(r_i = 1|s_i = 1)$, o tyle w praktyce – dysponując zestawem zmiennych pomocniczych $\boldsymbol{X}$ – oszacować można wyłącznie wielkości $\phi_{\boldsymbol{X}}(x_i) = P\big(r_i = 1|s_i = 1, \boldsymbol{X} = x_j\big)$. Innymi słowy, w empirii daje się ocenić skłonność do udziału w badaniu jednostek należących do pewnych warstw wyróżnionych z uwagi na wartości zmiennych wchodzących w skład wektora $\boldsymbol{X}$. Nie da się tym samym uchwycić żadnych innych zróżnicowań w wielkościach ϕ_i niż te, które ulokowane są pomiędzy klasami wyznaczonymi przez kategorie wartości zmiennych pomocniczych (por. Cobben i in. 2008: 6). A zatem, jeżeli mechanizm niedostępności kryjący się za niepełną realizacją próby badawczej byłby nielosowy w obrębie klas określonych przez zmienne $\boldsymbol{X}$ (a w zasadzie, jeśli nie byłby on *MCAR* lub *MAR* względem $\boldsymbol{X}$), to *R*-wskaźnik nie byłby dobrym miernikiem jakości pomiaru, gdyż zakładałby pozorną homogeniczność wewnątrz heterogenicznego podzbioru skłonności do udziału w badaniu. Zresztą świadomość ograniczeń związanych z możliwością empirycznej weryfikacji warunku pełnej reprezentatywności doprowadziła do przyjęcia znacznie mniej restrykcyjnego kryterium stwierdzającego, iż:

> zbiór odpowiedzi jest reprezentatywny w odniesieniu do $\boldsymbol{X}$, jeżeli skłonność do udziału w badaniu w warstwach populacji wyodrębnionych z uwagi na wartości wektora zmiennych pomocniczych jest stała. (Schouten i in. 2012: 384)

Chociaż kryterium reprezentatywności zostało zawężone do klas wyznaczonych przez wartości zmiennych wektora $\boldsymbol{X}$, to jednak nadal odnosi się ono do wielkości populacyjnych, to jest do $\phi_{\boldsymbol{X}}(x_i)$. Odpowiadający tej definicji *R*-wskaźnik można zapisać przy tym w postaci formuły $R(\phi_{\boldsymbol{X}}) = 1 - 2S(\phi_{\boldsymbol{X}})$,

---

[258] W modelowaniu jednostkowych skłonności do udziału w badaniu fundamentalne znaczenie ma zatem dostęp do zmiennych pomocniczych charakteryzujących jednostki próby. W literaturze metodologicznej wyróżnia się dwie ogólne kategorie takich zmiennych. Pierwsza obejmuje te charakterystyki, których wartości dostępne są dla wszystkich jednostek populacji, druga zawiera zmienne, których jednostkowe wartość nie są już wprawdzie znane dla całej populacji, ale pozostają dostępne dla wszystkich wylosowanych osób. Dla odróżnienia obu typów zmiennych wykorzystuje się określenia *population based* oraz *sample based* (por. Kalton i in. 1986: 1–16; Brick i in. 1996: 215–238). Z kolei Lundström i in. (1999: 305–327) oraz Särndal i in. (2005: 53–56) pierwszy typ takich charakterystyk pomocniczych określali mianem *InfoU*, drugi z kolei *InfoS*.

opartej na parametrze odchylenia standardowego w zbiorze $\phi_{\boldsymbol{X}}(x_i)$. Ponieważ wielkości te są w praktyce nieznane, zupełnie tak samo jak ϕ_i, to w analizie reprezentatywności jakiejś konkretnej próby badawczej trzeba oprzeć się – mimo wszystko – na estymatorach $\hat{\phi}_{\boldsymbol{X}}(x_i)$. Warto jednocześnie wskazać, że w szacowaniu *quasi*-prawdopodobieństw udziału w badaniu wykorzystać można bogaty zestaw procedur parametrycznych (por. Brick i in. 2009: 177–178) oraz nieparametrycznych (por. da Silva i in. 2009: 165–176; da Silva i in. 2006: 563–569; da Silva i in. 2004: 45–55). Powszechnie stosowaną metodą estymacji $\phi_{\boldsymbol{X}}(x_i)$ jest szacowanie tych wielkości poprzez model regresji logistycznej (por. Lee i in. 2008: 178–179) lub jej modyfikację dedykowaną oryginalnie ekonometrycznym modelom skoringowym (por. Jabkowski i in. 2010: 59–72). Nie mniej jednak nie wypracowano żadnej uniwersalnej metody modelowania skłonności do udziału w badaniu.

Powracając do zagadnień związanych z *R*-wskaźnikiem, można zatem wskazać, że jeżeli w analizie reprezentatywności zbioru odpowiedzi badacze posługują się estymatorami jednostkowych skłonności do udziału w badaniu, to oszacowaniem parametru $S(\phi_{\boldsymbol{X}})$ jest wielkość[259] (por. Schouten i in. 2011: 8):

$$\hat{S}(\hat{\phi}_{\boldsymbol{X}}) = \sqrt{\frac{1}{N-1}\sum_{i\in s}\frac{1}{\pi_i}\left(\hat{\phi}_{\boldsymbol{X}}(x_i) - \hat{\bar{\phi}}_{\boldsymbol{X}}\right)^2}, \tag{V.10.}$$

w której π_i jest szansą doboru *i*-tej jednostki *N*-elementowej populacji do próby badawczej, natomiast $\hat{\bar{\phi}}_{\boldsymbol{X}} = \frac{1}{N}\sum_{i\in s}\frac{1}{\pi_i}\hat{\phi}_{\boldsymbol{X}}(x_i)$ jest oszacowaniem parametru $\bar{\phi}_{\boldsymbol{X}}$ odpowiadającym przeciętnej skłonności do udziału w badaniu[260].

Wynika z tego, że reprezentatywność zbioru odpowiedzi musi być zawsze odnoszona do zmiennych służących estymacji $\phi_{\boldsymbol{X}}(x_i)$. Im bardziej wyrafinowa-

---

[259] Oczywiście $\hat{S}(\hat{\phi}_{\boldsymbol{X}})$ będzie nieobciążonym estymatorem $S(\phi_{\boldsymbol{X}})$, jeśli tylko $E(\hat{\phi}_{\boldsymbol{X}}) = \phi_{\boldsymbol{X}}$.

[260] A zatem, ponieważ w empirii wskaźnik reprezentatywności oparty jest na estymatorze $\hat{S}(\hat{\phi}_{\boldsymbol{X}})$, to sam jest również estymatorem wielkości $R(\phi_{\boldsymbol{X}}) = 1 - 2S(\phi_{\boldsymbol{X}})$. W konsekwencji posiada on określoną precyzję oraz może charakteryzować się błędem systematycznym (por. Cobben i in. 2008: 9–10). Można zauważyć, że źródłem takiego błędu systematycznego jest przede wszystkim wypaczenie oszacowań $\phi_{\boldsymbol{X}}(x_i)$. Z kolei wariancja *R*-wskaźnika wynika z jednej strony z samej natury procesu próbkowania (nawet jeśli znane byłyby populacyjne wielkości $\phi_{\boldsymbol{X}}(x_i)$, to w odniesieniu do konkretnej próby badawczej $\hat{R}(\phi_{\boldsymbol{X}})$ posiadałoby określoną precyzję), z drugiej zaś z tego, że estymator *R*-wskaźnika wykorzystuje wielkości $\hat{\phi}_{\boldsymbol{X}}(x_i)$, które same posiadają już jakąś precyzję. W konsekwencji wariancja estymatora wskaźnika reprezentatywności jest pochodną próbkowania oraz wariancji oszacowań skłonności do udziału w badaniu. Oznacza to, iż błędy standardowe *R*-wskaźników nie mogą być ustalane w sposób bezpośredni. F. Cobben i in. (2008: 9–10) podają przykład wykorzystania w tym celu doskonale znanej nieparametrycznej procedury bootstrapowej (por. na przykład Mooney 2008: 65–67; Davison i in. 2007: 373–375), z kolei N. Shlomo i in. (2012: 206) proponują oszacowanie wielkości błędów standardowych poprzez linearyzację Taylora (por. na przykład Demnati i in. 2004: 17–26). Co więcej, w artykule Shlomo i in. (2012: 204–205) odnaleźć można formułę służącą ocenie wielkości błędu systematycznego estymatora $\hat{S}(\hat{\phi}_{\boldsymbol{X}})$, który można następnie uwzględnić w szacunkach wielkości wskaźnika reprezentatywności.

na będzie bowiem rozdzielczość wektora $\boldsymbol{X}$, a zmienne wchodzące w skład tego wektora okażą się lepszymi predykatorami skłonności do udziału w badaniu, tym więcej zróżnicowania da się zaobserwować w populacyjnym zbiorze ϕ_i. Jednakże wysoka wartość wskaźnika reprezentatywności i tym samym niska wariancja w zbiorze oszacowań jednostkowych skłonności do udziału w badaniu wynikać może z jednej strony z rzeczywiście wysokiej reprezentatywności pomiaru (to znaczy z losowego mechanizmu niedostępności), z drugiej zaś, z doboru takiego zestawu zmiennych do wektora $\boldsymbol{X}$, które są słabymi predykatorami indywidualnych skłonności do udziału w badaniu[261]. Istotnie, jeśli mechanizm niedostępności byłby *MAR* względem zmiennych innych niż te uwzględnione w wektorze $\boldsymbol{X}$, lub w ogóle gdyby był *NMAR*, to estymacja quasi-prawdopodobieństw udziału w badaniu w oparciu o $\boldsymbol{X}$ wskazywałaby – niezgodnie z prawdą – na reprezentatywność zbioru pomiaru. Wybór właściwego zestawu zmiennych pomocniczych służących estymacji skłonności do udziału w badaniu jest więc przedsięwzięciem fundamentalnym w analizie reprezentatywności próby, bowiem „wykorzystanie mniej złożonego modelu […] prowadzi do 'wygładzenia' oszacowań skłonności do udziału w badaniu oraz w konsekwencji do wzrostu wartości *R*-wskaźnika" (Shlomo i in. 2012: 208).

Najwłaściwszym rozwiązaniem tego problemu byłby – rzecz jasna – dobór zmiennych będących silnymi predykatorami skłonności do udziału w badaniu[262], niezwykle rzadko dysponuje się jednak zestawem zmiennych najbardziej pożądanych, znacznie częściej trzeba oprzeć się na informacjach dostępnych (to jest danych zawartych w operatach doboru próby lub zebranych w trakcie terenowej realizacji badań)[263]. Pamiętać należy przy tym, że wyko-

---

[261] W tym drugim przypadku *R*-wskaźnik charakteryzowałby się po prostu znacznym błędem systematycznym.

[262] Ale też takich, które opisują główne kategorie jednostek populacji i są silnie powiązane z kluczowymi zmiennymi poddawanymi pomiarowi (por. Schouten i in. 2011: 15). Postulaty takie sformułowali wcześniej Särndal i in. (2005: 110–111) w pracy poświęconej estymatorom kalibracyjnym. W monografii Särndal i in. (2005: 116–129) oraz artykule Särndal i in. (2008: 167–191) odnaleźć można również definicje wskaźników ułatwiających selekcję zmiennych pomocniczych w taki sposób, aby były one w największym stopniu powiązane z kluczowymi charakterystykami poddawanymi pomiarowi oraz w możliwie najlepszy sposób wyjaśniały wariancję w zbiorze skłonności wylosowanych jednostek do udziału w badaniu. Z kolei w pracy B. Schoutena (2007: 51–68) omówiona została niezwykle ciekawa metoda doboru zmiennych pomocniczych powiązana z koncepcją wskaźnika reprezentatywności. Przykłady innych procedur służących selekcji zmiennych pomocniczych odnaleźć można również w opracowaniu Rizzo i in. (1996: 43–53).

[263] Problem polega oczywiście na tym, że informacje zawarte w operatach, takie jak płeć, wiek, typ i wielkość miejscowości, czy też inne, mogą w niewielkim stopniu wyjaśniać jednostkowe skłonności do udziału w badaniu. Tym samym będą one w ograniczonym stopniu przydatne do redukcji błędu nielosowego (por. Peytcheva i in 2009: 193–201). Warto wskazać, że poza danymi z rejestrów populacyjnych opisujących jednostki próby, w wektorze zmiennych pomocniczych $\boldsymbol{X}$ zawrzeć można również charakterystyki techniki badawczej, liczbę prób nawiązania kontaktu z respondentem, obserwowalne cechy sąsiedztwa, itp. (por. Schouten i in. 2011: 4).

rzystanie *R*-wskaźnika do celów porównawczych pociąga za sobą konieczność wystandaryzowania zestawu zmiennych służących estymacji skłonności do udziału w badaniu. Nie ma bowiem sensu porównywanie wartości wskaźników reprezentatywności bazujących na różnych zestawach zmiennych pomocniczych. Problemy związane z wyborem takiego wspólnego zbioru charakterystyk są szczególnie widoczne w badaniach o charakterze międzykrajowym. Wystarczy podać przykład Europejskiego Sondażu Społecznego, by zauważyć, że wykorzystywane w poszczególnych krajach operaty doboru próby rzadko kiedy zawierają ekwiwalentne charakterystyki jednostek[264]. Zresztą w literaturze odnaleźć można postulat odnoszący się wprost do ESS-u, zakładający konieczność podjęcia działań na rzecz wypracowania standardów umożliwiających przeprowadzenie międzykrajowych porównań jakości realizacji prób badawczych opartych na takich samych charakterystykach jednostek próby (por. Schouten i in. 2012: 387).

Wybór optymalnego zestawu zmiennych pomocniczych służących oszacowaniu jednostkowych skłonności wylosowanych osób do udziału w badaniu ma decydujące znaczenie nie tylko w ocenie poziomu reprezentatywności zbioru odpowiedzi, ale także – a może przede wszystkim – w działaniach zmierzających do wyeliminowania (lub redukcji) systematycznego błędu niepełnej realizacji próby badawczej poprzez zastosowanie procedur ważenia danych. Z definicji błędu braku danych zgodnej z założeniami paradygmatu probabilistycznego wynika bowiem wyraźnie, iż w celu skutecznego ograniczenia wielkości błędu niepełnej realizacji próby należałoby skonstruować wagi w taki sposób, by były one powiązane zarówno z jednostkowymi skłonnościami do udziału w badaniu, jak też ze zmienną/zmiennymi poddawanymi pomiarowi. Problem ten charakteryzuje w sposób niezwykle trafny G. Kalton (1983), zwracając uwagę na to, że:

> spośród potencjalnych zmiennych możliwych do wykorzystania w ustalaniu wielkości wag, najbardziej efektywne w redukcji systematycznego błędu braku danych będą zmienne silnie skorelowane zarówno z badanami zmiennymi, jak też z 0–1 zmienną [opisującą mechanizm niedostępności jednostek – P.J]. (Kalton 1983: 63)

Tego typu rekomendacje stwarzają oczywiście niemałe wyzwania dla badaczy. Należy bowiem zwrócić uwagę, że o ile dość łatwo jest ustalić, które ze zmiennych pomocniczych pozostają w silnej korelacji ze zmienną opisującą udział wylosowanych osób w badaniu (por. Matsuo i in. 2010: 167; Iannacchio-

---

[264] Zupełnie bezpodstawne jest porównywanie poziomu reprezentatywności prób dobieranych z rejestrów o odmiennych poziomach agregacji jednostek (np. prób imiennych oraz adresowych).

ne 2003: 31–43; Dunkan i in. 2001: 121–130), o tyle problemem staje się już ustalenie siły zależności pomiędzy zmiennymi *Y* poddawanymi pomiarowi oraz zmiennymi zawartymi w wektorze $\boldsymbol{X}$. Wystarczy wskazać, że wielkości *Y* znane są zazwyczaj jedynie w odniesieniu do zbioru respondentów. Tym samym, na podstawie wyników w próbie uzyskuje się niepełny obraz zależności pomiędzy $\boldsymbol{X}$ oraz *Y*, który jest ograniczony wyłącznie do zrealizowanej części próby badawczej. M.J. Brick (2013) w opublikowanym niedawno artykule *Unit Nonresponse and Weighting Adjustments: A Critical Review* sugeruje wprawdzie, że „zmienne pomocnicze powiązane z kluczowymi charakterystykami poddawanymi pomiarowi powinny być uwzględnione w estymacji [skłonności do udziału w badaniu – P.J.] o ile tylko są dostępne, niezależnie od tego, czy są rzeczywiście powiązane ze skłonnościami do udziału w badaniu, czy też nie" (Brick 2013: 335), jednakże w praktyce preferuje się raczej modelowanie skłonności do udziału w badaniu oparte na zbiorze zmiennych pomocniczych silnie powiązanych z ϕ_i. Uzasadnienie dla takiego postępowania jest niezwykle proste. Otóż, ponieważ w badaniach surveyowych przedmiotem pomiaru pozostaje zazwyczaj wiele zmiennych jednocześnie, to w praktyce trudno byłoby wyróżnić zestaw zmiennych powiązanych ze wszystkimi (lub choćby z wieloma) kluczowymi zmiennymi uwzględnionymi w pomiarze. W konsekwencji, skuteczność procedur ważenia danych w redukcji błędu niepełnej realizacji próby nie jest pewna, to znaczy ich efektywność zależy od tego, czy ustalone wartości wag będą silnie skorelowane z mechanizmem niedostępności oraz z rozpatrywanymi zmiennymi[265]. Innymi słowy, ważenie danych w oparciu o $\hat{\phi}_{\boldsymbol{X}}(x_i)$ może być efektywne w redukcji błędu braku danych dla pewnych zmiennych, ale innych

[265] Jedną z ciekawszych propozycji wykorzystania ważenia danych w celu redukcji wielkości systematycznego błędu braku danych jest procedura *predicted mean stratification* (por. Little 1986: 139–157). Polega ona na modelowaniu rozkładu zmiennej *Y* w oparciu o charakterystyki pomocnicze $\boldsymbol{X}$ powiązane zarówno z *Y*, jak też z mechanizmem niedostępności, a następnie na tworzeniu warstw zgodnych z oszacowaniami *Y*. J.A. Little zestawia tę metodę z procedurą *response propensity stratification*, w której modeluje się z kolei skłonności do udziału w badaniu. Autor przywołanego opracowania zwraca uwagę, że pierwsza metoda pozwala kontrolować (ograniczać) losowy oraz systematyczny komponent błędu braku danych, podczas gdy druga może prowadzić do redukcji błędu nielosowego, ale jednocześnie charakteryzuje się tendencją do obniżania precyzji estymacji (będzie to konsekwencją wykorzystania wektora zmiennych pomocniczych $\boldsymbol{X}$ niepowiązanego z *Y*). Problem polega jednak na tym, że *predicted mean stratification* wymaga konstrukcji oddzielnych modeli dla każdej zmiennej z osobna (por. Little 1986: 147). Stąd w praktyce wygodniej jest konstruować wagi, wykorzystując zmienne powiązane wyłącznie z mechanizmem niedostępności i stosować ważenie w odniesieniu do tych zmiennych, dla których pozwala ono zredukować błąd systematyczny oraz poziom wariancji. Problemy te analizowane były już w II rozdziale pracy w ramach charakterystyki alternatywnych (względem miernika *VIF*) współczynników służących ocenie wpływu ważenia danych na jakość wyników. Przywołując wówczas ustalenia Little i in. (2005) zwrócono uwagę, że warunkiem koniecznym efektywnej redukcji błędu systematycznego jest powiązanie wag z mechanizmem niedostępności oraz z analizowaną zmienną. Prowadzi to również do redukcji wariancji estymatorów.

już nie (por. Little i in. 2005: 161–164). Zresztą studia empiryczne ukazują, iż procedury ważenia wykorzystujące oszacowania jednostkowych skłonności do udziału w badaniu umożliwiają wyeliminowanie wyłącznie części wielkości błędu systematycznego (por. Micklewright i in. 2012: 915–938).

W literaturze badań sondażowych wyróżnia się przy tym kilka sposobów wykorzystania wielkości $\hat{\phi}_{\boldsymbol{X}}(x_i)$ w procesie redukcji błędu braku danych. Pierwszy z nich – chyba najmniej rozpowszechniony w praktyce sondażowej – jest co do swojej istoty podobny do metod imputacyjnych (por. Lee i in. 2008: 179). Polega on na znalezieniu najbliższego „sąsiada" pewnej jednostki niedostępnej, czyli na wskazaniu respondenta, który ma taką samą (lub możliwie najbliższą) wartość oszacowania skłonności do udziału w badaniu, a następnie na przypisaniu jednostce niedostępnej cech takiego respondenta. Nie jest to metoda wykorzystywana w praktyce sondażowej na szeroką skalę (głównie z uwagi na dość problematyczne założenia stojące u jej podstaw), ma jednak ciekawe zastosowania w badaniach ewaluacyjnych (por. na przykład Trzciński 2009).

Pozostałe sposoby wykorzystania oszacowań skłonności do udziału w badaniu celem redukcji błędu braku danych polegają już na ich implementacji w postaci wag korygujących rozkłady analizowanych zmiennych. Warto zatem zwrócić uwagę, iż każdemu respondentowi można przyporządkować wagi w taki sposób, by były one odwrotnościami szans jego udziału w badaniu. Jest to naturalne rozszerzenie idei estymatorów Horvitza-Thomsona, w których uwzględnia się nie tylko prawdopodobieństwa doboru jednostek do próby, ale też quasi-prawdopodobieństwa udzielenia odpowiedzi (por. wzór V.3. w części V.3. tego rozdziału). W takim przypadku zastosowana waga jest iloczynem wag wynikających ze schematu losowana (*design weights*) oraz wag korygujących zróżnicowany poziom skłonności do udziału w badaniu. Przykładem praktycznych implementacji takich procedur ważenia danych są studia autorstwa Iannacchione (2003: 31–43), Kwanga i in. (2007: 501–514), Lepkowskiego i in. (1989: 296–301) oraz Rosenbauma (1987: 387–394).

Takie bezpośrednie wykorzystanie quasi-prawdopodobieństw udziału w badaniu w procesie ważenia danych napotyka jednakże na wiele problemów. W niezwykle inspirującym artykule *Survey Nonresponse Adjustments for Estimates of Mean*, J.A. Little (1986: 139–157) zauważył bowiem, że może to prowadzić do niestabilnych estymatorów szacowanych parametrów. Będzie tak, gdy niektóre oszacowania skłonności do udziału w badaniu mieć będą wartości bliskie zeru. We wszystkich takich sytuacjach respondenci o niewielkich prawdopodobieństwach udzielenia odpowiedzi mieliby przypisane bardzo duże wartości wag, co w konsekwencji mogłoby prowadzić do znacznego przyrostu wariancji pewnych estymatorów. W związku z tym J.A. Little rekomendował podział wylosowanych jednostek zgodnie z oszacowanymi wartościami skłon-

ności do udziału w badaniu, a następnie przypisanie każdej jednostce, wewnątrz takich warstw, jednakowych wartości wag, niezależnie od tego, jakie były indywidualne oszacowania prawdopodobieństw ich udziału w badaniu. W jego opinii argumentem przemawiającym na rzecz takiego podziału – poza ograniczeniem zakresu wartości wag – jest również to, że:

> w mniejszym stopniu wymaga to prawidłowej estymacji skłonności do udzielenia odpowiedzi, gdyż, inaczej niż w ważeniu bezpośrednim, oszacowania wykorzystywane są jedynie do uporządkowania jednostek próby. Zatem model regresji służący estymacji skłonności do udzielenia odpowiedzi może być wystarczający do [prawidłowego uporządkowania jednostek wg ich skłonności do udziału w badaniu – P.J.], lecz niewystarczający do zdefiniowania wag w sposób bezpośredni. (Little 1986: 146)

A zatem za tak zwanym podejściem stratyfikacyjnym w wyznaczaniu wartości wag przemawiają w opinii tego metodologa przede wszystkim względy praktyczne oraz merytoryczne. Wystarczy bowiem – używając terminologii pomiaru psychometrycznego – aby estymatory quasi-prawdopodobieństw udziału w badaniu były trafne, niekoniecznie zaś muszą być one określone w sposób dokładny, to znaczy zgodny z prawdziwymi wartościami parametrów ϕ_i w populacji.

Mimo wszystko niemałym wyzwaniem staje się również sam podział jednostek próby w oparciu o oszacowania skłonności do udziału w badaniu. Powołując się na rekomendację W. Cochrana (1968: 295–313), niektórzy badacze, na przykład Rosenbaum i in. (1984: 522), Little (1986) oraz Lee (2006: 334) proponowali utworzenie pięciu warstw zawierających jednakową liczbę jednostek, zgodnie z kwintylami rozkładu $\hat{\phi}_{\boldsymbol{X}}(x_i)$, podczas gdy część metodologów (por. Loosveldt i in. 2008: 98; Matsuo i in. 2010: 167–165) sugerowała podział próby na 10 równolicznych warstw, inni natomiast uważali, że zastosowany podział powinien mieć uzasadnienie statystyczne (por. Little i in. 2003: 1589–1599). Tak naprawdę chodzi jednak o to, aby wariancja quasi-prawdopodobieństw udziału w badaniu wewnątrz wyróżnionych kategorii jednostek była jak najmniejsza, to znaczy aby zasadnicza część zróżnicowania ϕ_i ulokowana była między warstwami, a nie wewnątrz nich. Innymi słowy najlepszy jest taki podział, w którym wewnątrzgrupowa wariancja zostaje zupełnie wyeliminowania[266]. Niezwykle interesujące sposoby ustalania wartości wag w wyróżnionych

[266] Zupełnie tak samo, jak w modelu wspólnej przyczyny Grovesa (2006), czy też w losowym względem pewnej zmiennej mechanizmie niedostępności jednostek w ujęciu Little i in. (1987). W takiej sytuacji ważenie danych byłoby najbardziej efektywne w redukcji wielkości błędu braku danych.

wcześniej kategoriach jednostek próby odnaleźć można w opracowaniach Lee i in. (2008: 170–183)[267] oraz Matsuo i in. (2010: 165–178)[268].

Powracając do analiz związanych z *R*-wskaźnikiem, należy wskazać, że jego wykorzystanie do oceny reprezentatywności próby jest tylko jednym z kilku możliwych obszarów zastosowań tego miernika w praktyce sondażowej (por. Bethlehem i in. 2008: 4–7). Co więcej, wykonanie takiej postbadawczej oceny zbioru odpowiedzi pozostaje w gruncie rzeczy dużo mniej istotne niż możliwość prowadzenia bieżącej kontroli jakości realizacji badania. Zarazem to właśnie w tym wymiarze zastosowań wskaźnika reprezentatywności ujawniają się jego kolejne ograniczenia. Można bowiem zauważyć, iż *R*-wskaźnik – jako charakterystyka poziomu reprezentatywności całej próby badawczej – nie pozwala w ogóle na wyróżnienie tych kategorii jednostek próby, których oddziaływanie na odchylenie zbioru odpowiedzi od warunku pełnej reprezentatywności jest największe. Innymi słowy, wskaźnik ten nie daje możliwości opisania tego, jaki jest brzegowy lub warunkowy wpływ pojedynczych zmiennych na zróżnicowanie w zbiorze jednostkowych skłonności do udziału w badaniu. Zresztą to właśnie na monitorowaniu procesu realizacji próby oraz wykorzystaniu w tym celu wskaźników reprezentatywności skupiały się w ostatnich latach prace podejmowane w ramach projektu *RISQ*. Rezultaty tych niezwykle interesujących działań omówione zostały w dwóch znakomitych publikacjach autorstwa Schoutena i in. (2011: 1–24) oraz Schoutena i in. (2012: 382–399).

W pracach tych zdefiniowano między innymi częściowe *R*-wskaźniki – w ich wersji brzegowej (*unconditional partial indicator*) oraz warunkowej (*conditional partial indicator*) – służące zarówno analizie wpływu pojedynczej zmiennej z wektora $\boldsymbol{X}$ na odchylenie zbioru pomiaru od warunku pełnej reprezentatywności, jak też ocenie wpływu takiej zmiennej przy równoczesnym uwzględnieniu oddziaływania wszystkich pozostałych charakterystyk z wektora zmiennych pomocniczych. Odnosząc się do przywołanych tu artykułów, należy rozpocząć od tego, że idea wskaźników częściowych opiera się na doskonale

---

[267] S. Lee i in. (2008) zaproponowali dwa sposoby ustalania wartości wag. Po pierwsze, aby w każdej warstwie wartości wag były ilorazem sumy wag *design weights* (łącznie dla respondentów i jednostek niedostępnych w danej warstwie) oraz sumy wag *design weights* w warstwie (ale tylko dla zbioru respondentów). Drugi sposób ustalania wielkości wag polega na wyznaczaniu, w każdej warstwie, przeciętnej ważonej (tzn. uwzględniającej *design weights*) odwrotności skłonności do udziału w badaniu.

[268] H. Matsuo i in. (2010) proponowali metody wyznaczania wartości wag w warstwach próby zgodne z ideą wag poststratyfikacyjnych. Waga jest ilorazem odsetka liczby respondentów i jednostek niedostępnych w danej warstwie oraz odsetka liczby respondentów znajdujących się w takiej warstwie. Dokładne wzory odnaleźć można w Matsuo i in. (2010: 168–169), którzy podają je za Lee (2006: 329–349). Warto wskazać, że S. Lee wykorzystał tę metodę oryginalnie w celu ograniczenia dwóch klas błędów związanych z brakiem obserwacji (niepełnego pokrycia oraz niepełnej realizacji próby), z kolei Matsuo i in. (2010) implementowali ją w badaniach sondażowych celem redukcji błędu braku danych.

znanej właściwości wariancji, umożliwiającej jej dekompozycję na składnik zróżnicowania ulokowany między grupami oraz wewnątrz grup wyróżnionych z uwagi na wartości jakiejś zmiennej kategorialnej[269]. A zatem dla dowolnego podziału populacji na *K* warstw przeprowadzonego względem wartości pewnej charakterystyki *Z* wchodzącej w skład wektora zmiennych pomocniczych $\boldsymbol{X}$, wariancję w zbiorze jednostkowych skłonności do udziału w badaniu można zapisać jako sumę zróżnicowania międzygrupowego oraz wewnątrzgrupowego, to znaczy jako:

(V.11.) $S^2(\phi_{\boldsymbol{X}}) = S_b^2(\phi_{\boldsymbol{X}}|Z) + S_w^2(\phi_{\boldsymbol{X}}|Z)$ (por. Schouten i in. 2011: 5),

gdzie:

(V.12.) $$S_b^2(\phi_{\boldsymbol{X}}|Z) \approx \sum_{k=1}^{K} \frac{N_k}{N}(\bar{\phi}_{\boldsymbol{X};k} - \bar{\phi}_{\boldsymbol{X}})^2$$

jest wariancją międzygrupową w zbiorze $\phi_{\boldsymbol{X}}(x_i)$, natomiast:

(V.13.) $$S_w^2(\phi_{\boldsymbol{X}}|Z) = \frac{1}{N-1}\sum_{k=1}^{K}\sum_{i=1}^{N_k}(\phi_{\boldsymbol{X}}(x_i) - \bar{\phi}_{\boldsymbol{X};k})^2$$

komponentem zróżnicowania wewnątrzgrupowego[270].

Pierwszy ze wspomnianych wskaźników częściowych, tak zwany brzegowy *R*-wskaźnik, definiowany jest przy tym jako pierwiastek z międzygrupowej wariancji w zbiorze indywidualnych skłonności jednostek do udziału w badaniu:

(V.14.) $P_u(Z, \phi_{\boldsymbol{X}}) \overset{\text{def}}{=} S_b(\phi_{\boldsymbol{X}}|Z)$ (por. Schouten i in. 2012: 385).

Im większa będzie wartość, jaką osiągnie taki wskaźnik, tym większy będzie też wpływ zmiennej *Z* zawartej w wektorze $\boldsymbol{X}$ na odchylenie zbioru pomiaru od pełnej reprezentatywności. Zatem, dążąc do zidentyfikowania czynników odpowiadających w stopniu najwyższym za niereprezentatywność zbioru odpowiedzi, należałoby wykorzystać zmienne o największych wartościach wskaźników brzegowych[271]. Co więcej, jeżeli dla jakiejś zmiennej *Z* wartość miernika

---

[269] Zresztą dokładnie tę samą charakterystykę wariancji wykorzystywano w czwartym rozdziale pracy rozważając efektywność schematu losowania warstwowego (por. wzory IV.7. oraz IV.7'.).

[270] Ponieważ $S_b^2(\phi_{\boldsymbol{X}}|Z)$ oraz $S_w^2(\phi_{\boldsymbol{X}}|Z)$ są wielkościami populacyjnymi, to w praktyce wymagają oszacowania na podstawie danych z próby. Estymatory tych parametrów odnaleźć można w artykule Schoutena i in. (2011: 8).

[271] W opracowaniu Schoutena i in. (2011: 5–6) odnaleźć można także definicję wskaźnika brzegowego odnoszącą się już nie do zmiennej *Z*, ale do każdej z jej wartości. Taki warstwowy miernik brzegowy przyjmuje postać wyrażenia $P_u(Z = k, \phi_{\boldsymbol{X}}) \overset{\text{def}}{=} \sqrt{\frac{N_k}{N}}(\bar{\phi}_{\boldsymbol{X};k} - \bar{\phi}_{\boldsymbol{X}})$, przy czym jego wartości mieszczą się w przedziale [–0,5;+0,5]. Wartości dodatnie świadczą o nadreprezentacji (wyższej od przeciętnej skłonności do udziału w badaniu) jednostek z *k*–tej warstwy, z kolei wartości ujemne o niedoreprezentacji takiej warstwy (niższej przeciętnej skłonności do udziału

brzegowego osiągnęłaby poziom wariancji całkowitej[272], to zmienna taka wyjaśniałaby w stopniu zupełnym zróżnicowanie w zbiorze skłonności jednostek próby do udziału w badaniu[273]. Ponieważ w takiej sytuacji zmienność w zbiorze $\phi_{\boldsymbol{X}}$ ulokowana byłaby wyłącznie pomiędzy warstwami Z (zróżnicowanie wewnątrzgrupowe wyniosłoby zero), to mechanizm niedostępności jednostek byłby *MAR* względem zmiennej Z. A zatem wskaźnik reprezentatywności oraz brzegowy wskaźnik częściowy pozwala na empiryczną weryfikację mechanizmu niedostępności jednostek rozumianą zgodnie z koncepcją Little'a i in. (1987). O ile bowiem $S(\phi_{\boldsymbol{X}})$ opisuje stopień, w jakim mechanizm niedostępności odstaje od wzorca całkowicie losowego dla wszystkich zmiennych, o tyle $S_b(\phi_{\boldsymbol{X}}|Z)$ charakteryzuje już skalę w jakiej proces niedostępności różni się od wzorca *MCAR* dla pojedynczej zmiennej Z, natomiast $S_w(\phi_{\boldsymbol{X}}|Z)$ opisuje stopień, w jakim mechanizm ten nie jest *MAR*[274] względem zmiennej Z (por. Schouten i in. 2012: 385–386).

Wprowadzenie drugiego ze wskaźników częściowych wymaga wcześniejszego przedstawienia koncepcji warunkowej reprezentatywności zbioru odpowiedzi. W sposób najbardziej precyzyjny sens tego, czym dla pewnej zmiennej Z jest taka warunkowa reprezentatywność próby, oddaje definicja zamieszczona w tekście Schoutena i in. (2011). Autorzy tego artykułu wskazują bowiem, że:

> zbiór odpowiedzi nazywany jest warunkowo reprezentatywnym dla Z względem $\boldsymbol{X}^{-}$, jeżeli warunkowa skłonność do udziału w badaniu jest taka sama dla wszystkich L warstw wyróżnionych z uwagi na $\boldsymbol{X}^{-}$ (Schouten i in. 2011: 6).

W definicji tej pojawia się wektor $\boldsymbol{X}^{-}$ opisujący zbiór wszystkich zmiennych wchodzących w skład wektora $\boldsymbol{X}$ z wyłączeniem jednak Z, to znaczy $\boldsymbol{X} = (\boldsymbol{X}^{-}; Z)$. Ważne jest przede wszystkim to, że jeżeli zbiór odpowiedzi jest

---

w badaniu). Wskaźniki częściowe są zatem przydatne w procesie monitorowania realizacji próby. Ujemne wartości $P_u(Z = k, \phi_{\boldsymbol{X}})$ wskazują na takie kategorie jednostek, które wymagają dodatkowych działań badawczych na rzecz zwiększenia szans realizacji wywiadów z jednostkami z tych warstw.

[272] Jest to zresztą wartość maksymalna, której wskaźnik brzegowy nigdy nie przekroczy. A zatem $P_u(Z, \phi_{\boldsymbol{X}}) \in [0; S(\phi_{\boldsymbol{X}})]$.

[273] Pamiętać należy przy tym, że w odniesieniu do konkretnej próby można dokonać takiego podziału populacji, w którym wewnątrzwarstwowe wariancje skłonności do udziału w badaniu będą zawsze zerowe. Wystarczy przeprowadzić dekompozycję względem wartości wektora zmiennych pomocniczych $\boldsymbol{X}$ służących estymacji skłonności jednostek próby do udziału w badaniu. Częściowe wskaźniki reprezentatywności charakteryzują się zatem tymi wszystkimi ograniczeniami interpretacyjnymi, które wcześniej przypisywane zostały R–wskaźnikowi w jego postaci ogólnej. Jeśli bowiem rzeczywisty mechanizm niedostępności będzie nielosowy lub losowy względem innych zmiennych niż te uwzględnione w $\boldsymbol{X}$, to wskaźniki częściowe nie będą dobrymi charakterystykami jakości realizacji próby

[274] Ujmując to nieco inaczej: $S_w^2(\phi_{\boldsymbol{X}}|Z)$ charakteryzuje stopień, w jakim mechanizm niedostępności jest *MNAR* dla Z.

warunkowo reprezentatywny dla Z, to skłonności jednostek próby do udziału w badaniu ustalane względem $\boldsymbol{X}$ są dokładnie takie same, jak te odnoszone do zmiennych $\boldsymbol{X}^{-}$. Sytuacja warunkowej reprezentatywności dla charakterystyki Z oznacza bowiem, iż po skontrolowaniu wpływu $\boldsymbol{X}^{-}$ na $\phi_{\boldsymbol{X}}$, zmienna Z nie ma już żadnego przełożenia na niereprezentatywność próby i można ją pominąć w analizach $\phi_{\boldsymbol{X}}$. A zatem warunkowy R-wskaźnik jest miarą wpływu zmiennej Z na $\phi_{\boldsymbol{X}}$ po skontrolowaniu oddziaływania $\boldsymbol{X}^{-}$ na te wielkości. Oparty jest on przy tym na wariancji wewnątrzgrupowej[275] w populacyjnym zbiorze $\phi_{\boldsymbol{X}}$ przyjmując postać wyrażenia:

(V.15.) $$P_c(Z, \phi_{\boldsymbol{X}}) \stackrel{\text{def}}{=} S_b(\phi_{\boldsymbol{X}}|\boldsymbol{X}^{-})$$ (por. Schouten i in. 2012: 385).

Im większa będzie wartość, jaką osiągnie $P_c(Z, \phi_{\boldsymbol{X}})$[276], tym większa część wariancji w zbiorze skłonności do udziału w badaniu ulokowana będzie wewnątrz warstw zdefiniowanych przez $\boldsymbol{X}^{-}$. Z drugiej strony, jeżeli $P_c(Z, \phi_{\boldsymbol{X}})$ przyjmie wartość równą zeru, to całość zróżnicowania w zbiorze $\phi_{\boldsymbol{X}}$ wyjaśniana będzie międzygrupowym zróżnicowaniem w warstwach zmiennych $\boldsymbol{X}^{-}$. Z uwagi na sposób definiowania wektora $\boldsymbol{X}^{-}$ będzie to oznaczało, iż Z nie posiada żadnego przełożenia na $\phi_{\boldsymbol{X}}$. Stąd $P_c(Z, \phi_{\boldsymbol{X}})$ interpretować można jako miarę rzeczywistego (a nie tylko brzegowego) udziału Z w wartości R–wskaźnika. Co oczywiste, takie mierniki mają swoje konkretne przełożenie na praktykę badawczą pozwalając między innymi na zidentyfikowanie zmiennych o największych warunkowych oddziaływaniach na niereprezentatywność próby[277]. Dla przykładu, jeżeli pewna zmienna charakteryzowałaby się wysoką wartością wskaźnika brzegowego, a jednocześnie wpływ tej zmiennej zniknąłby warunkowo, to nie miałoby w ogóle sensu koncentrowanie uwagi na takiej zmiennej, gdyż jej oddziaływanie miałoby charakter pozorny, a inne zmienne posiadałyby o wiele większe przełożenie na poziom reprezentatywność zbioru odpowiedzi[278].

---

[275] Zgodnie z rozwarstwieniem na L kategorii według wartości wektora zmiennych $\boldsymbol{X}^{-}$.

[276] Podobnie jak to było w przypadku wskaźnika brzegowego, tak i tutaj $P_c(Z, \phi_{\boldsymbol{X}}) \in [0; S(\phi_{\boldsymbol{X}})]$.

[277] Dla celów praktycznych można wyznaczyć wartości mierników warunkowych odrębnie dla każdej z k–wartości zmiennej Z. Takie warstwowe wskaźniki definiowane są w oparciu o wewnątrzgrupową wariancje $\boldsymbol{X}^{-}$ ograniczoną do jednostek k-tej warstwy. W sposób formalny wskaźnik taki zapisuje się jako $P_c(Z = k, \phi_{\boldsymbol{X}}) = \sqrt{\frac{1}{N-1}\sum_{j=1}^{L}\sum_{i \in U_j}\delta_{k;i}\left(\phi_{\boldsymbol{X}}(x_i) - \bar{\phi}_{\boldsymbol{X},j}\right)^2}$, gdzie U_j oznacza j–tą warstwę populacji wyznaczoną przez $\boldsymbol{X}^{-}$ natomiast $\delta_{k;i} = \begin{cases} 1, \text{jeśli } z_i = k \\ 0, \text{jeśli } z_i \neq k \end{cases}$. Mierniki takie wspomagają monitorowanie jakości realizacji próby oraz umożliwiają podjęcie skutecznych działań na rzecz poprawy reprezentatywności survey'u. Można bowiem zauważyć, iż dążąc do wyższego poziomu jakości realizacji próby powinno się skupić największą uwagę na tych kategoriach jednostek, dla których warstwowe wskaźniki brzegowe charakteryzują się wartościami ujemnymi, a mierniki warunkowe osiągają znaczne wartości dodatnie (por. Schouten i in. 2011: 14–15).

[278] Niezwykle pouczające przykłady praktycznych implementacji R-wskaźnika oraz wskaźników częściowych odnaleźć można w artykułach Schouten i in. (2011: 1–24), Schouten i in. (2012: 382–399), Luiten i in. (2013: 169–189), a także Shlomo i in. (2013: 1–11).

* * *

Metodologiczne właściwości wskaźników reprezentatywności zbioru odpowiedzi oraz możliwość wykorzystania tych mierników w ocenie jakości realizacji sondażowej próby badawczej warto zobrazować kilkoma przykładami empirycznymi. Podobnie jak czyniono wcześniej, tak i tutaj analizy oparte zostaną na danych pozyskanych w ramach realizacji polskiego komponentu piątej rundy Europejskiego Sondażu Społecznego. Do badań z tej edycji ESS-u odwoływano się już zresztą w czwartym rozdziale pracy (rozważając komplikacje wynikające z empirycznej oceny efektywności wielostopniowych schematów doboru prób badawczych), jak też w części pierwszej rozdziału piątego (uzasadniając postrealizacyjny podział próby na kategorie jednostek wyróżniane z uwagi na przyczynę ich niedostępności). Warto zatem przypomnieć, iż jedną z konkluzji sformułowanych w oparciu o przeprowadzone wówczas analizy było stwierdzenie wyraźnej odmienności pomiędzy zbiorem jednostek niedostępnych z powodu braku kontaktu oraz zbiorem jednostek niedostępnych z powodu odmowy udziału w badaniu. A zatem, ponieważ mechanizm stojący za szansą nawiązania kontaktu posiada odmienny charakter od mechanizmu kształtującego gotowość wylosowanych jednostek do współpracy z ankieterem, to skłonność do udziału w badaniu – wykorzystywana do zdefiniowania reprezentatywności zbioru odpowiedzi oraz określania błędu niepełnej realizacji próby w duchu założeń paradygmatu probabilistycznego – pozostaje wypadkową obu wymiarów niedostępności. Innymi słowy, ponieważ są one od siebie niezależne, to wielkości ϕ_i powinny być wyznaczane jako iloczyn quasi-prawdopodobieństw nawiązania kontaktu oraz quasi-prawdopodobieństw kooperacji respondenta z ankieterem.

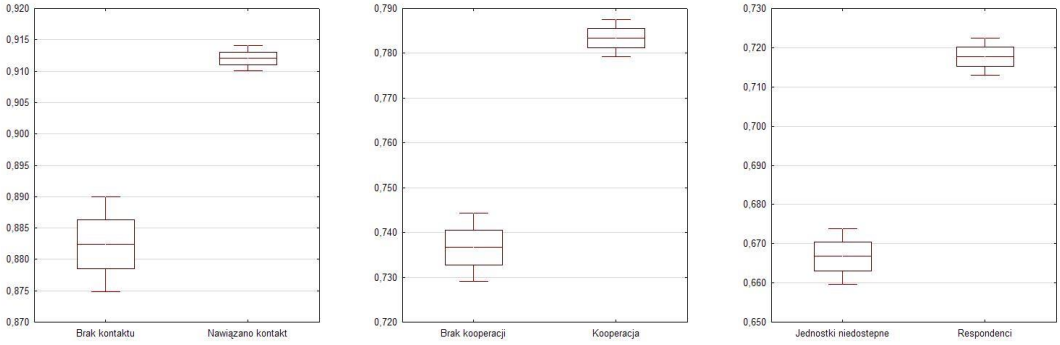
W celu wyznaczenia (estymacji) tych wielkości skonstruowane zostały dwa modele regresji logistycznej. Pierwszy z nich służył do oszacowania szans nawiązania kontaktu (zmienna 0–1) w zbiorze wylosowanych jednostek (z wyłączeniem osób nienależących do populacji), drugi zaś służył do estymacji pozio-

Tabela V.5. Postrealizacyjna struktura jednostek próby w ESS5–PL

Kategoria jednostek próby	Liczebność	Rozkład procentowy
1. Jednostki nienależące do populacji	170	6,4%
2. Brak kontaktu	220	8,3%
3. Odmowa udziału w badaniu	463	17,4%
4. Inne powody niezrealizowania wywiadu	40	1,5%
5. Zbiór respondentów	1768	66,4%
Razem	**2661**	**100,0%**

Źródło: obliczenia własne na podstawie repozytorium danych ESS5–PL

mu gotowości do kooperacji z ankieterem (zmienna 0–1) w zbiorze osób, z którymi udało się nawiązać kontakt, niezależnie od tego, czy następnie wywiad został przeprowadzony, czy też nie. W obu wypadkach zestaw zmiennych pomocniczych ograniczony został do trzech – dostępnych w operacie doboru próby – społeczno-demograficznych charakterystyk jednostek, to znaczy do płci, kategorii wiekowej[279] oraz typu i wielkości miejsca zamieszkania przez wylosowane osoby[280]. Wprawdzie zbiór zmiennych był dość ubogi (nie wyjaśniał w pełni zróżnicowania w zbiorze jednostkowych skłonności do udziału w badaniu)[281], jednak o wybranych charakterystykach wiadomo przynajmniej tyle, że różnicują one w stopniu znaczącym prawdopodobieństwo nawiązania kontaktu oraz szanse wylosowanych osób na kooperację z ankieterem (por. wyniki zaprezentowane w tabelach V.1. oraz V.2.). Co więcej, chociaż w empirii trudno jest zweryfikować, czy i na ile model służący estymacji skłonności do udziału w badaniu oddaje ich rzeczywiste wielkości populacyjne, to uzyskane oszacowania quasi-prawdopodobieństw udziału w badaniu wydają się przynajmniej trafne fasadowo. Doskonałym tego potwierdzeniem są dane zaprezentowane



(a) nawiązanie kontaktu (b) kooperacja z ankieterem (c) udział w badaniu

Ryc. V.9. Zróżnicowanie quasi-prawdopodobieństw nawiązania kontaktu, kooperacji oraz udziału w badaniu wg typów jednostek próby badawczej

Źródło: obliczenia własne na podstawie repozytorium danych ESS5–PL

---

[279] Wyróżnione kategorie obejmowały osoby: (1) do 24 roku życia, (2) od 25 do 34 roku życia, (3) od 35 do 49 roku życia, (4) od 50 do 64 roku życia, (5) 65 lat i więcej.

[280] Zmienna ta wykorzystywana jest w ESS–PL do rozwarstwienia populacji i obejmuje jej podział na: (1) wieś, (2) miasto do 10 tys. mieszkańców, (3) miasto od 10 do 19 tys. mieszkańców, (4) miasto od 20 do 49 tys. mieszkańców, (5) miasto od 50 do 99 tys. mieszkańców, (6) miasto od 100 do 199 tys. mieszkańców, (7) miasto od 200 do 499 tys. mieszkańców, (8) miasto od 500 do 999 tys. mieszkańców, (9) miasto pow. 1 mln osób (tj. Warszawa).

[281] Wybrane charakterystyki pomocnicze umożliwiały podział próby na 90 różnych warstw powstałych po skrzyżowaniu płci wylosowanych osób, ich kategorii wiekowej oraz typu i wielkości zamieszkiwanych przez nie miejscowości.

na trzech kolejnych wykresach ryciny V.9. Zamieszczono na nich informacje o średnich (przeciętnych) szansach: (a) nawiązania kontaktu (w zbiorze osób, z którymi się to udało, oraz takich, do których nie udało się dotrzeć), (b) kooperacji z ankieterem (w zbiorze osób niechętnych oraz gotowych do współpracy), oraz (c) udziału w badaniu (odpowiednio w zbiorze respondentów oraz jednostek niedostępnych), wraz z odpowiadającymi im błędami standardowymi (ramki) oraz 95-procentowymi przedziałami ufności (wąsy).

Przyglądając się powyższym wykresom, należy zauważyć, że uzyskane wyniki są rzeczywiście zgodne z oczekiwaniami. Istotnie, model empiryczny ukazuje, iż osoby niedostępne (w tym odmawiające oraz takie, z którymi kontakt okazał się niemożliwy), charakteryzują się przeciętnie mniejszym poziomem skłonności do udziału w badaniu (odpowiednio mniejszą gotowością do kooperacji oraz dostępnością) od osób, do których udało się dotrzeć i zrealizować z nimi wywiady (lub odpowiednio skłonić do kooperacji oraz nawiązać kontakt). Opierając się zatem na oszacowaniach jednostkowych skłonności do udziału w badaniu, można teraz poszukać odpowiedzi na trzy ważne pytania. Po pierwsze, w jakim zakresie wyniki pomiaru ESS5-PL nie spełniają warunku pełnej reprezentatywności (tj. całkowicie losowego mechanizmu niedostępności jednostek)? Po drugie, które zmienne poddawane pomiarowi obciążone są największymi błędami niepełnej realizacji próby? Po trzecie, które ze zmiennych pomocniczych (i w jaki sposób) oddziałują w największym stopniu na odchylenie zbioru pomiaru od warunku pełnej reprezentatywności?

Częściowej odpowiedzi na pierwsze z tych pytań dostarcza tabela V.5. Przedstawiono w niej wartości odchyleń standardowych w zbiorze jednostko-

Tabela V.6. Wskaźniki reprezentatywności zbioru danych ESS5–PL

	$\hat{\mathbf{S}}(\hat{\boldsymbol{\phi}}_{\boldsymbol{X}})$ [i)]	$\hat{\mathbf{R}}(\hat{\boldsymbol{\phi}}_{\boldsymbol{X}})$ [ii)]	$\mathbf{CI}_{\mathbf{0,05}}^{\boldsymbol{BT}}$ [iii)]	$\hat{\mathbf{B}}_{\boldsymbol{max}}$ [iv)]	$\hat{\mathbf{E}}_{\boldsymbol{max}}$ [v)]
Nawiązanie kontaktu	nawiązano kontakt *vs*. brak kontaktu				
	0,049	0,903	0,901–0,905	2,7%	2,9%
Kooperacja z ankieterem	kooperacja z ankieterem *vs*. niechęć do udziału w badaniu				
	0,090	0,821	0,819–0,823	5,8%	5,9%
Skłonność do udziału w badaniu	respondenci *vs*. jednostki niedostępne				
	0,102	0,796	0,794–0,798	7,2%	7,3%

Źródło: obliczenia własne na podstawie repozytorium danych ESS5–PL

[i)] Odchylenie standardowe w zbiorze oszacowań jednostkowych skłonności do udziału w badaniu wyznaczone zostało zgodnie ze wzorem (V.10.).

[ii)] Wartość wskaźnika reprezentatywności próby (por. wzór V.8.) wyznaczona w oparciu o $\hat{S}(\hat{\phi}_X)$.

[iii)] Przedział ufności estymatora *R*-wskaźnika wyznaczony został w oparciu o metodę bootstrapową z 200-krotną replikacja próby badawczej.

[iv)] Oszacowanie maksymalnej wielkości błędu systematycznego dla zmodyfikowanych estymatorów proporcji wyznaczono przyjmując we wzorze (V.9.) wartość $S(Y) = 0{,}5$.

[v)] Oszacowanie całkowitego błędu pomiaru dla zmodyfikowanych estymatorów proporcji wyznaczono zakładając maksymalną wielkość błędu losowego i systematycznego (por. wzór 13 w: Cobben i in. 2008: 8).

wych szans nawiązania kontaktu, gotowości do kooperacji oraz skłonności do udziału w badaniu, wraz z odpowiadającymi im *R*-wskaźnikami oraz 95-procentowymi przedziałami ufności dla oszacowań $\widehat{\mathrm{R}}(\hat{\phi}_X)$. W tabeli zawarto również informację o maksymalnych wielkościach błędów (systematycznych oraz całkowitych), które w najmniej korzystnej sytuacji wpływałyby na proces estymacji parametrów wskaźników struktury.

W pierwszej kolejności należy wskazać, że uzyskane w ESS5-PL wartości *R*-wskaźników osiągnęły poziom podobny do odnotowywanego w innych projektach (por. Schouten i in. 2012: 387)[282]. Wyniki te ukazują również, że odmowy udziału w badaniu mają znacznie większy wpływ na odchylenie zbioru pomiaru od warunku pełnej reprezentatywności, niż ma to miejsce w przypadku braku kontaktu. Oznacza to, że w celu poprawy reprezentatywności zbioru odpowiedzi najbardziej efektywne okazuje się dążenie do ograniczenia liczby odmów udziału badaniu.

Niezwykle pouczające okazują się też informacje zawarte w dwóch ostatnich kolumnach tabeli V.6. Wielkość $\widehat{\mathrm{B}}_{max}$ podaje bowiem oszacowanie maksymalnego poziomu błędu systematycznego estymatorów wskaźników struktury, natomiast $\widehat{\mathrm{E}}_{max}$ – oszacowanie maksymalnego błędu całkowitego estymatorów takich parametrów (jest to *de facto* pierwiastek z wielkości błędu średniokwadratowego). Porównując obie statystyki, można stwierdzić, że systematyczne zniekształcenie wyników jest rzeczywiście o wiele poważniejszą konsekwencją niepełnej realizacji próby aniżeli wynikający z tego przyrost wariancji. Warto przy tym pamiętać, że w estymacji wielkości $\widehat{\mathrm{B}}_{max}$ oraz $\widehat{\mathrm{E}}_{max}$ zakłada się przypadek skrajnie niekorzystny, to znaczy pełne skorelowanie skłonności do udziału w badaniu z wartościami zmiennych poddawanych pomiarowi, a także maksymalną wariancję estymatorów wskaźników struktury. A zatem w rzeczywistości – na co zwracano uwagę już wcześniej – błędy braków danych będą mniejsze, niż wskazują na to wartości maksymalne. Wielkości takich błędów, dla konkretnych zmiennych, można oszacować, wykorzystując estymatory jednostkowych skłonności do udziału w badaniu (por. wzór V.1. w sekcji V.3. tego rozdziału). W tabeli V.7. przedstawiono zmienne – ze zbioru ESS5-PL[283] – dla których wielkości błędów okazały się największe.

---

[282] Chociaż takie międzysurveyowe porównania *R*-wskaźników utrudnione są zazwyczaj z uwagi na odmienny sposób estymacji skłonności do udziału w badanu, to akurat w przywołanych studiach Schoutena i in. (2012) oraz w przedstawionych tutaj analizach danych ESS5-PL wykorzystano jednakowy zestaw zmiennych pomocniczych. Istotną komplikację stanowi jednak fakt, że wspomniani badacze zdefiniowali w nieco odmienny sposób kategorie wiekowe oraz kategorie typu i wielkości miejsca zamieszkania wykorzystywane w szacowaniu $\phi_{\boldsymbol{X}}(x_i)$. Tym samym, choć zestaw zmiennych pomocniczych był jednakowy, wartości *R*-wskaźników nie są w pełni porównywalne.

[283] Wielkości błędów systematycznych wyznaczone zostały dla tych samych zmiennych, które uwzględniono już we wcześniejszych analizach przeprowadzonych w rozdziale IV (wykaz zmiennych wraz z ich opisem zawarty został w tabeli aneksowej A2.).

Tabela V.7. Wielkości systematycznych błędów niepełnej realizacji próby dla wybranych zmiennych ESS5-PL

Numer pytania	Charakterystyka zmiennej	$\mathbf{Corr}(\hat{\boldsymbol{\phi}}_{\boldsymbol{X}}, \hat{\boldsymbol{y}})$[i]	$\widehat{\mathbf{B}}(\hat{\boldsymbol{y}})$[ii]
F41	Dochód łączny gospodarstwa domowego ze wszystkich źródeł	–0,164	–0,061
C21	Poziom religijności	0,102	0,043
B28	Opinia o stanie poziomu kształcenia, edukacji w Polsce	0,101	0,030
A3	Słuchanie radia, czas całkowity w przeciętny dzień tygodnia	–0,067	–0,025
F1	Liczba osób zamieszkujących na stałe w gospodarstwie domowym	0,112	0,024

Źródło: Obliczenia własne na podstawie repozytorium danych ESS5–PL

[i] $\mathrm{Corr}(\hat{\phi}_X, \hat{y})$ jest współczynnikiem korelacji pomiędzy estymatorami jednostkowych skłonności do udziału w badaniu oraz wartościami pomiaru zmiennych w zbiorze ESS5.

[ii] $\widehat{\mathrm{B}}(\hat{y})$ jest oszacowaniem poziomu błedu systematycznego wyznaczonym zgodnie ze wzorem (V.1.).

W tabeli tej, poza oszacowaniem błędu systematycznego (dla zmodyfikowanych estymatorów Horvitza-Thomsona parametrów średnich arytmetycznych), zawarto również informacje o poziomie skorelowania skłonności do udziału w badaniu ($\hat{\phi}_X$) z wartościami zmiennych ($\hat{y}$). Zauważyć można, że chociaż poziom korelacji jest statystycznie istotny (to znaczy znacząco różny od zera), to jednak siła związku okazuje się niezbyt duża i to nawet dla zmiennych o największych poziomach obciążenia błędem braku danych. Wystarczy przy tym dodać, że przeciętny poziom korelacji (tj. średnia bezwzględna z wartości współczynników $\mathrm{Corr}(\hat{\phi}_X, \hat{y})$ dla wszystkich rozpatrywanych zmiennych) wyniósł jedynie 0,049. Przyglądając się wielkościom błędów systematycznych, można również zobaczyć, że w realizacji próby ESS5-PL najbardziej niedoszacowane są jednostki o wyższym poziomie dochodu, osoby mniej religijne, zamieszkujące mniej liczne gospodarstwa domowe, wyrażające bardziej krytyczny stosunek wobec systemu edukacji w Polsce, jak też częściej słuchające radia. Z drugiej strony, niepełna realizacja próby ESS5-PL nie wpłynęła w ogóle na wyniki pomiaru (wielkość błędu systematycznego wyniosła zero) w odniesieniu do pytań G4 (*Kobieta powinna być gotowa ograniczyć pracę zawodową dla dobra rodziny*), D35 (*Należy ściśle przestrzegać przepisów prawa*), A8 (*Większości ludzi można ufać, czy też ostrożności nigdy nie za wiele*) oraz D25 (*Jak często policjanci w Polsce przyjmują łapówki*). Generalnie można wskazać, że średni (bezwzględny) poziom błędu systematycznego był nieznaczny, osiągając wartość równą 0,01.

Na zakończenie prowadzonych analiz warto jeszcze raz przyjrzeć się wartościom wskaźników reprezentatywności w zbiorze ESS5-PL i spróbować określić charakter mechanizmu niedostępności oraz zmienne oddziałujące w największym stopniu na odchylenie zbioru pomiaru od warunku całkowitej reprezentatywności. Wskazywano już wcześniej, że im większą wartość – dla pewnej

Tabela V.8. Częściowe R-wskaźniki dla zmiennych pomocniczych wykorzystanych w estymacji skłonności do udziału w badaniu jednostek wylosowanych do próby ESS5–PL

Zmienna pomocnicza	Wskaźniki brzegowe			Wskaźniki warunkowe		
	Nawiązanie kontaktu	Kooperacja	Udział w badaniu	Nawiązanie kontaktu	Kooperacja	Udział w badaniu
Płeć:	**0,016**	**0,016**	**0,003**	**0,028**	**0,057**	**0,063**
mężczyzna	–0,012	0,012	0,002	0,022	0,037	0,043
kobieta	0,011	–0,011	–0,002	0,018	0,043	0,046
Wiek:	**0,026**	**0,039**	**0,040**	**0,034**	**0,067**	**0,074**
do 24 lat	–0,004	0,032	0,033	0,009	0,035	0,037
od 25 do 34 lat	–0,013	0,002	–0,008	0,017	0,023	0,029
od 35 do 49 lat	–0,012	–0,013	–0,021	0,018	0,033	0,040
od 50 do 64 lat	0,005	–0,001	0,003	0,012	0,029	0,033
65 lat i więcej	0,018	–0,018	–0,004	0,018	0,029	0,026
Typ i wielkość miejscowości:	**0,029**	**0,057**	**0,069**	**0,038**	**0,079**	**0,094**
wieś	0,012	0,036	0,042	0,018	0,048	0,056
miasto do 10 tys.	0,005	0,016	0,019	0,007	0,020	0,024
miasto 10–19 tys.	0,007	0,000	0,005	0,009	0,016	0,019
miasto 20–49 tys.	0,000	–0,004	–0,004	0,008	0,021	0,023
miasto 50–99 tys.	–0,001	–0,019	–0,018	0,008	0,028	0,030
miasto 100–199 tys.	–0,003	–0,017	–0,018	0,007	0,024	0,025
miasto 200–499 tys.	–0,011	–0,013	–0,020	0,015	0,021	0,028
miasto 500–999 tys.	–0,023	–0,023	–0,037	0,025	0,028	0,041
miasto pow. 1 mln.	0,002	–0,018	–0,015	0,002	0,018	0,016

Źródło: Obliczenia własne na podstawie repozytorium danych ESS5–PL

zmiennej pomocniczej – osiągnie wskaźnik brzegowy, tym bardziej znaczny będzie wpływ takiej zmiennej na odchylenie zbioru pomiaru od warunku pełnej reprezentatywności. Podobnie, im większą wartość osiągnie częściowy wskaźnik warunkowy, tym większa część wariancji w zbiorze $\hat{\phi}_{\boldsymbol{X}}$ ulokowana będzie wewnątrz warstw wyznaczonych przez wartości zmiennych pomocniczych. Z uwagi na sposób definiowania wskaźnika warunkowego pozostaje on miarą rzeczywistego wpływu pewnej zmiennej na reprezentatywność zbioru odpowiedzi, tj. wpływu uwzględniającego oddziaływanie wszystkich zmiennych pomocniczych. Innymi słowy, dzięki zestawieniu *R*-wskaźnika z wskaźnikami częściowymi będzie można (1) ustalić część wariancji w zbiorze skłonności do udziału w badaniu ulokowaną pomiędzy warstwami wyznaczonymi przez wartości zmiennych pomocniczych (to znaczy sprawdzić, czy mechanizm niedostępności jest *MAR* względem rozpatrywanych zmiennych), (2) określić poziom

współliniowości charakterystyk pomocniczych (tj. rzeczywisty wpływ tych zmiennych na zróżnicowanie skłonności do udziału w badaniu), (3) jak też wyróżnić kategorie respondentów mające najbardziej znaczący wpływ na odchylenie zbioru pomiaru od warunku pełnej reprezentatywności (tj. oznaczyć warstwy, na których powinny koncentrować się dodatkowe działania terenowe skutkujące ograniczeniem zróżnicowania w zbiorze skłonności do udziału w badaniu).

Porównując wartości wskaźników brzegowych z przedstawionymi już wcześniej w tabeli V.6. wielkościami $\hat{S}(\hat{\phi}_{\boldsymbol{X}})$, należy stwierdzić, że mechanizm niedostępności w próbie ESS5-PL nie jest raczej *MAR* względem wyróżnionych zmiennych pomocniczych. Istotnie, zarówno dla szans nawiązania kontaktu, jak i dla gotowości do kooperacji, czy też skłonności do udziału w badaniu, znaczna część wariancji w zbiorze $\hat{\phi}_{\boldsymbol{X}}$ mieści się wewnątrz warstw wyznaczonych przez zmienne pomocnicze. Oznacza to, że międzygrupowe zróżnicowania $\hat{\phi}_{\boldsymbol{X}}$ ustalone względem wartości tych charakterystyk wyjaśniają niewielką część całkowitej wariancji w zbiorze jednostkowych skłonności do udziału w badaniu. Należy przy tym wskazać, że płeć oddziałuje w stopniu najmniejszym na reprezentatywność pomiaru ESS5-PL, z kolei największy wpływ ma typ i wielkość miejscowości zamieszkania. Potwierdza to słuszność wykorzystania tej ostatniej zmiennej do rozwarstwienia próby w polskim komponencie badań ESS-u.

Przyglądając się wartościom wskaźników częściowych, można również zobaczyć, że mierniki warunkowe są większe od brzegowych. Oznacza to, że zmienne pomocnicze oraz warstwy wyznaczone przez te zmienne nie są współliniowe, to znaczy każda z nich ma odmienny wpływ na odchylenie pomiaru od warunku pełnej reprezentatywności, wyjaśniając inną część zróżnicowania w zbiorze jednostkowych skłonności do udziału w badaniu. Łatwo również dostrzec, że w wymiarze związanym z brakiem kontaktu 8 spośród 16 warstw uformowanych przez wartości zmiennych pomocniczych charakteryzowało się ujemną wartością wskaźników brzegowych (było niedoreprezentowanych), natomiast w wymiarze gotowości do kooperacji oraz skłonności do udziału w badaniu niedoszacowanych było już 10 warstw. Biorąc dodatkowo pod uwagę wartości mierników warunkowych, można już wskazać, że w celu poprawy reprezentatywności zbioru danych, w wymiarze związanym z brakiem kontaktu, należałoby zwrócić uwagę przede wszystkim na mieszkańców miast od 500 do 999 tys. osób, mężczyzn oraz osoby w wieku od 25. do 49. roku życia, z kolei w wymiarze niedostępności związanej z odmowami udziału w badaniu na kobiety, osoby zamieszkujące w miastach od 500 do 999 tys. mieszkańców oraz powyżej 65. roku życia, a w będącym ich wypadkową wymiarze skłonności do udziału w badaniu, na mieszkańców największych polskich miast (z wyłączeniem Warszawy) oraz osoby w wieku od 25. do 49. roku życia.

V.5. Uwagi końcowe

Studia nad konsekwencjami niepełnej realizacji sondażowej próby badawczej prowadzone w duchu założeń probabilistycznego paradygmatu błędu braku danych ukazały dość wyraźnie, że jednym z największych wyzwań praktycznych związanych z oceną jakości terenowej realizacji próby staje się identyfikacja mechanizmu kształtującego charakter niedostępności jednostek próby. W zasadzie od tego, czy i w jakim stopniu uda się odtworzyć „rzeczywiste" skłonności wylosowanych osób do udziału w badaniu, zależy w znacznej mierze skuteczność wszystkich działań badawczych oraz postbadawczych podejmowanych w celu wyeliminowania zniekształcenia wyników pomiaru na skutek niepełnej realizacji próby. Innymi słowy, przyjęcie paradygmatu probabilistycznego narzuca badaczom obowiązek zarówno rozpoznania czynników powiązanych z procesem niedostępności jednostek, jak też zgromadzenia – w toku prowadzonego badania – odpowiednich informacji potrzebnych do modelowania jednostkowych skłonności do udziału w badaniu.

Zasadniczą różnicą wynikającą z przyjęcia modelu probabilistycznego, zamiast deterministycznego nie jest jednak ani odmienny sposób definiowania wielkości systematycznego błędu braku danych, ani nawet krytyczny stosunek do wskaźnika realizacji próby – jako miernika jakości danych, lecz zupełnie inne podejście w ocenie efektywności pewnych procedur terenowych służących zwiększaniu szans realizacji wywiadów z osobami wylosowanymi do próby. W rozdziale wykazane zostało w sposób jednoznaczny, że wszystkie takie działania mają uzasadnienie nie tyle w tym, że pozwalają zwiększyć wielkość wskaźnika realizacji próby, ile że poprzez zwiększenie szans nawiązania kontaktu z osobami trudno dostępnymi, czy też zwiększenie szans realizacji wywiadów z osobami mniej skłonnymi do kooperacji z ankieterem, pozwalają zmniejszyć wariancję w zbiorze jednostkowych skłonności do udziału w badaniu, prowadząc tym samym do ograniczenia, lub nawet do wyeliminowania, błędu systematycznego.

Zakończenie

Nieco ponad 60 lat temu Herbert H. Hyman przestrzegał socjologów-badaczy przed ignorowaniem błędów, na jakie narażone jest każde dochodzenie naukowe. We wprowadzeniu do monografii *Interviewing in Social Research* autor ten stwierdził dość wymownie, iż „nieświadomość błędu nie oznacza jego nieistnienia" (Hyman 1954: 4). Czterdzieści lat później Paweł B. Sztabiński zwrócił uwagę na inne zjawisko dotykające naukowej praktyki badawczej. Podsumowując rezultaty swoich autorskich analiz nad wpływem oraz efektem ankieterskim, badacz ten zauważył, że „[wraz ze wzrostem – P.J.] liczby badań socjologicznych […] nastąpiło […] rozdzielenie funkcji badawczych od funkcji związanych z opracowaniem badań […]. Rezultatem tego stał się trwający do dziś brak zainteresowania części socjologów-badaczy fazą terenową badań […]. Wielu badaczy poczuwa się do odpowiedzialności wyłącznie za kształt kwestionariusza […] oraz za analizy statystyczne. Pozostałe czynności, związane z przygotowaniem i realizacją badań […] powierzają wyspecjalizowanej agencji badawczej. Oczywiste jest, że w takiej sytuacji ich znajomość 'terenu' […] jest często znikoma" (Sztabiński, P.B. 1997: 217–218). Chociaż autor książki *Ankieterzy i ich respondenci* odnosił ową znikomą znajomość „terenu" do niewielkiego zainteresowania socjologów-badaczy pracą ankieterów oraz reakcjami respondentów na problematykę wywiadu – a więc do procesu pomiaru – to jednak odseparowanie funkcji badawczych od realizacyjnych jest również (a może przede wszystkim) przypadłością procesu losowania oraz terenowej realizacji próby sondażowej. Wprawdzie takie postępowanie nie jest niczym nagannym (a nawet nadzwyczajnym), to może jednak oznaczać, że socjolog-badacz niezainteresowany terenową fazą badań traktować będzie (zbyt optymistycznie) otrzymane wyniki jako pewne (nawet jeśli takie nie będą), co w pewnym sensie można nawet zrozumieć, bowiem opis społecznej rzeczywistości nie miałby przecież większego sensu, jeśli reprezentatywność próby miałaby być niska, a proces pomiaru nietrafny. Za tym brakiem zainteresowania kryje się jednak znacznie częściej brak wiedzy o tym, w jaki sposób pewne decyzje badawcze oraz wynikające z nich działania terenowe i postterenowe przekładają się na reprezentatywność próby oraz jakość pomiaru. A zatem, taka

outsourcingowa postawa badaczy czyni dochodzenie naukowe jeszcze bardziej niepewnym niż wynika to z jego natury.

Zaprezentowane w tej pracy analizy metodologiczne oraz empiryczne wydają się wnosić znaczny wkład w rozwój metodologii badań sondażowych przynajmniej w kilku jej kluczowych wymiarach.

Po pierwsze, udało się wykazać, że większość działań podejmowanych przez badaczy w celu poprawy jakości operatów losowania prób surveyowych (w tym omówione w tej pracy procedury sieciowania jednostek, przedziałów półotwartych oraz operatów wielokrotnych) skoncentrowanych jest przede wszystkim na wyeliminowaniu błędu niepełnego i/lub nadmiarowego pokrycia, w mniejszym stopniu natomiast na ograniczeniu błędów będących efektem zwielokrotnienia szans doboru pewnych jednostek i/lub ich uzespołowienia. O ile bowiem błędy niepełnego oraz nadmiarowego pokrycia mogą prowadzić do systematycznego wypaczenia wyników pomiaru, o tyle zróżnicowanie prawdopodobieństw selekcji (będące efektem dwóch pozostałych klas błędów) skutkuje „jedynie" koniecznością przeprowadzenia ważenia danych oraz obniżeniem precyzji wnioskowania statystycznego (wymaga tym samym zwiększenia liczebności próby). Innymi słowy, ponieważ błędów systematycznych nie da się wyeliminować standardowymi procedurami trenowymi, to badacze preferują minimalizowanie szans ich wystąpienia, nawet jeśli kosztem tego miałoby być wykorzystanie operatu wiążącego się z obniżeniem efektywności próby (to znaczy zmniejszeniem precyzji pomiaru). Zaprezentowane w pracy analizy ukazały również, że stopień pokrycia jednostek zespołowych (gospodarstw domowych lub budynków mieszkalnych) rejestrami adresowymi nie ma większego przełożenia na poziom pokrycia jednostek indywidualnych (osób tworzących populację docelową). Zwrócono także uwagę, iż korzyści wynikające z eliminacji systematycznego błędu niepełnego pokrycia poprzez wykorzystanie pewnych terenowych procedur badawczych niwelowane są zazwyczaj przez pojawienie się innych źródeł błędów.

Po drugie, w pracy podkreślono konieczność weryfikacji kryteriów uprawomocniających stosowanie pewnych popularnych w metodologii badań sondażowych mierników oceny efektywności schematu doboru próby. Wprawdzie definicja powszechnie stosowanego miernika efektywności losowania (*design effect*) jest wyjątkowo prosta i intuicyjna, jednak jej zastosowanie praktyczne okazuje się już niezwykle kłopotliwe. W praktyce schematy losowania przyjmują bowiem niezwykle złożoną postać i rzadko kiedy ich efektywność daje się oszacować bez przyjęcia jakichś upraszczających założeń. Badacze wykorzystują wprawdzie zredukowaną postać estymatora *design effect* – polegającą na niezależnym wyznaczeniu efektu losowania stratyfikacyjnego, zespołowego oraz doboru z nierównymi szansami selekcji jednostek, a następnie na prze-

mnożeniu takich cząstkowych wskaźników, jako miernikowi efektu całkowitego – problem polega jednak na tym, że możliwość zastosowania takiej uproszczonej wersji wskaźnika efektywności schematu losowania pozostaje obwarowana na tyle restrykcyjnymi wymogami formalnymi, że tylko w nielicznych przypadkach schematy doboru próby warunki takie spełniają. W konsekwencji, uproszczone mierniki mogą prowadzić do niewłaściwego oszacowania rzeczywistego przyrostu wariancji w danym schemacie losowania próby, i co za tym idzie, w sposób nieprawidłowy określać efektywność próby badawczej. W pracy wykazane zostało, że powszechnie stosowany miernik oceny efektywności schematów doboru próby prowadzić może do znacznego niedoszacowania skali przyrostu wariancji. W polskiej metodologii badań sondażowych zagadnienia te podjęte zostały z całą pewnością po raz pierwszy. Zresztą również w literaturze o zasięgu światowym niewiele było dotąd prób empirycznej oceny konsekwencji wynikających z przyjęcia uproszczonych metod szacowania efektywności schematów doboru prób badawczych.

Po trzecie, w pracy tej podaję w wątpliwość ekwiwalentny charakter terenowej realizacji prób imiennych, adresowych oraz gospodarstw domowych. Szczegółowe zestawienie postbadawczych wzorców terenowej realizacji próby ze wszystkich krajów uczestniczących w badaniach ESS-u uwidoczniło niezwykle niepokojące tendencje. Ukazało ono, że w próbach adresowych znacznie większy odsetek respondentów stanowią osoby łatwej dostępne oraz bardziej skłonne do udziału w badaniu, jednocześnie odnotowuje się w nich mniejszy (niż w próbach imiennych) odsetek odmów. Tym samym pokazano, że praktyka stoi w sprzeczności z tym, czego należałoby oczekiwać, gdyby tylko proces selekcji respondentów pozostał (w próbach adresowych) całkowicie wolny od systematycznych zniekształceń. Zarówno w polskiej, jak i światowej literaturze metodologicznej problemy te podjęte zostały po raz pierwszy.

Po czwarte, zaletą pracy są analizy poziomu reprezentatywności terenowej realizacji próby w duchu probabilistycznego paradygmatu błędu jej niepełnej realizacji. Przynajmniej w polskiej metodologii badań sondażowych paradygmat ten nie był dotąd przedmiotem jakiegoś szczególnego zainteresowania. Studia literaturowe wykazały, że zdecydowana większość procedur terenowych mających na celu zarówno szacowanie, jak i też ograniczenie wielkości błędów braku danych opiera się na założeniach przyjmowanych w – niezwykle upraszczającym rzeczywistość – paradygmacie deterministycznym. Konsekwencją tego pozostaje nadmierna koncentracja uwagi wielu badaczy na wskaźniku realizacji próby jako mierniku jakości badania, jak też przecenianie skuteczności działań służących minimalizacji ryzyka błędu systematycznego poprzez maksymalizację poziomu realizacji próby. Analizy prowadzone w duchu probabilistycznego paradygmatu błędu wykazały przy tym, że wszystkie

procedury terenowe ukierunkowane na zwiększanie wskaźników realizacji próby będą miały sens tylko wtedy, kiedy ich konsekwencją będzie wyeliminowanie/zmniejszenie zróżnicowania w zbiorze jednostkowych skłonności wylosowanych osób do udziału w badaniu lub (przynajmniej) ograniczenie poziomu kowariancji pomiędzy wartościami analizowanej zmiennej oraz skłonnościami wylosowanych jednostek do udziału w badaniu.

Po piąte, w pracy udało się wykazać, że odpowiednie zdefiniowanie wag (poprzez wykorzystanie empirycznych oszacowań jednostkowych skłonności do udziału w badaniu) pozwala nie tylko na zredukowanie wielkości systematycznego błędu braku danych, ale również na poprawę precyzji wnioskowania statystycznego. Pokazano, że najbardziej efektywne jest wykorzystanie wag powiązanych zarówno z badaną zmienną (im bardziej będą one skorelowane, tym lepiej dla poprawy precyzji pomiaru), jak też z jednostkowymi skłonnościami do udziału w badaniu (im bardziej pozostaną one skorelowane, tym lepiej dla eliminacji błędu nielosowego). A zatem, jeżeli ważenie danych stosuje się w celu redukcji błędu braku danych, to ma ono sens wyłącznie w odniesieniu do zmiennych pozostających w silnej korelacji z wartościami wag. Nawet jeśli nie da się ograniczyć wielkości błędu systematycznego (nie zawsze wiadomo, czy udało się poprawnie wyestymować jednostkowe skłonności do udziału w badaniu), to przynajmniej poprawi się w ten sposób precyzję pomiaru. W innym przypadku ważenie danych nie spowoduje redukcji błędu systematycznego, doprowadzi natomiast do przyrostu wariancji.

W książce udało się również wskazać na kilka obszarów wymagających dalszych pogłębionych analiz. Wydaje się, że do najważniejszych wyzwań, z jakimi przyjdzie się zmierzyć w najbliższej przyszłości, należeć będzie przede wszystkim wypracowanie uniwersalnych standardów identyfikacji mechanizmów kształtujących niedostępność wylosowanych jednostek oraz ujednolicenie procedur służących ocenie reprezentatywności próby w ramach wyznaczonych przez probabilistyczny paradygmat błędu jej niepełnej realizacji. Do tej pory udało się jedynie wypracować i upowszechnić standardy wyznaczania wartości wskaźników realizacji próby. Wymaga to również rewizji „starych" oraz zaproponowaniu „nowych" strategii badawczych i postbadawczych zmierzających do ograniczenia negatywnych konsekwencji niedostępności pewnych jednostek wylosowanych do próby. Dużą rolę w tym względzie odgrywać będzie zapewne schemat badawczy dopuszczający zastosowanie technik mieszanych (*mixed mode design*) – polegający na zbieraniu tych samych informacji od różnych osób za pomocą różnych technik – w którym upatruje się nadziei na zwiększenie szans realizacji wywiadów z większą liczbą wylosowanych jednostek, niż ma to miejsce w przypadku zastosowania tylko jednej techniki. Problem braku uczestnictwa osób wylosowanych do badania stał się już powszechny, nic nie

wskazuje na to, aby tendencja ta miała się znacząco odwrócić. Możliwość upowszechnienia technik mieszanych, wpisujących się w ogólną ideę dopasowywania schematu terenowej realizacji próby do konkretnej osoby (*tailoring design*), wymaga jednak refleksji nad tym, w jaki sposób połączenie wielu technik przekłada się na jakość całego badania (może pojawić się tak zwany *mode effect*, skutkujący kumulacją błędów specyficznych dla danej techniki badawczej).

Kolejnym obszarem wymagającym dalszych analiz pozostają zagadnienia porównywalności terenowej realizacji prób imiennych oraz adresowych. Oczywiście za decyzją o wyborze konkretnego operatu losowania jednostek i powiązanego z nim typu próby badawczej stoją zazwyczaj względy merytoryczne (na przykład zniwelowanie ryzyka systematycznego błędu niepełnego pokrycia populacji), jednak z przeprowadzonych w tej pracy studiów wynika jasno, że kluczowe znaczenie dla zapewnienia odpowiedniej jakości realizacji próby – zwłaszcza w losowaniu wielostopniowym – ma nie tylko zobiektywizowanie procedur doboru jednostek, ale też (a może przede wszystkim) terenowa kontrola przebiegu takiej selekcji.

* * *

Ponieważ praca ta ma przede wszystkim wymiar metodologiczny i użytkowy – a ma również ambicje edukacyjne – to w ramach podsumowania przedstawione zostaną rekomendacje ułatwiające socjologom-badaczom metodologiczną ocenę poziomu reprezentatywności sondażowej próby badawczej.

Rekomendacje w zakresie oceny jakości operatów doboru sondażowych prób badawczych:

1. Uchybienia operatu wykorzystanego w doborze sondażowej próby badawczej – odstępstwa od operatu idealnego – powinny zostać uznane za poważne źródło błędu losowego oraz systematycznego.
   W metodologii badań sondażowych wyróżnia się cztery główne klasy ułomności operatów doboru prób badawczych: (1) niepełne pokrycie, (2) nadmiarowe pokrycie, (3) multiplikowanie jednostek oraz (4) uzespołowienie jednostek. W każdym z tych przypadków operat losowania nie przystaje w pełni do populacji docelowej (odstaje od operatu idealnego), przy czym niepełne oraz nadmiarowe pokrycie może skutkować błędami systematycznymi, a multiplikowanie jednostek oraz ich uzespołowienie – prowadzić do zróżnicowania szans selekcji oraz przyrostu wariancji estymatorów.
2. Przystępując do wyboru konkretnego operatu (lub operatów) doboru próby badawczej, należy opisać relacje zachodzące pomiędzy populacją

będącą przedmiotem badania oraz populacją docelową (to znaczy taką, która w rzeczywistości zostanie objęta badaniem reprezentatywnym).

3. Należy w sposób precyzyjny scharakteryzować wykorzystany operat (lub operaty) doboru próby badawczej. Minimalny zestaw informacji powinien obejmować: (1) dane dotyczące zakresu czasowego oraz przestrzennego operatu, (2) zasady aktualizowania danych zawartych w operacie doboru próby oraz (3) stopnień pokrycia populacji docelowej.
4. Konieczne jest zidentyfikowanie słabości operatu poprzez rozpoznanie jego potencjalnych uchybień przekładających się na błędy systematyczne:
   a. należy ustalić stopień niepełnego i/lub nadmiarowego pokrycia jednostek populacji docelowej przez operat doboru próby. Istotne jest to, że błędy niepełnego oraz nadmiarowego pokrycia okazują się domeną nie tylko badań realizowanych na próbach reprezentatywnych, ale również badań pełnych na całych populacjach;
   b. w ramach oceny jakości operatu należy określić warstwy populacji docelowej charakteryzujące się istotnie większym (odbiegającym od innych) stopniem niepełnego i/lub nadmiarowego pokrycia. Wskaźniki niepełnego i/lub nadmiarowego pokrycia powinny zostać wyznaczone zwłaszcza w ramach warstw posiadających kluczowe znaczenie dla opisu przedmiotu badania;
   c. należy scharakteryzować działania wykorzystane w celu ograniczenia błędu systematycznego będącego efektem uchybień operatu, to znaczy procedury służące zwiększeniu pokrycia i/lub eliminacji pokrycia nadmiarowego. Ważne jest to, że zdecydowana większość procedur wykorzystywanych w celu poprawy jakości operatu służy przede wszystkim ograniczeniu błędu niepełnego lub nadmiarowego pokrycia, natomiast w mniejszym stopniu ich zadaniem jest wyeliminowanie uzespołowienia lub multiplikowania szans selekcji pewnych jednostek do próby badawczej. A zatem tym, co powinno decydować o wyborze konkretnego operatu (lub operatów), jest stopień pokrycia populacji docelowej.
5. Konieczne jest zidentyfikowanie słabości operatu poprzez rozpoznanie jego potencjalnych uchybień przekładających się na precyzję pomiaru:
   a. należy ustalić stopień multiplikowania szans doboru pewnych jednostek do próby badawczej oraz opisać procedury ważenia danych służące wyrównywaniu szans selekcji;
   b. należy rozpoznać stopień uzespołowienia jednostek. W sytuacji wykorzystania operatu grupującego jednostki indywidualne konieczne jest podanie informacji o metodach wielostopniowej selekcji jednostek

(w tym o sposobach terenowej kontroli takiej selekcji) oraz opisanie procedur ważenia danych wyrównujących nierówne szanse losowania jednostek z zespołów o różnej liczbie elementów.

6. W praktyce badań sondażowych wyróżnić można trzy typy prób badawczych określonych z uwagi na charakter wykorzystanych operatów: (1) adresowe próby budynków mieszkalnych, (2) próby gospodarstw domowych oraz (3) próby imienne. W przypadku dwóch pierwszych typów prób badawczych zachodzi konieczność wielostopniowego losowania jednostek populacji. Chociaż w praktyce dostępnych jest wiele procedur służących zrandomizowanemu lub quasi-losowemu doborowi jednostek z wylosowanych zespołów, to jednak kluczową rolę w zapewnieniu odpowiedniego poziomu reprezentatywności próby odgrywa terenowa kontrola procesu selekcji. Przeprowadzone w tej pracy analizy pokazały wyraźnie, że zastosowanie prób adresowych (budynków lub gospodarstw domowych) oraz związany z tymi próbami etap wielostopniowej selekcji jednostek prowadzić może do znacznego zniekształcenia próby badawczej poprzez dobór osób łatwiej dostępnych oraz bardziej skłonnych do udziału w badaniu.

Rekomendacje w zakresie oceny efektywności schematów doboru sondażowych prób badawczych:

1. Proces próbkowania reprezentatywnego powinien zostać uznany – sam z siebie – za główne źródło błędu losowego przekładającego się na precyzję pomiaru.
2. Precyzję pomiaru zmniejsza lub zwiększa zastosowany schemat doboru próby badawczej.
   W metodologii badań sondażowych wyróżnia się trzy główne schematy losowania prób badawczych różniące się od (uznawanego za wzorcowy) schematu losowania prostego. Zalicza się do nich: (1) dobór warstwowy, (2) zespołowy oraz (3) losowanie z nierównymi prawdopodobieństwami selekcji. W praktyce schematy losowania mogą przyjmować niezwykle złożone formy. A zatem, jeżeli nie zastosowano schematu doboru próby prostej, niezbędne jest opisanie różnic pomiędzy wybranym schematem doboru próby oraz charakterystykami losowania prostego.
3. Nie ma większego uzasadnienia dla podawania wielkości błędów statystycznych (wyznaczanych z założeniem losowania zgodnego ze schematem próby prostej) bez wcześniejszego zweryfikowania efektywności zastosowanego schematu doboru próby.
   Przedstawiając informacje dotyczące błędów statystycznych (precyzji pomiaru), należy opisać relacje zachodzące pomiędzy wielkością próby

dobranej zgodnie z wybranym schematem oraz odpowiadającą jej efektywną liczebnością prostej próby losowej. Minimalna liczebność wylosowanej próby badawczej powinna uwzględniać (uśrednioną) efektywność wybranego schematu doboru próby. Wielkości błędów statystycznych powinny zostać wyznaczone nie tylko dla estymatorów wskaźników struktury, ale również dla statystyk innego typu (np. średnich, median, stosunków itd.).

4. W analizach efektywności schematu doboru próby można (za punkt wyjścia) przyjąć metodę oceny polegającą na niezależnej od siebie estymacji mierników efektu doboru warstwowego, zespołowego oraz/lub z nierównymi prawdopodobieństwami selekcji. Szczegółowa analiza efektywności schematu losowania powinna zostać jednak poprzedzona weryfikacją kryteriów uprawomocniających określony sposób oceny jego efektywności oraz prowadzić do wykorzystania mierników posiadających umocowanie metodologiczne.

Rekomendacje w zakresie oceny terenowej realizacji sondażowej próby badawczej:

1. Niepełna realizacja próby badawczej (braki danych na poziomie jednostki oraz pojedynczych pytań) powinna zostać uznana za poważne źródło błędu systematycznego oraz znaczące źródło błędu losowego skutkującego obniżeniem precyzji estymacji.
2. Minimalna liczebność wylosowanej próby badawczej powinna uwzględniać przewidywany stopień jej terenowej realizacji. Uzupełnianie jednostek niedostępnych innymi osobami (dobieranymi celowo lub kwotowo) nie ma żadnego umocowania metodologicznego.
3. Informacje o wartościach wskaźników realizacji próby badawczej (w tym o wartościach wskaźników kontaktu, kooperacji, odmów itd.) powinny zostać wyznaczone nie tylko na poziomie całej próby badawczej, ale również w kluczowych (zarówno dla opisu przedmiotu badania, jak i charakterystyki mechanizmu niedostępności jednostek) warstwach populacji.
4. Postrealizacyjna klasyfikacja jednostek próby badawczej oraz procedury wyznaczania wskaźników realizacji próby powinny opierać się na ujednoliconych standardach wypracowanych w literaturze światowej. Zabieg taki umożliwi międzysurveyowe porównywanie wskaźników realizacji próby.
5. Należy precyzyjnie określić relacje zachodzące pomiędzy próbą wylosowaną oraz zrealizowaną poprzez podanie informacji na temat frakcji: (1) respondentów (tj. osób dostępnych), (2) jednostek niedostępnych

(w podziale na główne przyczyny niezrealizowania wywiadów, tj. (a) brak kontaktu, (b) odmowę, (c) inny powód niedostępności), (3) osób, co do których nie udało się zweryfikować statusu ich przynależności do populacji docelowej, (4) jednostek zidentyfikowanych w trakcie badań jako elementy nienależące do populacji docelowej (jednostki nadmiarowo pokryte przez operat).

Postrealizacyjna klasyfikacja jednostek próby badawczej (zwłaszcza w zbiorze osób niedostępnych) ma uzasadnienie merytoryczne wynikające z tego, że mechanizm niedostępności oddziałujący na gotowość jednostek do udziału w badaniu ma odmienną naturę od mechanizmu warunkującego możliwość dotarcia do wylosowanych osób. Jednostki niedostępne nie tworzą kategorii homogenicznej.

6. Terenowa faza badań powinna zostać scharakteryzowana poprzez podanie informacji na temat: (a) liczby podejmowanych prób nawiązania kontaktu z osobami wylosowanymi do próby (w doborze dwustopniowym informacja taka powinna uwzględniać również wizytę, podczas której spisywano jednostki oraz prowadzono selekcję wewnątrzzespołową), a także (b) liczby wizyt (licząc od nawiązania kontaktu) podjętych przez ankieterów w celu nakłonienia wylosowanych respondentów do udziału w badaniu. Należy również wyznaczyć wskaźniki: (a) wysokiej/niskiej dostępności, (b) wysokiej/niskiej gotowości do kooperacji, (c) odmów udziału w badaniu itd. Wielkości te pozwalają ocenić jakość terenowej realizacji próby sondażowej.
7. Przyjmując probabilistyczny paradygmat błędu niepełnej realizacji próby badawczej – w którym każdej wylosowanej osobie przypisuje się niezerową skłonność do udziału w badaniu – konieczne jest określenie zestawu tak zwanych zmiennych pomocniczych służących estymacji jednostkowych skłonności do udziału w badaniu. Oszacowania takie powinny zostać następnie wykorzystane w celu identyfikacji mechanizmu niedostępności jednostek oraz oceny poziomu reprezentatywności próby (z zastrzeżeniem jednak, że wnioskowanie ogranicza się do rozdzielczości zmiennych pomocniczych).
8. Należy opisać procedury wykorzystane w celu zwiększenia odsetka realizacji próby (tj. wskaźników kontaktu oraz kooperacji), a także monitorować, czy działania takie przekładają się na zmniejszenie zróżnicowania w zbiorze jednostkowych skłonności do udziału w badaniu (tj. czy zmniejszają ryzyko błędu systematycznego będącego skutkiem niepełnej realizacji próby).
9. Należy scharakteryzować procedury wykorzystane w celu szacowania wielkości błędów systematycznych będących następstwem niepełnej

realizacji próby badawczej, a także przedstawić i opisać zmienne najbardziej narażone na takie błędy.

10. Należy opisać procedury – ważenia i/lub imputacji danych – wykorzystane w celu ograniczenia negatywnych skutków niepełnej realizacji próby. Konieczne jest zweryfikowanie skuteczności takich procedur w eliminacji błędu systematycznego lub przynajmniej skontrolowanie ich przełożenia na precyzję pomiaru.

Literatura cytowana

Akacha, Mouna, Jane L. Hutton, 2010, *Modelling the Rate of Change in a Longitudinal Study with Missing Data, Adjusting for Contact Attempts*, „Statistics in Medicine", nr 30(10), s. 1072–1089

Aliaga, Alfredo, Ruilin Ren, 2006, *Optimal sample sizes for two-stage cluster sampling in demographic and health surveys*, „Demographic and Health Surveys Working Papers", nr 30, s. 1–18

Alreck, Pamela, Robert Settle, 1995, *The Survey Research Handbook*, McGraw-Hill/Irwin Series in Marketing, New York

Alwin, Duane F., 2007, *Margin of Errors. A Study of Reliability in Survey Measurement*, John Wiley & Sons, Inc., New York

Andresen, Ronald, Judith Kasper, Martin R. Frankel, 1979, *Total Survey Error*, Jossey-Boss Publishers, San Francisco

Atkinson, Dale, Dennis Schwanz, Karl W. Sieber, 1999, *Reporting sources of error in analytic publications*, „Statistical Policy Working Paper 28: Seminar on Interagency Coordination and Cooperation", U.S. Office of Management and Budget, Washington, DC, s. 329–341

Atrostic, Barbara N., Nancy Bates, Geraldine Burt, Adriana Silberstein, 2001, *Nonresponse in U.S. Government Household Surveys: consistent measures, recent trends, and new insights*, „Journal of Official Statistics", nr 17(2), s. 209–226

Babbie, Earl, 2007, *The Practice of Social Research*, 11thed., Thomson Learning Inc., Belmont

Babiński, Grzegorz, 1980, *Wybrane zagadnienia z metodologii socjologicznych badań empirycznych*, Wydawnictwo Uniwersytetu Jagiellońskiego, Kraków

Babiński, Grzegorz, 2004, *Metodologia a rzeczywistość społeczna. Dylematy badań etnicznych,* Zakład Wydawniczy NOMOS, Kraków

Baillargeon, Sophie, Louis-Paul Rivest, 2011, *The Construction of stratified design in R with the package stratification*, „Survey Methodology", nr 37(1), s. 53–65

Bankier, Michael D., 1986, *Estimators based on several stratified sample with applications to multiple frame surveys*, „Journal of the American Statistical Association", nr 81(396), s. 1074–1079

Barnett, Vic, 1974, *Elements of Sampling Theory*, English Universities Press, London

Barnett, Vic, 1982, *Elementy teorii pobierania prób,* Państwowe Wydawnictwo Naukowe, Warszawa

Bates, Nancy, James Dohlhamer, Eleanor Singer, 2008, *Privacy concerns, too busy, or just not interested: using doorstep concersn to predict survey nonresponse*, „Journal of Official Statistics", nr 24(4), s. 591–612

Batorski, Dominik, 2011, *Korzystanie z technologii informacyjno-komunikacyjnych*, „Contemporary Economics", nr 5(3), s. 299–327

Beaumont, Jean-François, 2005, *On the use of data collection process information for the treatment of unit nonresponse thought weight adjustment*, „Survey Methodology", nr 31(2), s. 227–231

Berger, Yves G., Yves Tillé, 2009, *Sampling with unequal probabilities*, w: *Sample Surveys: Design, Methods and Applications, Handbook of Statistics*, (red.) D. Pfeffermann, R.C. Calyampudi, nr 29A, s. 39–54
Best, Samuel C., Brian Krueger, 2002, *New approaches to assessing opinion: the prospects for electronic mail surveys*, „Journal of Public Opinion Research", nr 14(1), s. 73–92
Bethlehem, Jelke G., 1988, *Reduction of nonresponse bias thought regression estimation*, „Journal of Official Statistics", nr 4(3), s. 251–260
Bethlehem, Jelke G., 1999, *Cross-sectional research*, w: *Research Methodology in the Social, Behavioural and Life Sciences*, (red.) H.J. Adér, G.J. Mellenbergh, SAGE Publications, Inc., London, s. 110–142
Bethlehem, Jelke G., 2002, *Weighting nonresponse adjustments based on auxiliary information*, w: *Survey Nonresponse*, (red.) R. Groves, D.A. Dillman, J.L. Eltinge R.J.A. Little, John Wiley & Sons, Inc., New York, s. 275–287
Bethlehem, Jelke, Fannie Coben, Barry Schouten, 2008, *Indicators for the representativeness of survey response*, „Proceedings of Statistics Canada Symposium 2008", s. 1–8
Bethlehem, Jelke G., H.M.N. Kersten, 1985, *On the treatment on nonresponse in sample surveys*, „Journal of Official Statistics", nr 1(3), s. 287–300
Bethlehem, Jelke G., H.M.N. Kersten, 1986, *Werken met non-respons*, University of Amsterdam, Amsterdam
Biemer, Paul P., 2001, *Nonresponse bias and measurement bias in a comparison of face to face and telephone interviewing*, „Journal of Official Statistics", nr 17(2), s. 295–320
Biemer, Paul P., 2010a, *Total survey error. design, implementation, and evaluation*, „Public Opinion Quarterly", nr 74(5), s. 817–848
Biemer, Paul P., 2010b, *Overview of design issues: total survey error*, w: *Handbook of Survey Research*, (red.) P.V. Mardsen, J.D. Wright, Emerald Group Publishing Limited, New York City, NY s. 27–58
Biemer, Paul P., 2011, *Latent Class Analysis of Survey Error*, John Wiley & Sons, Inc., New York
Biemer, Paul P., Robert M. Groves, Nancy A. Mathiowetz, Seymour Sudman (red.), 1991, *Measurement Errors in Surveys*, John Wiley & Sons, Inc., New York
Biemer, Paul P., Lars E. Lyberg, 2003, *Introduction to Survey Quality*, John Wiley & Sons, Inc., New York
Billiet, Jaak, Hodeko Matsuo, Koen Beullens, Vasja Vehovar, 2009, *Non-response bias in cross-national surveys: design for detection and adjustment in the ESS*, „ASK. Research&Methods", nr 18(1), s. 3–43
Billiet, Jaak, Hodeko Michael Philippens, Rory Fitzgerald, Ineke Stoop, 2007, *Estimation of nonresponse bias in the European social survey: using information from reluctant respondents*, „Journal of Official Statistics", nr 23(2), s. 135–162
Binson, Diane, Jesse A. Canchola, Joseph A. Catania, 2000, *Random selection in a National Telephone Survey: a comparison of the Kish, Next-Birthday, and Last-Birthday methods*, „Journal of Official Statistics", nr 16(1), s. 53–59
Blumberg, Tephen J., Julina V. Luke, 2007, *Coverage bias in traditional telephone surveys of low-income and young adults*, „Public Opinion Quarterly", nr 71(5), s. 734–749
Bohrnstedt, George W., 1983, *Measurement*, w: *Handbook of Survey Research*, (red.) P.H. Rossi, J.D. Wright, A.B. Anderson, Academic Press, New York, s. 70–122
Brick, Michael J., 2008, *Random digital dialling*, w: *Encyclopedia of Survey Research Methods* (red.) P.J. Lavrakas, SAGE Publications, Inc., Thousand Oaks, CA, s. 675–678
Brick, Michael J., 2011, *Nonsampling errors in dual frame telephone surveys*, „Survey Methodology", nr 37(1), s. 1–12

Brick, Michael J., 2013, *Unit nonresponse and weighting adjustment: a critical review*, „Journal of Official Statistics", nr 29(3), s. 329–353
Brick, Michael J., Michael E. Jones, 2008, *Propensity to respond and nonresponse bias*, „METRON", nr LXVI(1), s. 51–73
Brick, Michael J., Graham Kalton, 1996, *Handling missing data in survey research*, „Statistical Methods in Medical Research", nr 5(3), s. 215–238
Brick, Michael J., Jill M. Montaquila, 2009, *Nonresponse and weighting*, w: *Sample Surveys*: *Design, Methods and Applications*, Handbook of Statistics, (red.), D. Pfeffermann, R.C. Calyampudi, Nr 29A, s. 163–246
Brick, Michael J., Joseph Waksberg, Dale Kulp, Amy Starer, 1995, *Bias in list-assisted telephone samples*, „Public Opinion Quarterly", nr 59(2), s. 218–235
Brill, Jonathan E., 2008, *Representative sample*, w: *Encyclopedia of Survey Research Methods*, (red.) P.J. Lavrakas, SAGE Publications, Inc., Thousand Oaks, CA, s. 101–102
Brooks, Camilla A., Barbara Bailar, 1978, *An Error Profile: Employment as Measurer by the Current Population Survey*, Statistical Policy Working Paper 3, Office of Federal Statistical Policy and Standards, U.S. Department of Commerce, Washington
Bryant, Barbara E., 1975, *Respondent selection in a time of changing household composition*, „Journal of Marketing Research", nr 12, s. 129–135
Brzezińska, Anna I., Jerzy Brzeziński, 2011, *Skale szacunkowe w badaniach diagnostycznych*, w: *Metodologia badań psychologicznych*, (red.) J. Brzeziński, Wydawnictwo Naukowe PWN, Warszawa, s. 299–399
Brzeziński, Jerzy, Anna I. Brzezińska, 1984, *Elementy metodologii badań psychologicznych*, Wydawnictwo Naukowe PWN, Warszawa
Buskirk, Trent, D., 2008, *Dual-frame sampling*, w: *Encyclopedia of Survey Research Methods*, (red.) J.P. Lavrakas, SAGE Publications, Inc., Thousand Oaks, CA, s. 212–215
Busse, Britta, Marek Fuchs, 2012, *The components of landline telephone survey coverage bias. the relative importance of no-phone and mobile-only populations*, „Quality&Quantity", nr 46(1), s. 1209–1225
Campbell, Donald T., Donald W. Fiske, 1959, *Convergent and doscriminant validation by the multitriat – multimethod matrix*, „Psychological Bulletin", nr 56, s. 81–105
Cannell, Charles, Floyd Fowler, 1963, *A Study of the Reporting Visit to Doctors in the National Health Surveys*, Research Report, Survey Research Centre, Ann Arbor, Michigan
Centrum Badania Opinii Społecznej, 2010, *Po wyborach prezydenckich*, raport nr BS/110/ 2010, Fundacja Centrum Badania Opinii Społecznej, Warszawa
Christman, Mary C., 2009, *Sampling of rare populations*, w: *Sample Surveys: Design, Methods and Applications, Handbook of Statistics*, (red.) D. Pfeffermann, R.C. Calyampudi, Nr 29A, s. 109–124
Chromy, James R., 2008, *Probability proportional to size (PPS) sampling*, w: *Encyclopedia of Survey Research Methods*, (red.) P.J. Lavrakas, SAGE Publications, Inc., Thousand Oaks, CA, s. 619–621
Chu, Adam, Michael J. Brick, Graham Kalton, 1999, *Weights for combining surveys across time or space*, „Bulletin of the International Statistical Institute", Contributed Papers, nr 2, s. 103–104
Cichocki, Piotr, Piotr Jabkowski, 2009, *Poczucie związku z UE w nowych krajach członkowskich Europy Środkowo-Wschodniej*, „Przegląd Zachodni", nr 3, s. 191–212
Cichomski, Bogdan, Tomasz Jerzyński, Marcin Zieliński, 2009, *Polskie Generalne Sondaże Społeczne: struktura skumulowanych wyników badań 1992–2008*, Instytut Studiów Społecznych, Uniwersytet Warszawski, Warszawa

Clark, Robert G., 2013, *Sample design using imperfect design data*, „Journal of Survey Statistics and Methodology", nr 1, s. 6–23
Clark, Robert G., David G. Steel, 2002, *The effect of using household as a sampling unit*, „International Statistical Review", nr 70(2), s. 289–314
Cobben, Fannie, Barry Schouten, 2008, *An empirical validation of R-indicators*, Discussion paper 08006, Statistics Netherlands, Voorburg/Heerlen
Cochran, William G., 1968, *The planning of observational studies of human population*, „Biometrics", nr 24(2), s. 295–313
Cochran, William G., 1977, *Sampling Techniques*, John Wiley & Sons, Inc., New York
Couper, Mick P., Arie Kapteyn, Matthias Schonlau, Joachim Winter, 2007, *Noncoverage and nonresponse in an Internet survey*, „Social Science Research", nr 36(1), s. 131–148
Cox, Brenda G., 2008, *Target populaton*, w: *Encyclopedia of Survey Research Methods*, (red.) P.J. Lavrakas, SAGE Publications, Inc., Thousand Oaks, CA, s. 875–876
Cronbach, Lee J., 1951, *Coefficient alpha and the iinternal structure of tests*, „Psychometrika", nr 16, s. 297–334
Curtin, Richard, Stanley Presser, Eleanor Singer, 2000, *The effects of response rate changes on the index of consumer sentiment*, „Public Opinion Quarterly", nr 64(2), s. 413–428
Curtin, Richard, Stanley Presser, Eleanor Singer, 2005, *Changes in telephone survey nonresponse over the past quarter century*, „Public Opinion Quarterly", nr 69(1), s. 87–98
Czaja, Ronald, Johnny Blairm, Jutta P. Sebestik, 1982, *Respondent selection in a telephone survey: a comparison of three techniques*, „Journal of Marketing Research", nr 19, s. 381–385
Czapiński, Janusz, Tomasz Panek, 2009, *Diagnoza Społeczna 2009. Warunki i jakość życia Polaków*, Wyższa Szkoła Finansów i Zarządzania w Warszawie, Warszawa
Dalenius, Tore, 1950, *The problem of optimum stratification*, „Scandinavian Actuarial Journal", nr 3–4, s. 203–213
Dalenius, Tore, Joseph L. Hodges Jr, 1959, *Minimum variance stratification*, „Journal of the American Statistical Association", nr 54(285), s. 88–101
Daniłowicz, Paweł, Franciszek Sztabiński, 1992, *Nowe spojrzenie na ankietę pocztową. Jak uzyskano 70% zwrotów*, w: *Analizy prób i technik badawczych w socjologii*, t. IX: *Problemy humanizacji procesu badawczego*, (red.) Z. Gostkowski, Wydawnictwo IFiS PAN, Warszawa, s. 122–141
Davern, Edward, 2008, *Representative sample*, w: *Encyclopedia of Survey Research Methods*, (red.) P.J. Lavrakas, SAGE Publications, Inc., Thousand Oaks, CA, s. 720–722
Davison, Anthony C, Sylvain Sardy, 2007, *Resampling variance estimation in survey with missing data*, „Journal of Official Statistics", nr 23(3), s. 371–386
da Silva, Damião N., Jean D. Opsomer, 2004, *Properties of the weighting cell estimator uner a nonparametric response mechanism*, „Survey Methodology", nr 30(1), s. 45–55
da Silva, Damião N., Jean D. Opsomer, 2006, *A Kernel smoothing method of adjusting for unit non-response in sample surveys*,: „The Canadian Journal of Statistics", nr 34(4), s. 563–579
da Silva, Damião N., Jean D. Opsomer, 2009, *Nonparametric propensity weighting for survey nonresponse through local polynomial regression*, „Survey Methodology", nr 35(2), s. 165–176
de Heer, Wim, 1999, *International response trends: results of an international survey*, „Journal of Official Statistics", nr 15(2), s. 129–142
de Leeuw, Edith D., Joop Hox, Mark Huisman, 2003, *Prevention and treatment of item nonresponse*, „Journal of Official Statistics", nr 19(2), s. 153–176

Deming, Edwards W., 1944, *On errors in surveys*, „American Sociological Review", nr 9(4), s. 359–369
Demnati, Abdellatif, J.N.K. Rao, 2004, *Linearization variance estimators for survey data*, „Survey Methodology", nr 30(1), s. 17–26
Designing Household Survey Samples: Practical Guidelines, 2005, Studies in Methods, Seria F, nr 98, Department of Economic and Social Affairs, United Nations, New York City, NY
Dever, Jill A., Ann Rafferty, Richard Valliant, 2008, *Internet surveys: can statistical adjustments eliminate coverage bias?*, „Survey Research Methods", nr 2(2), s. 47–62
Dillman, Don A., 1991, *The design and administration of mail surveys*, „Annual Review of Sociology", nr 17, s. 225–249
Dillman, Don A., John L. Eltinge, Robert M. Groves, Roderick J.A. Little, 2002, *Survey nonresponse in design, data collection, and analysis*, w: *Survey Nonresponse*, (red.) R. Groves, D.A. Dillman, J.L. Eltinge R.J.A. Little, John Wiley & Sons, Inc., New York, s. 3–26
Domański, Henryk, 1999, *Jednostki niedostępne. Problem wpływu na wyniki badań*, „ASK. Społeczeństwo. Badania. Metody", nr 8, s. 67–92
Domański, Henryk, 2006, *Liczba wizyt i czas trwania badania*, „ASK. Społeczeństwo. Badania. Metody", nr 15, s. 29–49
Domański, Henryk, Zbigniew Sawiński, Kazimierz M. Słomczyński, 2007, *Nowe klasyfikacje i skale zawodów*, Wydawnictwo IFiS PAN, Warszawa
Dorofeev, Sergey, Peter Grant, 2006, *Statistics for Real-Life Sample Surveys. Non-Simple-Random Samples and Weighted Data*, Cambridge University Press, Cambridge
Duncan, Kristin B., Elizabeth A. Stasny, 2001, *Using propensity scores to control coverages bias in telephone surveys*, „Survey Methodology", nr 27(2), s. 121–130
Durrant, Gabriele B., 2009, *Imputation methods for handling item-nonresponse in practice: methodological issues and recent debates*, „International Journal of Social Research Methodology", nr 12(4), s. 293–304
Durrant, Gabriele B., Chris Skinner, 2006, *Using missing data methods to correct for measurement error in a distribution function*, „Survey Methodology", nr 32(1), s. 25–36
Eckman, Stephanie, Colm O'Muircheartaigh, 2011, *Performance of the half-open interval missed housing unit procedure*, „Survey Research Methods", nr 5(3), s. 125–131
Eldridge, Sandra M., Obioha C. Ukoumunne, John B. Carlin, 2009, *The intra-cluster correlation coefficient in cluster randomized trials: a review of definitions*, „International Statistical Review", nr 77(3), s. 378–394
European Social Survey: ESS1, 2002, *Sampling for the European Social Survey – Round 1*, European Social Data Archive, Norwegian Social Science Data Service, Bergen, http://www.europeansocialsurvey.org/docs/round1/methods/ESS1_sampling_report.pdf [data pobrania danych 24 października 2012 roku]
European Social Survey: ESS2, 2004, *Sampling for the European Social Survey – Round 2*, European Social Data Archive, Norwegian Social Science Data Service, Bergen, http://www.europeansocialsurvey.org/docs/round2/methods/ESS2_sampling_guidelines.pdf [data pobrania danych 24 października 2012 roku]
European Social Survey: ESS4, 2008, *Documentation Report. Edition 5.0*, Bergen, European Social Data Archive, Norwegian Social Science Data Service, http://www.europeansocialsurvey.org/docs/round4/survey/ESS4_data_documentation_report_e05_3.pdf [data pobrania danych 24 października 2012 roku]
European Social Survey: ESS5, 2009, *Specification for Participating Countries*. Centre for Comparative Social Surveys, City University London, London, http://www.europeansocial

sur vey.org/ docs/round5/methods/ESS5_project_specification.pdf [data pobrania danych 22 października 2012 roku]
European Social Survey: ESS5, 2010, *Documentation Report. Edition 2.0*, Bergen, European Social Data Archive, Norwegian Social Science Data Service, http://www.europeansocialsurvey.org/docs/round5/survey/ESS5_data_documentation_report_e02_0.pdf [data pobrania danych 22 października 2012 roku]
European Social Survey: ESS5, 2010, *Sampling design in ESS5-PL*, Ośrodek Realizacji Badań Społecznych IFiS PAN, Warszawa
European Social Survey: ESS5, 2010, *Sampling for the European Social Survey Round V: Principles and Requirements*, The Sampling Expert Panel of the ESS, GESIS, Mannheim, http://www.europeansocialsurvey.org/docs/round5/methods/ESS5_sampling_guidelines.pdf [data pobrania danych 7 czerwca 2010 roku]
European Social Survey: ESS6, 2012, *Sampling for the European Social Survey Round VI: Principles and Requirements*, The Sampling Expert Panel of the ESS, GESIS, Mannheim, http://www.europeansocialsurvey.org/docs/round6/methods/ESS6_sampling_guidelines.pdf [data pobrania danych 22 października 2012 roku]
Fahimi, Mansour, 2008, *Cluster sample*, w: *Encyclopedia of Survey Research Methods*, (red.) P.J. Lavrakas, SAGE Publications, Inc., Thousand Oaks, CA, s. 98–99.
Fieldwork jest sztuką, (red.) P.B. Sztabiński, Z. Sawiński, F. Sztabiński, Wydawnictwo IFiS PAN, Warszawa
Flores-Cervantes, Ismael, Michael J. Brick, 2009, *Efficacy of poststratification in complex sample design*, „Survey Research Methods", JSM, s. 4642–4655
Frankfort-Nachmias, Chava, Davis Nachmias, 2001, *Metody badawcze w naukach społecznych*, Zysk i S-ka, Poznań
Fuchs, Marek, 2008, *Total survey error*, w: *Encyclopedia of Survey Research Methods*, (red.) P.J. Lavrakas, SAGE Publications, Inc., Thousand Oaks, CA, s. 896–902
Funkhouser, G. Ray, Edwin B. Parker, 1968, *Analyzing coding reliability: the random systematic-error coefficient*, „Public Opinion Quarterly", nr. 32(1), s. 122–128
Gabler, Siegfried, Öztas Ayhan, 2007, *Gewichtung bei Erhebungen im Festnetz und über Mobilfunk: Ein Dual-Frame-Ansatz*, w: *Mobilfunktelefonie – Eine Herausforderung für die Umfrageforschung, ZUMA-Nachrichten*, Spezial Band 13, (red.) S. Gabler, S. Häder, Mannheim, s. 39–46
Gabler, Siegfried, Matthias Ganninger, Sabine Häder, Ralf Munnich, 2008, *Design effect (*DEFF*)*, w: *Encyclopedia of Survey Research Methods*, (red.) P.J. Lavrakas, SAGE Publications, Inc., Thousand Oaks, CA, s. 193–197
Gabler, Siegfried, Sabine Häder, Partha Lahiri, 1999, *A model based justification of Kish's formula for design effects for weighting and clustering*, „Survey Methodology", nr 25(1), s. 105–106
Gabler, Siegfried, Sabine Häder, Peter Lynn, 2006, *Design effects for multiple design samples*, „Survey Methodology", nr 32(1), s. 115–120
Ganninger, Matthias, 2008, *The ESS sample design data file (SDDF)*, Documentation of the European Social Survey, http://www.europeansocialsurvey.org/docs/round1/methods/ESS1_sddf_documentation.pdf (data pobrania: 25 czerwca 2014 roku)
Gaziano, Cecilie, 2005, *Comparative analysis of within-household respondent selection techniques*, „Public Opinion Quarterly", nr 69(1), s. 124–157
Gaziano, Cecilie, 2008, *Within-unit coverage*, w: *Encyclopedia of Survey Research Methods*, (red.) P.J. Lavrakas, SAGE Publications, Inc., Thousand Oaks, CA, s. 961–962

Gillikin, Jason E., 2008, *Interpenetrated design*, w: *Encyclopedia of Survey Research Methods*, (red.) P.J. Lavrakas SAGE Publications, Inc., Thousand Oaks, CA, s. 359–360
Goldberg, Caren B., 2003, *Who responds to surveys?: Assessing the effects of nonresponse in cross-sectional dyadic research*, „Assessment", nr 10(1), s. 41–48
Goyder, John, 1987, *The Silent Minority. Nonrespondents on Sample Surveys*, Polity Press, Cambridge
Goyder, John, Jean Lock, Trish McNair, 1992, *Urbanization effects on survey nonresponse: a test within and across cities*, „Quality & Quantity", nr 26, s. 39–48
Groves, Robert M., 1989, *Survey Errors and Survey Costs*, John Wiley & Sons, Inc., New York
Groves, Robert M., 2006, *Nonresponse rates and nonresponse bias in household surveys*, „Public Opinion Quarterly", nr 70(5), s. 646–675
Groves, Robert M., Mick P. Couper, 1995, *Theoretical motivation for post-survey nonresponse adjustment in household surveys*, „Journal of Official Statistics", nr 11(1), s. 93–106
Groves, Robert M., Mick P. Couper, 1998, *Nonresponse in Household Interview Surveys*, John Wiley & Sons, Inc., New York
Groves, Robert M., Mick P. Couper, Stanley Presser, Eleanor Singer, Roger Tourangeau, Giorgina Piani Acosta, Lindsay Nelson, 2006, *Experiments in producing nonresponse bias*, „Public Opinion Quarterly", nr 70(5), s. 720–736
Groves, Robert M., Don. A. Dillman, John L. Eltinge, Roderick J.A. Little, 2002, *Survey Nonresponse,* John Wiley & Sons, Inc., New York
Groves, Robert M., Emilia Peytcheva, 2008, *The impact of nonresponse rates on nonresponse bias. A meta-analysis*, „Public Opinion Quarterly", nr 72(2), s. 167–189
Groves, Robert M., Floyd J. Fowler, Mick P. Couper, James M. Lepkowski, Elanor Singer, Roger Tourangeau, 2004, *Survey Methodology*, John Wiley & Sons, Inc., New York
Groves, Robert M., Lars E. Lyberg, 1988, *An overview of nonresponse issues in telephone surveys*, w: *Telephone Survey Methodology*, (red.) R.M. Groves, P.P. Biemer, L.E. Lyberg, J.T. Massey, W.L. Nichols II, J. Waksberg, John Wiley&Sons Inc., New York, s. 191–212
Groves, Robert M., Lars E. Lyberg, 2010, *Total survey error. past, present, and future*, „Public Opinion Quarterly", nr 74(5), s. 849–879
Groves, Robert M., Stanley Presser, Sarah Dipko, 2004, *The role of topic interest in survey participation decisions*, „Public Opinion Quarterly", nr 68(1), s. 2–31
Grzeszkiewicz-Radulska, Katarzyna, 2009, *Respondenci niedostępni w badaniach sondażowyc*h, „Analizy i Próby Technik Badawczych w Socjologii", t. XII, Wydawnictwo Uniwersytetu Łódzkiego, Łódź
Gunning, Patricia, Jane M. Horgan, 2004a, *A new algorithm for the construction of stratum boundaries in skewed populations*, „Survey Methodology", nr 30(2), s. 159–166
Gunning, Patricia, Jane M. Horgan, William Yancey, 2004b, *Geometric stratification of accounting data*, „Revista Contaduria Administration", nr 214, Septiembre-Diciembre 2004
Häder Sabine, Siegfred Gabler, 2003, *Sampling and estimation*, w: *Cross Cultural Survey Methods*, (red.) J.A. Harkness, F.J.R. van de Vijver, P.Ph. Mohler, John Wiley & Sons, Inc., New York, s. 117–136
Häder Sabine, Iris Lehnhoff, Elisabeth Mardian, 2010, *Mobile phone surveys: empirical findings from a research project*, „ASK. Research&Methods", nr 19(1), s. 3–19
Hagan, Dan E., Charlotte M. Collier, 1983, *Must respondent selection procedures for telephone surveys be invasive?*, „Public Opinion Quarterly", nr 47(4), s. 547–556
Haines, Dawn E., Kenneth H. Pollock, Sastry G. Pantula, 2000, *Population size and total estimation when sampling from incomplete list frames with heterogeneous inclusion probabilities*, „Survey Methodology", nr 26(2), s. 121–129

Hall, John, 2008, *Area probability sample*, w: *Encyclopedia of Survey Research Methods*, (red.) P.J. Lavrakas, SAGE Publications, Inc., Thousand Oaks, CA, s. 33–36
Hansen, Morris H., William N. Hurwitz, 1946, *The problem of non-response in sample surveys*, „Journal of the American Statistical Association", nr 41(236), s. 517–529
Hartley, Herman O., 1962, *Multiple frame surveys*, „Proceedings of the Social Statistics Section", American Statistical Association, s. 203–206
Hartley, Herman O., 1974, *Multiple frame methodology and selected applications*, „The Indian Journal of Statistics ", nr 36(3), s. 99–118
Haziza, David, 2009, *Imputation and inference in the presence of missing data*, w: *Sample Surveys: Design, Methods and Applications, Handbook of Statistics*, (red.) D. Pfeffermann, R.C. Calyampudi, Nr 29A, s. 215–246
Hedlin, Dan, 2000, *A procedure for stratification by an extended ekman rule*, „Journal of Official Statistics", nr 16(1), s. 15–29
Heerwegh, Dirk, Geert Loosveldt, 2008, *Face-to-face versus web surveying in a high-internet coverage population: difference in response quality*, „Public Opinion Quarterly", nr 72(5), s. 836–846
Henry, Gary T., 1990, *Practical Sampling*, Sage Publications Inc., Thousand Oaks, CA
Hidiroglou, Michael A., Douglas J. Drew, Gerald B. Gray, 1993, *A framework for measuring and reducing nonresponse in surveys*, „Survey Methodology", nr 19(1), s. 81–94
Holt, Daniel, Fred T.M. Smith, *Post stratification*, „Journal of the Royal Statistical Society. Series A", nr 142(1), s. 44–46
Hornowska, Elżbieta, 2007, *Testy psychologiczne. Teoria i praktyka*, Wydawnictwo Naukowe SCHOLAR, Warszawa
Horvitz, Daniel G., Donovan J. Thompson, 1952, *A generalization of sampling without replacement from a finite universe*, „Journal of the American Statistical Association", nr 47(260), s. 663–685
Hyman, Herbert H., 1954, *Interviewing in Social Research*, University of Chicago Press, Chicago
Iannacchione, Vincent G., 2003, *Sequential weight adjustments for the location and cooperation propensity for the 1995 National Survey of Family Growth*, „Journal of Official Statistics", nr 19(1), s. 31–43
Inkmann, Joachim, 2010, *Estimating firm size elasticities of product and process R&D*, „Economica", nr 77(306), s. 384–402
Jabkowski, Piotr, 2007, *Wpływ niezrealizowania części wywiadów na trafność wnioskowania statystycznego w badaniach społecznych*. *Technika wywiadu kwestionariuszowego oraz telefonicznego w świetle błędów nielosowych*, „ASK. Społeczeństwo. Badania. Metody", nr 16, s. 67–86
Jabkowski, Piotr, 2011, *Do more contact-attempts reduce non-response bias in representative face-to-face interviews? Findings from a PAPI Survey with a low response rate*, „ASK. Research&Methods", nr 20(1), s. 27–58
Jabkowski, Piotr, 2013, *How (not) to estimate the design effect of a complex sampling scheme: a case study of the Polish section of the European Social Survey, Round 5*, „ASK. Research&Methods", nr 22(1), s. 55–77
Jabkowski, Piotr, Piotr K. Potempa, 2010, *Wykorzystanie modeli skoringowych w obsłudze portfeli wierzytelności dłużników biznesowych*, w: Materiały na konferencję *Skoring w zarządzaniu ryzykiem*, Wydawnictwo StatSoft, Kraków, s. 59–72
Jäckle, Annette, Peter Lynn, Jennifer Sinibaldi, Sarah Tipping, 2013, *The effect of interviewer experience, attitudes, personality and skills on respondent co-operation with face-to-face surveys*, „Survey Research Methods", nr 7(1), s. 1–15

Johnson, Timothy P., Young Ik Cho, Richard T. Campbell, Allyson L. Holbrook, 2006, *Using community-level correlates to evaluate nonresponse effects in a telephone survey*, „Public Opinion Quarterly", nr 70(5), s. 704–719
Johnston, Ron, Richard Harris, 2006, *Do survey respondents and non-respondents differ? Ecological analyses of the 2005 British election study*, „International Journal of Market Research", nr 48(3), s. 277–303
Jowell, Roger, 2009, *How comparative is comparative research?*, „American Behavioural Scientist", nr 42(2), s. 168–167
Kahn Robert L., Charles F. Cannell, 1957, *The Dynamics of Interviewing: Theory, Technique, and Cases*, John Wiley & Sons, Inc., New York
Kaldjian, Lauric C., J.F. Jekel, J.L. Bernene, G.E. Rosenthal, M. Vaughau-Sarrazin, T.P. Duffy, 2004, *Internists' attitudes towards terminal sedation in end of life care*, „Journal of Medical Ethics", nr 30, s. 499–503
Kalsbeek, William D., 2008, *Stratified sampling*, w: *Encyclopedia of Survey Research Methods*, (red.) P.J. Lavrakas, SAGE Publications, Inc., Thousand Oaks, CA, s. 849–850.
Kalton, Graham, 1983, *Compensating for Missing Survey Data*, Institute for Social Research University of Michigan, Michigan
Kalton, Graham, 2009, *Methods for overssampling rare subpopulations in social surveys,* „Survey Methodology", nr 35(2), s. 125–141
Kalton, Graham, Dallas W. Anderson, 1986, *Sampling rare populations,* „Journal of the Royal Statistical Society", nr 159(1), s. 65–82
Kalton, Graham, Ismael Flores-Cervantes, 2003, *Weighting methods*, „Journal of Official Statistics", nr 19(2), s. 81–97
Kalton, Graham, Daniel Kasprzyk, 1986, *The treatment of missing survey data*, „Survey Methodology", nr 12(1), s. 1–16
Kaminska, Olena, Allan L. Mccutcheon, Jaak BIlliet, 2010, *Satosficing among reluctant respondenrs in a cross-national context*, „Public Opinion Quarterly", nr 74(5), s. 956–984
Kasprzyk, Daniel, Lee Giesbrecht, 2003, *Reporting sources of error in i.s. federal government surveys*, „Journal of Official Statistics", nr 19(4), s. 343–363
Keeter, Scott, Courtney Kennedy, Michael Dimock, Jonathan Best, Peyton Craighill, 2006, *Gauging the impact of growing nonresponse on estimate from a national rdd telephone survey*, „Public Opinion Quarterly", nr 70(5), s. 759–779
Keeter, Scott, Carolyn Miller, Andrew Kohut, Robert M. Groves, Stanley Presser, 2000, *Consequences of reducing nonresponse in a national telephone survey*, „Public Opinion Quarterly", nr 64(1), s. 125–148
Kendall, Maurice G., Alan Stuart, 1979, *The advanced theory of statistics. Vol. 2. Inference and relationship*, 4th ed., Griffin, London
Kersten, Hubert M.P., Jelke G. Bethlehem, 1984, *Exploring and reducing the nonresponse bias by asking the basic question*, „Statistical Journal of the United Nations Economic Commission for Europe", nr 2(4), s. 369–380
Kim, Sun W., 2008, *Half-open interval*, w: *Encyclopedia of Survey Research Methods*, (red.) P.J. Lavrakas, SAGE Publications, Inc., Thousand Oaks, CA, s. 310
Kish, Leslie, 1919, *Populations for survey sampling*, „Survey Statistician", nr 1, s. 14–15
Kish, Leslie, 1949, *A procedure for objective respondent selection within the household*, „Journal of the American Statistical Association", nr 44(247), s. 380–387
Kish, Leslie, 1962, *Studies of interviewer variance for attitudinal variables*, „Journal of the American Statistical Association", nr 57, s. 92–115
Kish, Leslie, 1965, *Survey Sampling*, John Wiley & Sons, Inc., New York

Kish, Leslie, 1987, *Statistical Design for Research*, John Wiley & Sons, Inc., New York
Kish, Leslie, 1992, *Weighting for unequal P_i*, „Journal of Official Statistics", nr 8(2), s. 183–200
Kish, Leslie, Martin R. Frankel, 1974, *Inference from complex samples*, „Journal of the Royal Statistical Society. Series B (Methodological)", nr 36(1), s. 1–37
Koch, Achim, Michael Blohm, 2009, *Item nonresponse in the European Social Survey*, „ASK Research & Methods", nr 18(1), s. 45–65
Kohler, Urlich, 2007, *Survey from inside: an assessment of unit nonresponse bias with internal criteria*, „Survey Research Methods", nr 2(1), s. 55–67
Korns, Alexander, 1977, *Coverage issues raised by comparisons between CPS and establishment employement*, „Proceedings of the Social Statistics Section, American Statistical Accociation", s. 60–69
Kowal, Jolanta, 1988, *Metody statystyczne w badaniach sondażowych rynku*, Wydawnictwo Naukowe PWN, Warszawa
Kozak, Marcin, 2004, *Optimal stratification using random search method in agricultural surveys,* „Statistics Transition", nr 6(5), s. 797–806
Kozak, Marcin, Med. Ram Verma, 2006, *Geometric versus optimization approach to stratification: a comparision of efficiency*, „Survey Methodology", nr 32(2), s. 157–163
Kreuter, Frauke, Urlich Kohler, 2009, *Analizing contact sequences in call record data. potential and limitation of sequence indicators for nonresponse adjustments in the European Social Survey*, „Journal of Official Statistics", nr 25(2), s. 203–226
Kreuter, Frauke, Gerrit Muller, Mark Trappmann, 2010, *Nonresponse and measurement error in employment research. making use of administrative data*, „Public Opinion Quarterly", nr 74(5), s. 880–906
Kreuter, Frauke, Kristen Olson, James Wagner, Ting Yan, Trena M. Ezzati-Rice, Carolina Casas-Cordero, Michael Lemay, Andy Peytchev, Robert. M. Groves, Trivellore E. Raghunathan, *Using proxy measures and other correlates of survey outcomes to adjust for nonresponse: examples from multiple surveys*, „Journal of the Royal Statistical Society. Series A", nr 173(2), s. 389–407
Krosnick, Jon A., 1999, *Survey research*, „Annual Review of Psychology", nr 50, s. 537–567
Kruskal, William, Frederick Mosteller, 1979a, *Representative sampling, i: non-scientific literature*, „International Statistical Review", nr 47(1), s. 13–24
Kruskal, William, Frederick Mosteller, 1979b, *Representative sampling, II: scientific literature, excluding statistics*, „International Statistical Review", nr 47(2), s. 111–127
Kruskal, William, Frederick Mosteller, 1979c, *Representative sampling, III: the current statistical literature*, „International Statistical Review", nr 47(3), s. 245–265
Kruskal, William, Frederick Mosteller, 1980, *Representative sampling, IV: the history of the concept in statistics, 1895–1939*, „International Statistical Review", nr 48(2), s. 169–195
Kuha, Jouni, David Firth, 2011, *On the index of dissimilarity for lack of fit in loglinear and log-multiplicative models*, „Computational Statistics and Data Analysis", nr 55(1), s. 375–388
Kwang, Jae, Jay J. Kim, 2007, *Nonresponse weighting adjustment using estimated response probability*, „The Canadian Journal of Statistic", nr 35(4), s. 501–514
Kviz, Frederick J., 1977, *Toward a standard definition of response rate*, „Public Opinion Quarterly", nr 41(2), s. 265–267
Laaksonen, Seppo, 2007, *Weighting for two-phase surveyed data*, „Survey Methodology", nr 33(2), s. 121–130
Lachapelle, Réjean, Don Kerr, 2000, *Census coverage error: A demographic evaluation*, „Survey Methodology", nr 26(1), s. 43–52

Laska, Eugene, Morris Meisner, 1993, *A plant-capture method for estimating the size of a population from a single sample*, „Biometrics", nr 49(1), s. 209–220
Lavallée, Pierre, 1995, *Cross-sectional weighting of longitudinal surveys of individuals and households using the weight share method*, „Survey Methodology", nr 21(1), s. 25–32
Lavallée, Pierre, Michel A. Hidiroglou, 1988, *On the stratification of skewed population*, „Survey Methodology", nr 14(1), s. 3–43
Lavrakas, Paul L., Sandra L. Bauman, Daniel M. Merkle, 1993, *The last-birthday selection method & within-unit coverage problems*, „Proceedings of the Section on Survey Reearch Methods", s. 1107–1112
Lavrakas, Paul L., Elizabeth A. Stasny, Brian Harpuder, 2000, *A further investigation of the last-birthday respondent selection method and within-unit coverage error*, „Proceedings of the Section on Survey Reearch Methods", s. 890–895
Lednicki, Bronisław, Robert Wieczorkowski, 2003, *Optimal stratification and sample allocation between subpopulations and strata*, „Statistics in Transition", nr 6(2), s. 287–305
Lee, Geon, 2008, *Network sampling*, w: *Encyclopedia of Survey Research Methods*, (red.) P.J. Lavrakas SAGE Publications, Inc., Thousand Oaks, CA, s. 506–507
Lee, Hyunshik, 2012, *How should one find out the contributions to the design effect (variance) made by each of the design components (stratification, clustering, weighting) of a complex sample design?*, „Survey Statistician", nr 66, s. 16–20
Lee, Sunghee, 2006, *Propensity score adjustment as a weighting scheme for volunteer panel web surveys*, „Journal of Official Statistics", nr 22(2), s. 329–349
Lee, Sunghee, Richard Valliant, 2008, *Weighting telephone samples using propensity scores*, w: *Advances in Telephone Survey Methodology*, (red.) J.M. Lepkowski, C. Tucker, J.M. Brick, E.D. De Leeuw, L.Japec, P.J. Lavrakas, M.W. Link, R.L. Sangster, John Wiley & Sons, Inc., New York, s. 170–183
Lepkowski, James M., 2008, *Population*, w: *Encyclopedia of Survey Research Methods*, (red.) P.J. Lavrakas, SAGE Publications, Inc., Thousand Oaks, CA, s. 590–591
Lepkowski, James M., Graham Kalton, Daniel Kasprzyk, 1989, *Weighting adjustments for partial nonresponse in the 1984 SIPP Panel*, „Proceedings of the Survey Research Methods Section", American Statistical Association, s. 296–301
Lepkowski, James M., William D. Mosher, Karen E. Davis, Robert M. Groves, John Van Hoewyk, 2010, *The 2006–2010 National Survey of Family Growth: sample design and analysis of a continuous survey*, „Vital Health Statistics", nr 2(150), s. 1–36
Lessler, Judith T., William D. Kalsbeek, 1987, 1992, *Nonsampling Error in Surveys*, John Wiley & Sons, Inc., New York
Liao, Dan, Richard Valliant, 2012, *Variance inflation factors in the analysis of complex survey data*, „Survey Methodology", nr 38(1), s. 53–62
Lin, I-Fen, Nora C. Schaeffer, *Using survey participants to estimate the impact of nonparticipation*, „Public Opinion Quarterly", nr 59(2), s. 236–258
Link, Michael W., Jennie W. Lai, 2011, *Cell-phone-only households and poblems of differential nonresponse using an address-based sampling design*, „Public Opinion Quarterly", nr 75(4), s. 613–635
Lissowski, Grzegorz, 1971, *Problem jednostek niedostępnych w reprezentatywnych badaniach socjologicznych*, w: *Metody matematyczne w socjologii – zagadnienia wybrane,* (red.) K. Szaniawski, Wydawnictwo IFiS PAN, Warszawa, s. 7–34
Lissowski, Grzegorz, Jacek Haman, Mikołaj Jasiński, 2008, *Podstawy statystyki dla socjologów*, Wydawnictwo Naukowe Scholar, Warszawa

Little, Roderick J.A., 1986, *Survey nonresponse adjustment for estimates of mean*, „International Statistical Review", nr 54(2), s. 139–157
Little, Roderick J.A., Donald Rubin, 1987, 2002, *Statistical Analysis with Missing Data*, John Wiley & Sons, Inc., New York
Little, Roderick J.A., Sonya Vartivarian, 2003, *On weighting the rates in non-response weights*, „Statistics in Medicine", nr 22(9), s. 1589–1599
Little, Roderick J.A., Sonya Vartivarian, 2005, *Does weighting for nonresponse increase the variance of surveys means?*, „Survey Methodology", nr 31(2), s. 161–168
Lohr, Sharon L., 1999, *Sampling: Design and Analysis*, Brools/Cole Publishing Company, Boston
Lohr, Sharon L., 2009, *Multiple-frame surveys*, *sample surveys: design, methods and applications, handbook of statistics*, (red.) D. Pfeffermann, R.C. Calyampudi, Nr 29A, s. 71–88
Lohr, Sharon L., 2011, *Alternative survey sample designs: sampling with multiple overlapping frames*, „Survey Methodology", nr 37(2), s. 197–213
Loosveldt, Geert, Nathalie Sonck, 2008, *An evaluation of the weighting procedures for an online access panel survey*, „Survey Research Methods", nr 2(2), s. 93–105
Luiten, Annemieke, 2011, *Personalisation in advance letters does not always increase response rates. Demographic correlates in a large scale experiment*, „Survey Research Methods", nr 5(1), s. 11–20
Luiten Annemieke, Barry Schouten, 2013, *Tailored fieldwork design to increase representative household survey response: an experiment in the survey of consumer satisfaction*, „Journal of Royal Statistics Society", nr 176(1), s. 169–189
Lundström, Sixten, Carl-Eric Särndal, 1999, *Calibration as a standard method for treatment on nonresponse*, „Journal of Official Statistics", nr 15(2), s. 305–327
Lutyńska, Krystyna, 1978, *Ankieterzy i badacze. Z badań nad wpływem ankieterskim*, „Przegląd Socjologiczny", nr 30, s. 143–173
Lutyńska, Krystyna, 1989, *Analiza odmów w polskich badaniach kwestionariuszowych w latach 1982–1985*, „Przegląd Socjologiczny", nr 37, s. 209–234
Lutyńska, Krystyna, 1993, *Surveye w Polsce. Spojrzenie socjologiczno-antropologiczne*, Wydawnictwo IFiS PAN, Warszawa
Lutyńska, Krystyna, 1997, *Wpływ ankieterski w pierwszej fazie badań kwestionariuszowych*, „ASK. Społeczeństwo. Badania. Metody", nr 1–2, s. 53–71
Lutyńska, Krystyna, 1998, *„Strategie" i postawy współczesnych ankieterów a reakcje i nowe obawy respondentów*, „ASK. Społeczeństwo. Badania. Metody", nr 7, s. 17–36
Lyberg, Lars E., 2012, *Survey quality*, „Survey Methodology", nr 38(2), s. 107–130
Lynn, Peter, 2003, *PEDAKSI: methodology for collecting data about survey non-respondents*, „Quality&Quantity", nr 37(3), s. 239–261
Lynn, Peter, Paul Clarke, Jean Martin, Patrick Sturgis, 2002, *The effect of extended interviewer efforts on nonresponse bias*, w: *Survey Nonresponse*, (red.) R. Groves, D.A. Dillman, J.L. Eltinge R.J.A. Little, John Wiley & Sons, Inc., New York, s. 135–147
Lynn, Peter, Siegfried Gabler, 2005, *Approximation of b^* in the prediction of design effects due to clustering*, „Survey Methodology" nr 31(1), s. 101–104
Lynn, Peter, Siegfried Gabler, Sabine Häder, Seppo Laaksonen, 2007, *Methods for achieving equivalence of samples in cross-national surveys*, „Journal of Official Statistics" nr 27(1), s. 107–124
Mabli, James, James C. Ohls, 2012, *Supplemental nutrition assistance program dynamics and employment transitions: the role of employment instability*, „Applied Economic Perspectives and Policy", nr 34(1), s. 187–213

Mahalanobis, Prasant C., 1946, *Recent experiments in statistical sampling in the Indian Statistical Institute*, „Journal of the Royal Statistical Society", nr 109, s. 325–378
Marella, Daniela, 2007, *Errors depending on costs in sample surveys*, „Survey Research Methods", nr 1(2), s. 85–96
Marker, David A., Don L. Stevens Jr, 2009, *Sampling and inference in environmental surveys*, w: *Sample Surveys*: *Design, Methods and Applications*, *Handbook of Statistics*, (red.) D. Pfeffermann, R.C. Calyampudi, Nr 29A, s. 487–512
Martin, Elizabeth, 1999, *Who knows who lives here? Within-household disagreements as a source of survey coverage error*, „Public Opinion Quarterly", nr 63(2), s. 220–236
Martin, Elizabeth, Eugene Laska, Kim Hopper, Morris Merisner, Joe Wanderling, 1997, *Issues in the use of a plant-capture method for estimating the size of the street dwelling population*, „Journal of Official Statistics", nr 13(1), s. 59–73
Martin, Peter, 2011, *A good mix? Mixed mode data collection and cross-national surveys*, „ASK Research & Methods", nr 20(1), s. 5–26
Mason, Robert, Virginia Lesser, Michael W. Traugott, 2002, *Effect of item nonresponse on nonresponse error and inference*, w: *Survey Nonresponse*, (red.) R. Groves, D.A. Dillman, J.L. Eltinge R.J.A. Little, John Wiley & Sons, Inc., New York, s. 149–161
Matsuo, Hideko, Jaaj Billiet, Geert Loosveldt, Frode Berglund, Øyven Kleven, 2010, *Measurement and adjustment of non-response bias based on non-response surveys: the case of Belgium and Norway in the European Social Survey Round 3*, „Survey Research Methods", nr 4(3), s. 165–178
Merkle, Daniel, Murray Edelman, 2002, *Nonresponse in exit polls: a comprehensive analysis*, w: *Survey Nonresponse*, (red.) R. Groves, D.A. Dillman, J.L. Eltinge R.J.A. Little, John Wiley & Sons, Inc., New York, s. 243–258
Micklewright, John, Schnepf, Sylke, Chris J. Skinner, 2012, *Non-response biases in surveys of schoolchildren: the case of the English Programme for International Student Assessment (PISA) samples*, „Journal of the Royal Statistical Society. Series A (Statistics in Society)", nr 175(4), s. 915–938
Montaquila, Lill M., J. Micheal Brick, Mary C. Hagedorn, Courtney Kennedy, Scott Keeter, 2008, *Aspects of nonresponse bias in RDD telephone surveys*, w: *Advances in Telephone Survey Methodology*, (red.) J.M. Lepkowski, C. Tucker, J.M. Brick, E.D. De Leeuw, L. Japec, P.J. Lavrakas, M.W. Link, R.L. Sangster, John Wiley & Sons, Inc., New York, s. 561–586
Mooney, Christopher Z., 2008, *Bootsrapping*, w: *Encyclopedia of Survey Research Methods*, (red.) P.J. Lavrakas, SAGE Publications, Inc., Thousand Oaks, CA, s. 65–67
Mulekar, Madhuri S., John C. Knutson, Jyoti A. Champanerkar, 2008, *How useful are approximations to mean and variance of the index of dissimilarity?*, „Computational Statistics & Data Analysis", nr 52(4), s. 2098–2109
Mulry, Mary H., 2007, *Summary of accuracy and coverage evaluation for the U.S. Census 2000*, „Journal of Official Statistics", nr 23(3), s. 345–370
Mulry, Mary H., 2008, *Coverage error*, w: *Encyclopedia of Survey Research Methods*, (red.) P.J. Lavrakas, SAGE Publications, Inc., Thousand Oaks, CA, s. 161–166
Neyman, Jerzy, 1933, *Zarys teorji i praktyki badania struktury ludności metodą reprezentacyjną*, Wydawnictwo Instytutu Spraw Społecznych, Warszawa
Neyman, Jerzy, 1934, *On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection*, „Journal of the Royal Statistical Society", nr 97(4), s. 558–625
Nowak, Lucyna, Dorota Szałtys, Zofia Kostrzewska, Aleksandra Pazderska, Alina Sobieszak, Joanna Stańczak, Ewa Jabłońska, 2007, *Poprawa jakości i dostępności statystyki migracji*

zagranicznych, Raport finalny w ramach projektu nr 12 Transition Facility 2004––19100.2005.001–2005.536, Główny Urząd Statystyczny, Warszawa
Nowak, Stefan, 2007, *Metodologia badań społecznych*, Wydawnictwo Naukowe PWN, Warszawa
O'Rourke, Diane, Johnny Blair, 1983, *Improving random respondent selection in telephone surveys*, „Journal of Marketing Research", nr 20, s. 428–432
Oh, Lock H., Frederick J. Scheuren, 1983, *Weighting adjustments for unit nonresponse*, w: *Incomplete Data in Sample Surveys. Vol. 2. Theory and Bibliographies*, (red.) W.G. Madow, H. Nisselson, I. Olkin, D. Rubin, Academic Press, New York, s. 143–187
Oldendick, Robert W., George F. Bishop, Susan B. Sorenson, Alfres J. Tuchfarber, 1988, *A comparison of the Kish and Last Birthday Methods of respondent selection in telephone surveys*, „Journal of Official Statistics", nr 4(4), s. 307–318
Olson, Kirsten, 2006, *Survey participation, nonresponse bias, measurement error bias, and total bias*, „Public Opinion Quarterly", nr 70(5), s. 737–758
Olson, Kristen, Ipek Bilgen, 2011, *The role of interviewer experience on acquiescence*, „Public Opinion Quarterly", nr 75(1), s. 99–114
Olson, Kirsten, James M. Lepkowski, David H. Garabrant, 2011a, *An experimental examination of the content of persuasion letters on nonresponse rates and survey estimates in a nonresponse follow-up study*, „Survey Research Methods", nr 5(1), s. 21–26
Olsson, Ulf H., Tron Foss, Einar Brevik, 2004, *Two equivalent discrepancy functions for maximum likelihood estimation: do their test statistics follow a non-central chi-square distribution under model misspecification?*, „Sociological Methods & Research", nr 32(4), s. 453–500
Opsomer, Jean, 2011, *Innovations in survey sampling design: discussion of three contributions presented at the U.S. Census Bureau,* „Survey Methodology", nr 37(2), s. 227–231
Panek, Tomasz, Janusz Czapiński, Irena E. Kotowska, 2011, *Metodologia badań*, w: *Diagnoza Społeczna 2011 Warunki i Jakość Życia Polaków – Raport,* (red.) J. Czapiński, T. Panek, „Contemporary Economics", nr 5(3), s. 35–44
Park, Inho, Hyunshik Lee, 2004, *Design effects for the weighted mean and total estimators under complex survey sampling*, „Survey Methodology", nr 30(2), s. 183–193
Paul, Sudhir R, Krishna K. Saha, Uditha Balasooriya, 2003, *An empirical investigation of different operating characteristics of several estimators of the intraclass correlation in the analysis of binary data*, „Journal of Statistical Computation & Symulation", nr 73(7), s. 507–523
Pawłowski, Tadeusz, 1969, *Metodologiczne zagadnienia humanistyki*, Wydawnictwo Naukowe PWN, Warszawa
Peytchev, Andy, 2009, *Survey breakoff*, „Public Opinion Quarterly", nr 73(1), s. 74–97
Peytchev, Andy, 2011, *Breakoff and unit nonresponse across web surveys*, „Journal of Official Statistics", nr 27(1), s. 33–47
Peytchev, Andy, Rodney K. Baxter, Lisa R. Carley-Baxter, 2009, *Not all survey effort is equal. reuction of nonresponse bias and nonresponse error*, „Public Opinion Quarterly", nr 73(4), s. 785–806
Peytchev, Andy, Mick P. Couper, Sean Esteban McCabe, Scott Crawford, 2006, *Web survey design: paging vs. scrolling*, „Public Opinion Quarterly", nr 70(4), s. 596–607
Pickery, Jan, Ann Carton, 2008, *Oversampling in relation to differential regional response rates*, „Survey Research Methods", nr 2(2), s. 83–92
Pike, Gary R., 2008, *Using weighting adjustments to compensate for survey nonresponse*, „Research in Higher Education", nr 49(2), s. 153–171

Platek Richard, Carl-Erik Särndal, 2001, *Can a statistician deliliver?*, „Journal of Official Statistics", nr 17(1), s. 1–20
Pokropek, Artur, 2011, *Missing by design: planned missing-data designs in social science*, „ASK. Research&Methods" nr 20(1), s. 81–105
Politz, Alfred, Willard Simmons, 1949, *An attempt to get the „not at homes" into the sample without callbacks*, „Journal of the American Statistical Association", nr 44(245), s. 9–16
Potter, Frank, 2008, *Multiplicity sampling*, w: *Encyclopedia of Survey Research Methods*, (red.) P.J. Lavrakas, SAGE Publications, Inc., Thousand Oaks, CA, s. 491–492
Rässler, Suzanne, Donald B. Rubin, Hathaniel Schenker, 2008, *Assessing the impact of initial nonresponse and attrition in the analysis of unemployment duration with panel surveys*, „Advances in Statistical Analysis", nr 92(3), s. 297–318
Renaud, Anne, 2007, *Estimation of the coverage of the 2000 census of population in Switzerland: methods and results*, „Survey Methodology", nr 33(2), s. 199–210
Rizzo, Louis, Michael J. Brick, Inho Park, 2004, *A minimally intrusive method for sampling persons in random digit dial surveys*, „Public Opinion Quarterly", nr 68(2), s. 267–274
Rizzo, Louis, Graham Kalton, Michael J. Brick, 1996, *A comparison of some weighting adjustments methods for panel nonresponse*, „Survey Methodology", nr 22(1), s. 43–53
Robins, Lee N., 1963, *The reluctant respondent*, „Public Opinion Quarterly", nr 27(2), s. 276–286
Rosenbaum, Paul R., 1987, *Model-based direct adjustment*, „Journal of the American Statistical Association", nr 82(398), s. 387–394
Rosenbaum, Paul R., Donald B. Rubin, *Reducing bias in observational studies using subclassification on the propensity score*, „Journal of the American Statistical Association", nr 79(387), s. 516–524
Rousseau, Michael, Marielle Simon, Richard Bertrand, Krystal Hachey, 2012, *Reporting missing data: a study of selected articles published from 2003–2007*, „Quality & Quantity", nr 46, s. 1393–1406
Rubin, Donald B., 1976, *Inference and missing data*, „Biometrika", nr 63(3), s. 581–592
Sakshaug, Joseph W., Frauke Kreuter, 2012, *Assessing the magnitude of non-consent biases in linked survey and administrative data*, „Survey Research Methods", nr 6(2), s. 113–122
Sakshaug, Joseph W., Ting Yan, Roger Tourangeau, 2010, *Nonresponse error, measurement error, and mode of data collection. tradeoffs in a multi-mode survey of sensitive and non-sensitive items*, „Public Opinion Quarterly", nr 74(5), s. 907–993
Salmon, Charles T., John S. Nichols, 1983, *The Next-birthday method of respondent selection*, „Public Opinion Quarterly", nr 47(2), s. 270–276
Särndal, Carl-Erik, Sixten Lundström, 2008, *Assessing auxilary vectors for control of non-response bias in the calibration estimators*, „Journal of Official Statistics", nr 24(2), s. 167–191
Särndal, Carl-Erik, Sixten Lundström, 2005, *Estimation in surveys with nonresponse*, John Wiley & Sons, Inc., New York
Särndal, Carl-Erik, Bengt Swensson, 1987, *A general view of estimation for two-phases of selection with applications to two-phase sampling and non-response*, „International Statistical Review", nr 55(3), s. 279–294.
Särndal, Carl-Erik, Bengt Swensson, Jan Wretman, 1992, *Model assisted survey sampling*, Springer-Verlag, New York
Sawińska, Monika, Zbigniew Sawiński, 2009, *Europejski Sondaż Społeczny 2008. Edycja 4. Schemat doboru próby wraz z oceną realizacji*, ORBS IFiS PAN, Warszawa
Sawiński, Zbigniew, 2005, *Metody doboru respondentów*, w: *Fieldwork jest sztuką*, (red.) P.B. Sztabiński, Z. Sawiński, F. Sztabiński, Wydawnictwo IFiS PAN, Warszawa, s. 79–118

Sawiński, Zbigniew, 2005a, *Wywiady osobiste z komputerem przenośnym*, w: *Fieldwork jest sztuką*, (red.) P.B. Sztabiński, Z. Sawiński, F. Sztabiński Wydawnictwo IFiS PAN, Warszawa, s. 225–236

Sawiński, Zbigniew, 2005b, *Program kontroli jakości pracy ankieterów*, w: *Fieldwork jest sztuką*, (red.) P.B. Sztabiński, Z. Sawiński, F. Sztabiński Wydawnictwo IFiS PAN, Warszawa, s. 401–410

Sawiński, Zbigniew, 2010, *Zastosowania tablic w badaniach zjawisk społecznych*, w: *Studia z socjologii ilościowej. Dane, pomiar, wskaźniki, analizy 2*, Wydawnictwo IFiS PAN, Warszawa

Sawiński, Zbigniew, 2011, *Intra-cluster homogeneity in survey samples: a neglected tool*, referat wygłoszony w ramach 4th Conference of the European Survey Research Association (ESRA). Lozanna, Szwajcaria, 18–22 lipca 2011 roku

Sawiński, Zbigniew, Franciszek Sztabiński, 2005, *Czy ankieterzy oszukują? Jak można to sprawdzić?*, w: *Fieldwork jest sztuką*, (red.) P.B. Sztabiński, Z. Sawiński, F. Sztabiński, Wydawnictwo IFiS PAN, Warszawa, s. 361–380

Schouten, Barry, 2007, *A selection strategy for weighting variables under a not-missing-at-random assumption*, „Journal of Official Statistics", nr 23(1), s. 51–68

Schouten, Barry, Jelke Bethlehem, Koen Beullens, Øyvin Kleven, Geert Loosveldt, Annemieke Luiten, Katja Rutar, Natalie Shlomo, Chris Skinner, 2012, *Evaluating, comparing, monitoring, and improving representativeness of survey response throught R-indicators and partial R-indicators*, „International Statistical Review", nr 80(3), s. 382–399

Schouten, Barry, Fannie Cobben, 2007, *R-indexes for the comparison of different fieldwork strategies and data collection modes*, Discussion paper 07002, Statistics Netherlands, Voorburg/Heerlen

Schouten, Barry, Fannie Cobben, Jelke Bethlehem, 2009, *Indicators for the representativeness of survey response*, „Survey Methodology", nr 35(1), s. 101–113

Schouten, Barry, Natalie Shlomo, Chris Skinner, 2011, *Indicators for monitoring and improving representativeness of response*, „Journal of Official Statistics", nr 27(2), s. 1–24

Shlomo, Natalie, Chris Skinner, Barry Schouten, 2012, *Estimation of an indicator of the representativeness of survey response*, „Journal of Statistical Planning and Inference", nr 142, s. 201–211

Singelton, Royce, Bruce Straits, 2005, *Approaches to Social Research*, 4th ed., Oxford University Press, New York

Singer, Elenaor 2006, *Nonresponse bias in household surveys*, „Public Opinion Quarterly", nr 70(5), s. 637–645

Sirken, Monroe G., 1970, *Household surveys wit multiplicity*, „Journal of the American Statistical Association", nr 65(329), s. 257–266

Sirken, Monroe G., 1972, *Stratified sample surveys with multiplicity*, „Journal of the American Statistical Association", nr 67(337), s. 224–227

Sirken, Monroe G., 2002, *Design effects of sampling frames in establishments surveys*, „Survey Methodology", nr 28(2), s. 183–190

Sirken, Monroe G., Barry J. Graubard, Richard W. LeValley, 1978, *Evaluation of census population coverage by network surveys*, „Proceedings of the Survey Research Methods Section", American Statistical Association, s. 239–244.

Słomczyński, Kazimierz M., 1999, *Międzykrajowe badania porównawcze,* w: *Encyklopedia socjologii*, (red.) H. Domański, A. Kojder, K. Koseła, K. Kowalewicz, H. Kubiak, W. Kwaśniewicz, J. Mucha, M. Ofierska, A. Piotrowski, J. Szacki, M. Ziółkowski, Oficyna Naukowa, Warszawa, s. 238–245

Słomczyński, Kazimierz M., 2004, *Europejski Sondaż Społeczny a inne międzynarodowe badania surveyowe*, „ASK. Społeczeństwo. Badania. Metody" nr 13, s. 9–25
Słomczyński, Kazimierz M., Krystyna M. Janicka, Włodzimierz Wesołowski, 1994, *Badania struktury społecznej Łodzi: doświadczenia i perspektywy*, Wydawnictwo IFiS PAN, Warszawa
Smith, Tom W., 1983, *The hidden 25 percent: an analysis of nonresponse on the 1980 general social survey*, „Public Opinion Quarterly" nr 47(3), s. 386–404
Smith, Tom W., 2002, *Developing nonresponse standards*, s. 27–40, w: *Survey Nonresponse*, (red.) R. Groves, D.A. Dillman, J.L. Eltinge R.J.A. Little, John Wiley & Sons, Inc., New York
Smith, Tom W., 2007, *Survey non-response procedures in cross-national perspective: the 2005 issp non-response survey*, „Survey Research Methods" nr 1(1), s. 45–54
Smith, Tom W., 2009, *A Review of Methods to Estimate the Status of Cases With Unknown Eligibility.* AAPOR, https://www.aapor.org/AAPORKentico/AAPOR_Main/media/MainSiteFiles/FindingE.pdf (data dostępu: 12 grudnia 2014 roku)
Smith, Tom W., 2011, *The report of the international workshop on using multi-level data from sample frames, auxilary databases, paradata and related souces to detect and adjust for nonresponse bias in surveys*, „International Journal of Public Opinion Research", nr 23(3), s. 389–402
Smith, Wayne, Tien Chey, Bin Jalaludin, Glenn Salkeld, T. Capon, 1995, *Increasing response rates in telephone surveys: a randomized trial*, „Journal of Public Health", nr 17(1), s. 33–38
Spiess, Martin, Jan Goebel, 2005, *In the effect of item nonresponse on the estimation of a two-panel-waves wage equation*, „Allgemeines Statistisches Archiv", nr 89, s. 63–74
Stec, Jeffrey A., Paul J. Lavrakas, Elizabeth A. Stasny, 1999, *Investigating unit non-response in an RDD survey*, „Proceedings of the Survey Research Methods Section", American Statistical Assocation, s. 919–924
Stoop, Ineke, 2004, *Surveying nonrespondents*, „Field Method" nr 16(1), s. 23–54
Stoop, Ineke, 2005, *The hunt for the last respondent. Nonresponse in sample surveys*, Social and Cultural Planning Office, Haga
Stoop, Ineke, Jaak Billiet, Achim Koch, Rory Fitzegarald, 2010, *Improving survey response. Lessons learned from the European Social Survey*, John Wiley & Sons, Ltd, New York
Sułek, Antoni, 2002, *Ogród metodologii socjologicznej*, Wydawnictwo Naukowe Scholar, Warszawa
Sztabiński, Franciszek, 2011, *Ocena jakości danych w badaniach surveyowych*, Wydawnictwo IFiS PAN, Warszawa
Sztabiński, Paweł B., 1995, *Ankieter jako źródło zniekształceń w procesie badawczym*, „ASK. Społeczeństwo. Badania. Metody", nr 2, s. 159–168
Sztabiński, Paweł B., 1997, *Ankieterzy i ich respondenci*, Wydawnictwo IFiS PAN, Warszawa
Sztabiński, Paweł B., 2001, *Wywiad telefoniczny ze wspomaganiem komputerowym (CATI): czy rzeczywiście idealna technika*?, „ASK. Społeczeństwo. Badania. Metody", nr 10, s. 65–90
Sztabiński, Paweł B., 2004, *Metodologia badania Europejski Sondaż Społeczny,* „ASK. Społeczeństwo. Badania. Metody", nr 16, s. 27–37
Sztabiński, Paweł B., 2005, *Dlaczego należy rygorystycznie przestrzegać zasad pracy ankieterskiej? Wywiad kwestionariuszowy jako technika standaryzowana*, w: *Fieldwork jest sztuką*, (red.) P.B. Sztabiński, Z. Sawiński, F. Sztabiński, Wydawnictwo IFiS PAN, Warszawa, s. 49–54
Sztabiński, Paweł B., 2006, *Dlaczego respondenci uczestniczą lub nie uczestniczą w badaniach? Porównanie 1994–2004*, „ASK. Społeczeństwo. Badania. Metody", nr 15(1), s. 7–28

Sztabiński, Paweł B., 2011, *How to prepare an advance letter? The ESS Experience in Poland*, „ASK Research & Methods", nr 20(1), s. 107–148

Sztabiński, Paweł B., Anna Dyjas-Pokorska, Teresa Żmijewska-Jędrzejczyk, 2008, *Understanding refusals*, „ASK. Społeczeństwo. Badania. Metody", nr 17(1), s. 39–84

Sztabiński, Paweł B., Franciszek Sztabiński, Dariusz Przybysz, 2007, *Are non-respondents similar to respondents? Findings from the ESS-2004 in Poland*, „ASK. Społeczeństwo. Badania. Metody", nr 18(1), s. 25–54

Sztabiński, Paweł B., Franciszek Sztabiński, Dariusz Przybysz, 2009, *How does lenght of fieldwork period influence non-response? Findings from ESS2 in Poland*, „ASK Research & Methods", nr 18(1), s. 67–95

Sztabiński, Paweł B., Franciszek Sztabiński, Dariusz Przybysz, 2012, *What do respondents and non-respondents think of incentives and how do they react to them? The ESS experience in Poland*, „ASK Research & Methods", nr 21(1), s. 87–122

The American Association for Public Opinion Research, 2011, *Standard Definitions: Final Disposition of Case Codes and Outcome Rates for Surveys. 7th edition*, AAPOR, https://www.esomar.org/uploads/public/knowledge-and-standards/codes-and-guidelines/ESOMAR_Standard-Definitions-Final-Dispositions-of-Case-Codes-and-Outcome-Rates-for-Surveys.pdf (data pobrania: 7 lutego 2015 roku)

Tortora, Robert D., Robert M. Groves, Emilia Peytcheva, 2008, *Multiplicity based sampling for the mobile telephone population: Coverage, nonresponse, and measurement issues*, w: *Advances in Telephone Survey Methodology*, (red.) J.M. Lepkowski, , C. Tucker, M.J. Brick, E.D. de Leeuw, L. Japec, P.J. Lavrakas, M.W. Link, R.L. Sangster, John Wiley & Sons, Inc., New York, s. 133–148

Tourangeau, Roger, Gary Shapiro, Anne Kearney, Lawrence Ernst, 1997, *Who lives here? Survey undercoverage and household roster quaestions*, „Journal of Official Statistics", nr 13(1), s. 1–18

Troldahl, Verling C., Roy E. Carter Jr., 1964, *Random selection of respondents within household in phone surveys*, „Journal of Marketing Research", nr 1, s. 71–76

Troxel, Andrea B., Stuart R. Lipsitz, Troyen A. Brennan, 1997, *Weighted estimationg equations with nonignorably missing response data*, „Biometrics", nr 53(3), s. 857–869

Trzciński, Rafał, 2009, *Wykorzystanie techniki propensity score matching w badaniach ewaluacyjnych*, Polska Agencja Rozwoju Przedsiębiorczości, Warszawa

Ukoumunne, Obioha C., 2002, *A Comparison of confidence interval methods for the intraclass correlation coefficient in cluster randomized trials*, „Statistics in Medicine", nr 21(24), s. 3757–3774

Valentine, Charles A., Betty L. Valentine, 1971, *Missing Men: A Comparative Methodological Study of Underenumeration and Related Problems*, Report prepared under contract for the U.S. Census Bureau, Washington D.C.

Vehovar, Vasja, 2007, *Non-response bias in the european social survey*, w: *Measuring meaningful data in social research*, (red.) G. Loosveldt, M. Swyngedouw, B. Cambr´e, Acco, Leuven, s. 335–356

Vehovar, Vasja, Zenel Bategelj, Katja Lozar Manfreda, Metka Zaletel, 2002, *Nonresponse in Web surveys*, w: *Survey Nonresponse*, (red.) R. Groves, D.A. Dillman, J.L. Eltinge R.J.A. Little. John Wiley & Sons, Inc., New York, s. 229–242

Vicente, Paula, Elizabeth Reis, 2009, *The mobile-only population in Portugal and its impact in a dual frame telephone survey*, „Survey Research Methods", nr 2(3), s. 105–111

von der Lippe, Elena, Patrick Schmich and Cornelia Lange, Robert Koch, 2011, *Advance letters as a way of reducing non-response in a National Health Telephone Survey: Differences between listed and unlisted numbers*, „Survey Research Methods", nr 5(3), s. 103–116

Voogt, Robert J.J., Hetty van Kempen, 2002, *Nonresponse bias and stimulus effects in the dutch national election study*, „Quality & Quantity", nr 36, s. 325–345

Weisberg, Herbert F., 2005, *The Total Survey Error Approach. A Guide to the New Science of Survey Research,* The University of Chicago Press, Chicago and London

Wesołowski Włodzimierz, Kazimierz M. Słomczyński, Krystyna Janicka, 1965, *Struktura społeczna mieszkańców Łodzi 1965*, komputerowy zbiór danych, Zespół Porównawczych Analiz Nierówności Społecznych, IFiS PAN, Warszawa

West, Brady T., 2011, *Paradata in survey research*, „Survey Practise", nr 4(4), s. 1–8

West, Brady T., Kristen Olson, 2010, *How much of interviewer variance is really nonresponse error variance?*, „Public Opinion Quarterly", nr 74(5), s. 1004–1026

Wiseman, Frederick, Philip McDonald, 1979, *Noncontact and refusal rates in consumer telephone surveys*, „Survey Methodology", nr. 16(4), 478–484

Wood, Angela M., Ian R. White, 2004, *Using number of failed contact attempts to adjust for non-ignorable non-response*, „Journal of the Royal Statistical Society", nr 169(3), s. 525–542

Yan, Ting, 2009, *A Meta-analysis of Within-Household Respondent Selection Methods*, AAPOR Annual Conference – May 14–17, 2009, s. 6134–6147, https://www.amstat.org/sections/srms/proceedings/y2009/Files/400064.pdf (data pobrania: 3 stycznia 2015 roku)

Zinniel, Sonja, 2008, *Within-unit coverage error*, w: *Encyclopedia of Survey Research Methods*, (red.) P.J. Lavrakas, SAGE Publications, Inc., Thousand Oaks, CA, s. 962–963

Aneks

Tabela A1. Charakterystyka populacji docelowych oraz operatów doboru prób badawczych w badaniach ESS1-2002 oraz ESS2-2004

Nazwa kraju	Populacja docelowa	Operat losowania
Typ I. Rejestry indywidualne – losowanie prób imiennych		
Belgia ESS1, ESS2	Osoby w wieku 15+ mieszkające w prywatnych gospodarstwach domowych. • *Kategorie wykluczone z populacji*: bezdomni, przebywający w instytucjach (szpitale, koszary wojskowe, więzienia i areszty śledcze).	Losowanie imienne z rejestru ludności. • *Wskaźnik pokrycia populacji:* wskaźnik pokrycia oraz sposób uaktualniania oceniany jest jako „idealny".
Dania ESS1, ESS2	Osoby w wieku 15+ będące obywatelami Danii oraz wszystkie osoby, które planują pobyt w Danii przez przynajmniej 3 miesiące. • *Kategorie wykluczone z populacji*: osoby bezdomne.	Losowanie imienne z oficjalnego rejestru ludności. • *Wskaźnik pokrycia populacji:* rejestr ludności pokrywa około 99,9% osób zamieszkujących w Danii.
Estonia ESS2	Osoby w wieku 15+ mieszkające w prywatnych gospodarstwach domowych. • *Kategorie wykluczone z populacji*: osoby przebywające w instytucjach: szpitale, koszary wojskowe, więzienia i areszty śledcze, domy dziecka, domy opieki społecznej, internaty szkolne, akademiki, klasztory.	Losowanie z oficjalnego rejestru zameldowanych na pobyt stały mieszkańców Estonii. Rejestr uaktualniany na bieżąco. • *Wskaźnik pokrycia populacji:* b.d.
Finlandia ESS1, ESS2	Obywatele Finlandii w wieku 15+ oraz cudzoziemcy posiadający status rezydenta. • *Kategorie wykluczone z populacji*: więźniowie, osoby przebywające w szpitalach, marynarze, osoby przebywające poza miejscem zamieszkania (za granicą) z uwagi na studiowanie lub pracę, osoby z nieprawidłowymi danymi adresowymi. Szacunkowa wielkość osób wykluczonych z populacji wynosi mniej więcej 1,8%.	Losowanie z rejestru imiennego • *Wskaźnik pokrycia populacji:* b.d.
Hiszpania ESS2	Osoby w wieku 15+ zameldowane na pobyt stały i mieszkające w prywatnych gospodarstwach domowych łącznie z mieszkańcami obszaru Cetua oraz miastem Melilla. • *Kategorie wykluczone z populacji*: b.d.	Operat losowania obwodów wyborczych na podstawie pełnego rejestru przygotowanego przez urząd statystyczny. Losowanie jednostek w ramach wybranych obwodów wyborczych ze spisów indywidualnych uaktualnianych przez samorządy lokalne. • *Wskaźnik pokrycia populacji:* b.d.

Nazwa kraju	Populacja docelowa	Operat losowania
Islandia ESS2	Osoby w wieku 15+ mieszkające w prywatnych gospodarstwach domowych. • *Kategorie wykluczone z populacji*: obcokrajowcy są włączeni jeżeli posiadają status rezydentów.	Losowanie imienne z oficjalnego rejestru ludności. • *Wskaźnik pokrycia populacji:* 100%.
Niemcy ESS1, ESS2	Osoby w wieku 15+ będące obywatelami Niemiec. Populacja podzielona na dwie niezależne części: a) obywatele dawnego NRD włącznie z mieszkańcami Berlina Wschodniego, b) obywatele dawnej RFN z mieszkańcami Berlina Zachodniego. • *Kategorie wykluczone z populacji*: b.d.	Rejestr jednostkowy zawierający dane o zameldowaniu czasowym. Z uwagi na regulacje prawne, każda osoba zmieniająca miejsce pobytu na okres dłuższy niż jeden tydzień, jest zobowiązana do rejestracji meldunkowej. • *Wskaźnik pokrycia populacji:* 100%
Norwegia ESS1, ESS2	Osoby w wieku 15+ mieszkające w Norwegii włącznie z zarejestrowanymi obcokrajowcami. • *Kategorie wykluczone z populacji*: studenci spoza Norwegii.	Uaktualniany miesięcznie oficjalny rejestr ludności. • *Wskaźnik pokrycia populacji:* b.d.
Polska ESS1, ESS2	Osoby w wieku 15+ zameldowane na pobyt stały na terytorium Rzeczpospolitej Polskiej. • *Kategorie wykluczone z populacji*: osoby przebywające w instytucjach, obcokrajowcy pracujący „na czarno", osoby bezdomne, osoby czasowo niedostępne.	Rejestr PESEL – imienny rejestr ludności uwzględniający urodzenia, zgony, migracje. • *Wskaźnik pokrycia populacji:* b.d.
Słowacja ESS2	Osoby w wieku 15+ mieszkające w prywatnych gospodarstwach domowych. • *Kategorie wykluczone z populacji*: osoby bezdomne, osoby przebywające w instytucjach.	Imienny rejestr ludności aktualizowany na bieżąco. • *Wskaźnik pokrycia populacji:* około 99,9%.
Słowenia ESS1, ESS2	Osoby w wieku 15+ mieszkające w prywatnych gospodarstwach domowych. • *Kategorie wykluczone z populacji*: osoby przebywające w instytucjach nie są wyłączone, ale zakłada się niski stopień realizacji wywiadów.	Imienny rejestr ludności zameldowanej na pobyt stały (obywatele oraz obcokrajowcy). • *Wskaźnik pokrycia populacji:* co najmniej 99%.
Szwecja ESS1, ESS2	Osoby w wieku 15+ mieszkające na terytorium Szwecji. • *Kategorie wykluczone z populacji*: b.d.	Urzędowy, imienny rejestr ludności, uaktualniany na bieżąco. • *Wskaźnik pokrycia populacji:* 100%.
Węgry ESS1, ESS2	Osoby w wieku 15+ zameldowane na pobyt stały na terytorium Węgier. • *Kategorie wykluczone z populacji*: Nie wyklucza się z góry żadnych kategorii, ale praktycznie nie prowadzi się wywiadów z osobami bezdomnymi, osobami przebywającymi poza granicą.	Ogólnokrajowy rejestr ludności uaktualniany na bieżąco (urodzenia, zgony, migracje). • *Wskaźnik pokrycia populacji:* 100%.
Typ II. Rejestry imienne – losowanie prób gospodarstw domowych lub punktów adresowych		
Irlandia ESS1, ESS2	Osoby w wieku 15+ mieszkające w prywatnych gospodarstwach domowych. • *Kategorie wykluczone z populacji*: b.d.	Cyfrowy rejestr wyborców 18+ aktualizowany corocznie, wykorzystywany do losowania punktów adresowych. • *Wskaźnik pokrycia populacji:* b.d.

Nazwa kraju	Populacja docelowa	Operat losowania
Włochy ESS1, ESS2	Osoby w wieku 15+ mieszkające obecnie w prywatnych gospodarstwach domowych na terenie Włoch. • *Kategorie wykluczone z populacji*: b.d.	Rejestr wyborców 18+ wykorzystywany do losowania punktów adresowych. • *Wskaźnik pokrycia populacji:* b.d.
Typ III. Rejestry abonentów telefonicznych oraz inne rejestry – losowanie prób gospodarstw domowych		
Austria ESS1, ESS2	Osoby w wieku 15+ mieszkające w prywatnych gospodarstwach domowych. • *Kategorie wykluczone z populacji*: bezdomni, przebywający w instytucjach (szpitale, koszary wojskowe, więzienia i areszty śledcze). Osoby nie posługujące się językiem niemieckim.	Losowanie gospodarstw domowych w oparciu o spis abonentów telefonicznych uaktualniany cztery razy w roku. • *Wskaźnik pokrycia populacji*: spis abonentów telefonicznych pokrywa około 90% gospodarstw domowych. Brak możliwości wylosowania gospodarstw bez dostępu do telefonii stacjonarnej oraz gospodarstw z zastrzeżonymi numerami – niepełne pokrycie redukowane poprzez metodę ustalonej ścieżki w celu wyszukania gospodarstw nieuwzględnionych w rejestrach.
Czechy ESS1	Osoby w wieku 15+ mieszkające w gospodarstwach domowych. • *Kategorie wykluczone z populacji*: b.d.	Losowanie gospodarstw domowych z bazy „SIPO" rozporządzanej przez prywatną firmę. Baza integruje repozytoria gospodarstw domowych korzystających z usług przedsiębiorstw dostarczających energię elektryczną, gaz, telefonię stacjonarną oraz sygnał telewizyjny. Baza danych uaktualniana miesięcznie. • *Wskaźnik pokrycia populacji:* baza „SIPO" zawiera informacje o mniej więcej 98% gospodarstwach domowych.
Hiszpania ESS1	Osoby w wieku 15+ zameldowane na pobyt stały i mieszkające w prywatnych gospodarstwach domowych. • *Kategorie wykluczone z populacji*: mieszkańcy miast Ceuta oraz Melilla.	Operatem jest tzw. *master sample* przygotowana przez hiszpański urząd statystyczny, zawierająca rejestr obwodów wyborczych (dane adresowe o 65 000 gospodarstwach domowych). • *Wskaźnik pokrycia populacji:* około 99,7% populacji.
Izrael ESS1	Osoby w wieku 15+ mieszkające w prywatnych gospodarstwach domowych włącznie z ludnością żydowską zamieszkującą Zachodni Brzeg Jordanu oraz Strefę Gazy. • *Kategorie wykluczone z populacji*: ludność palestyńska zamieszkująca Wschodnią Jerozolimę, Beduini, ludność arabska z małych społeczności wiejskich.	Losowanie z rejestru abonentów telefonicznych. • *Wskaźnik pokrycia populacji:* 95% wszystkich gospodarstw domowych.
Luksemburg ESS1, ESS2	Osoby w wieku 15+ mieszkające w prywatnych gospodarstwach domowych. • *Kategorie wykluczone z populacji*: b.d.	Rejestr systemu zabezpieczeń społecznych zawierający informacje jednostkowe lub grupujący osoby traktowane zbiorczo dla celów podatkowych. • *Wskaźnik pokrycia populacji:* około 91%–94% populacji.

Nazwa kraju	Populacja docelowa	Operat losowania
Szwajcaria ESS1, ESS2	Osoby w wieku 15+ mieszkające w prywatnych gospodarstwach domowych. • *Kategorie wykluczone z populacji*: osoby bezdomne, osoby przebywające w instytucjach.	Rejestr abonentów telefonicznych, którego depozytariuszem jest szwajcarski urząd statystyczny. Rejestr numerów telefonicznych zawiera dane gospodarstw domowych włącznie z gospodarstwami z numerami zastrzeżonymi. Użytkownicy telefonów komórkowych nieposiadający telefonii stacjonarnej są ujęci w rejestrze pod warunkiem, że mają podpisaną stałą umowę z operatorem. • *Wskaźnik pokrycia populacji:* co najmniej 95% gospodarstw domowych.
Typ IV. Rejestry kodów pocztowych lub rejestry budynków – losowanie prób adresowych		
Czechy ESS2	Osoby w wieku 15+ mieszkające w gospodarstwach domowych. • *Kategorie wykluczone z populacji:* b.d.	Losowanie punktów adresowych z bazy „UIR-ADR" – zawierającej adresy wszystkich budynków z przyporządkowanymi numerami. Rejestry opracowane przez samorządy lokalne oraz administrację centralną z darmowym dostępem przez Internet. • *Wskaźnik pokrycia populacji:* operat pozwala zidentyfikować budynki w obrębie ulic miast lub wsi, z wyłączeniem małych wsi, w których nie ma administracyjnego przyporządkowania ulic.
Holandia ESS1, ESS2	Osoby w wieku 15+ mieszkające w prywatnych gospodarstwach domowych. • *Kategorie wykluczone z populacji*: osoby hospitalizowane (1,3% populacji), marynarze, kierowcy.	Rejestr kodów pocztowych. • *Wskaźnik pokrycia populacji:* b.d.
Wielka Brytania ESS1, ESS2	Osoby w wieku 15+ mieszkające na stałe na terytorium Anglii, Walii, Szkocji oraz Irlandii Północnej. • *Kategorie wykluczone z populacji*: b.d.	Uaktualniany na bieżąco rejestr kodów pocztowych brytyjskiej poczty królewskiej, zawierający dane o wszystkich powiązanych z kodami pocztowymi adresach. Operat nie zawiera informacji o niewielkiej liczbie obszarów Irlandii Północnej. • *Wskaźnik pokrycia populacji:* około 96% wszystkich gospodarstw domowych oraz 97% mieszkańców.
Typ V. Brak rejestrów indywidualnych, gospodarstw domowych lub adresowych – losowanie adresowe w oparciu o mapy administracyjne		
Francja ESS1, ESS2	Osoby w wieku 15+ mieszkające w prywatnych gospodarstwach domowych. • *Kategorie wykluczone z populacji*: b.d.	Operat gmin. • *Wskaźnik pokrycia populacji:* b.d.
Grecja ESS1, ESS2	Osoby w wieku 15+ mieszkające w prywatnych gospodarstwach domowych. • *Kategorie wykluczone z populacji*: mieszkańcy wysp archipelagu Cyklad oraz Dodekanez (z wyłączeniem Rodos). Osoby bezdomne oraz przebywające w instytucjach.	Operat obwodów spisowych • *Wskaźnik pokrycia populacji:* b.d.

Nazwa kraju	Populacja docelowa	Operat losowania
Portugalia ESS1, ESS2	Osoby w wieku 15+, które zamieszkują w kontynentalnej części Portugalii. • *Kategorie wykluczone z populacji*: osoby przebywające w instytucjach, osoby mieszkające w miejscowościach o liczbie budynków zamieszkałych mniejszej niż 10 (około 3% populacji), osoby nieposługujące się językiem portugalskim, osoby niedołężne, osoby bezdomne, obywatele Portugalii przebywający poza granicami kraju, emerytowani oraz niepracujący obcokrajowcy.	Operat miejscowości. • *Wskaźnik pokrycia populacji:* b.d.
Ukraina ESS2	Osoby w wieku 15+ mieszkające w prywatnych gospodarstwach domowych. • *Kategorie wykluczone z populacji*: osoby bezdomne oraz przebywające w instytucjach.	Operat miejscowości. • *Wskaźnik pokrycia populacji:* b.d.

Źródło: Opracowanie własne na podstawie raportów *ESS1 Sampling report 2002* oraz *ESS2 Sampling report 2004*

Tabela A2. Opis zmiennych ESS5-PL (ed. 2010) wykorzystanych w analizach DEFF

Nr pytania	Charakterystyka zmiennych	Zakres wartości
	Moduł A – media, zaufanie do ludzi	
A1	Oglądanie telewizji, czas całkowity w przeciętny dzień tygodnia	0–7
A3	Słuchanie radio, czas całkowity w przeciętny dzień tygodnia	0–7
A5	Czytanie gazet, czas całkowity w przeciętny dzień tygodnia	0–7
A8	Większości ludzi można ufać, czy też ostrożności nigdy nie za wiele	0–10
A9	Większość ludzi stara się mnie wykorzystać, czy też starałaby się postępować uczciwie	0–10
A10	Ludzie starają się służyć pomocą innym, czy też dbają o własny interes	0–10
	Moduł B – polityka	
B1	Stopień zainteresowania polityką	1–4
B4	Zaufanie do polskiego parlamentu	0–10
B5	Zaufanie do systemu prawnego	0–10
B6	Zaufanie do policji	0–10
B7	Zaufanie do polityków	0–10
B8	Zaufanie do partii politycznych	0–10
B9	Zaufanie do Parlamentu Europejskiego	0–10
B10	Zaufanie do Organizacji Narodów Zjednoczonych	0–10
B23	Usytuowanie na skali lewica – prawica	0–10
B24	Zadowolenie z obecnego życia	0–10
B25	Zadowolenie z obecnego stanu polskiej gospodarki	0–10
B26	Zadowolenie z polskiego rządu	0–10
B27	Zadowolenie z funkcjonowania demokracji	0–10
B28	Opinia o stanie poziomu kształcenia, edukacji w Polsce	0–10
B29	Opinia o stanie służby zdrowia oraz dostępnych usług medycznych	0–10
B30	Rząd powinien redukować różnice w dochodach	1–5
B31	Geje i lesbijki powinny mieć swobodę układania życia wg własnych przekonań	1–5
B32	Powinno się zakazać działalności partiom dążącym do obalenia demokracji	1–5
B33	Można liczyć na współczesną naukę przy rozwiązywaniu problemów ekologicznych	1–5
B38	Imigracja korzystna, czy też niekorzystna dla polskiej gospodarki	0–10
B39	Imigracja szkodzi, czy też wzbogaca życie kulturalne w Polsce	0–10
B40	Imigranci sprawiają, że Polska staje się lepszym, czy też gorszym miejscem do życia	0–10
	Moduł C – poczucie dobrobytu, wykluczenie społeczne, religia	
C1	Poziom poczucia szczęścia	0–10
C4	Częstość udziału w spotkaniach, imprezach lub wydarzeniach towarzyskich	1–5
C6	Poczucie bezpieczeństwa po zmroku w okolicy miejsca zamieszkania	1–4

Nr pytania	Charakterystyka zmiennych	Zakres wartości
C15	Subiektywna ocena stanu zdrowia	1–5
C21	Poziom religijności	0–10
	Moduł D – zaufanie do wymiaru sprawiedliwości	
D1	Jak bardzo złe jest zawyżanie wartości szkód lub bezpodstawne odszkodowanie	1–4
D2	Jak bardzo złe jest kupowanie rzeczy pochodzącej z kradzieży	1–4
D3	Jak bardzo złe jest popełnianie wykroczeń drogowych	1–4
D4	Jak bardzo prawdopodobne jest ukaranie za zawyżanie wartości szkód lub bezpodstawne odszkodowanie	1–4
D5	Jak bardzo prawdopodobne jest ukaranie za kupowanie rzeczy pochodzącej z kradzieży	1–4
D6	Jak bardzo prawdopodobne jest ukaranie za popełnianie wykroczeń drogowych	1–4
D7	Policja wypełnia swoje zdania właściwie, czy też niewłaściwie	1–5
D12	Na ile skutecznie policja zapobiega przestępstwom z udziałem przemocy	0–10
D13	Na ile skutecznie policja łapie przestępców, którzy włamują się do domów	0–10
D14	Na ile szybko, lub wolno, policja dotarłaby na miejsce przestępstwa	0–10
D18	Obowiązkiem, czy też nie, jest uszanować decyzje podejmowane przez policję	0–10
D19	Obowiązkiem, czy też nie, jest robić to, co nakazuje policja, nawet gdy nie rozumie się powodów	0–10
D20	Obowiązkiem, czy też nie, jest robić to, co nakazuje policja, nawet gdy to osobiście nie odpowiada	0–10
D21	Policja ma takie samo jak ja poczucie tego, co jest dobre a co złe	1–5
D22	Policja stoi na straży wartości, które są ważne dla ludzi takich jak ja	1–5
D23	Ogólnie rzecz biorąc popieram sposoby działania policji	1–5
D24	Policja zbytnio ulega wpływowi partii politycznych i polityków	1–5
D25	Jak często policjanci w Polsce przyjmują łapówki	0–10
D26	Sądy wypełniają swoje zadania dobrze, czy też źle	1–5
D27	Jak często sądy popełniają błędy, w wyniku których winny unika kary	0–10
D28	Jak często w sądach zapadają sprawiedliwe, bezstronne wyroki oparte na dowodach	0–10
D31	Jak często sędziowie w Polsce przyjmują łapówki	1–5
D32	Sądy chronią bardziej ludzi bogatych i wpływowych, niż zwykłych ludzi	1–5
D33	Ludzie, którzy łamią prawo powinny dostawać dużo surowsze wyroki niż obecnie	1–5
D34	Każdy powinien uszanować prawomocne wyroki sądów	1–5
D35	Należy ściśle przestrzegać przepisów prawa	1–5
D36	Aby postąpić słusznie, czasami trzeba złamać prawo	1–5
D37	Sądy zbytnio ulegają wpływowi partii politycznych i polityków	1–5
D40	Jak bardzo prawdopodobne jest zadzwonienie na policję w sytuacji zaistnienia przestępstwa	1–4

Nr pytania	Charakterystyka zmiennych	Zakres wartości
D41	Czy chętnie wzięłoby się udział w identyfikacji sprawcy przestępstwa	1–4
D42	Czy chętnie zeznawałoby się jako świadek w sądzie przeciwko oskarżonemu	1–4
	Moduł F – charakterystyka społeczno-demograficzna	
F1	Liczba osób zamieszkujących na stałe w gospodarstwie domowym	$\in \mathbb{N}$
F41	Dochód łączny gospodarstwa domowego ze wszystkich źródeł	1–10
	Moduł G – praca, rodzina, dobrobyt	
G4	Kobieta powinna być gotowa ograniczyć pracę zawodową dla dobra rodziny	1–5
G5	Mężczyźni powinny mieć pierwszeństwo w uzyskaniu pracy, jeśli sytuacja na runku pracy jest trudna	1–5
G6	Państwo powinno robić więcej, aby nie dopuszczać do popadania ludzi w biedę	1–5
	Moduł H – skala ludzkich wartości	
H_a	Ważne jest wymyślanie nowych rzeczy, bycie kreatywnym	1–6
H_b	Ważne jest bycie bogatym, posiadanie dużo pieniędzy i kosztowności	1–6
H_c	Ważne jest, aby wszyscy ludzie na świecie traktowani byli równo	1–6
H_d	Ważne jest pokazywanie własnych zdolności, bycie podziwianym	1–6
H_e	Ważne jest życie w bezpiecznym otoczeniu	1–6
H_f	Ważne jest poszukiwanie nowych zajęć	1–6
H_g	Ważne jest postępowanie zgodnie z nakazami, zasadami i przepisami	1–6
H_h	Ważne jest wysłuchiwanie ludzi, zrozumienie ich	1–6
H_i	Ważne jest bycie skromnym i pokornym	1–6
H_j	Ważna jest dobra zabawa	1–6
H_k	Ważne jest, aby samemu podejmować decyzję we własnych sprawach	1–6
H_l	Ważne jest pomaganie ludziom, dbanie o ich dobro	1–6
H_m	Ważne jest odnoszenie znaczących sukcesów	1–6
H_n	Ważne jest, aby władza zapewniała ochronę przed wszelkimi zagrożeniami	1–6
H_o	Ważne jest poszukiwanie przygód i podejmowanie ryzyka	1–6
H_p	Ważne jest, aby zachowywać się poprawnie	1–6
H_q	Ważny jest szacunek ze strony innych	1–6
H_r	Ważne jest, aby być lojalnym wobec przyjaciół	1–6
H_s	Ważne jest, aby ludzie dbali o przyrodę, środowisko naturalne	1–6
H_t	Ważna jest tradycja, postępowanie zgodnie z tradycjami religijnymi lub rodzinnymi	1–6
H_u	Ważne jest poszukiwanie zabawy, robienie tego, co sprawia przyjemność	1–6

Źródło: opracowanie własne

Tabela A3. Kryteria warunkujące sposób estymacji DEFF dla danych ESS5-PL (ed. 2010)

Nr pytania	B1[i)]	A2[ii)]	~A2[iii)]			
		$\hat{\rho}$[iv)]	$\hat{\rho}_{I}$[v)]	$\hat{\rho}_{II}$	$\hat{\rho}_{III}$	$\hat{\rho}_{IV}$
Moduł A – media, zaufanie do ludzi						
A1	+	,065	,052	,160	,054	,045
A3	+	,096	,145	,091	,065	-,047
A5	+	,037	,015	,112	,100	,004
A8	+	,140	,131	,306	,087	,019
A9	+	,090	,122	,117	,033	-,036
A10	+	,163	,192	,211	,171	,033
Moduł B – polityka						
B1	+	,098	,164	,130	-,122	-,017
B4	+	,111	,127	-,118	,199	,145
B5	+	,137	,161	,232	,053	,066
B6	+	,071	,064	,148	,022	,095
B7	+	,131	,131	,077	,165	,140
B8	+	,135	,135	,153	,161	,115
B9	,016	,126	,110	,129	,160	,136
B10	+	,092	,088	,134	,067	,085
B23	,020	,121	,155	,038	,082	,105
B24	+	,073	,103	,065	,039	,030
B25	+	,157	,217	-,011	,264	-,002
B26	+	,136	,151	,108	,223	,033
B27	+	,118	,122	,215	,135	,062
B28	+	,158	,151	,329	,165	,047
B29	+	,199	,176	,220	,427	,115
B30	<,001	,155	,204	,125	,069	,039
B31	+	,131	,202	-,093	,068	,002
B32	,001	,128	,177	,271	,106	-,060
B33	,033	,070	,080	-,042	,053	,048
B38	+	,189	,197	,155	,168	,184
B39	+	,217	,226	,097	,264	,218
B40	+	,232	,257	,144	-,010	,325
Moduł C – poczucie dobrobytu, wykluczenie społeczne, religia						
C1	+	,069	,080	,162	,229	-,104
C4	+	,089	,100	,122	,141	,004
C6	,001	,145	,158	,072	,204	,029
C15	+	,001	-,012	,029	-,082	,066
C21	<,001	,088	,128	,206	,030	-,034

Nr pytania	B1[i)]	A2[ii)]	~A2[iii)]			
		$\hat{\rho}$[iv)]	$\hat{\rho}_{I}$[v)]	$\hat{\rho}_{II}$	$\hat{\rho}_{III}$	$\hat{\rho}_{IV}$
Moduł D – zaufanie do wymiaru sprawiedliwości						
D1	+	,147	,155	,170	,163	,090
D2	,040	,092	,105	,063	,150	,017
D3	+	,154	,166	,151	,172	,053
D4	,039	,280	,264	,219	,287	,375
D5	,003	,272	,247	,251	,376	,316
D6	,016	,282	,285	,330	,233	,310
D7	+	,137	,167	,226	,135	,040
D12	+	,193	,221	,204	,116	,174
D13	+	,227	,267	,222	,108	,179
D14	<,001	,367	,402	,164	,193	,485
D18	+	,161	,170	,092	,109	,255
D19	+	,183	,224	,135	,062	,194
D20	+	,164	,200	,165	,078	,118
D21	,002	,173	,188	,078	,217	,147
D22	+	,150	,196	-,032	,180	,124
D23	+	,169	,260	,291	,009	-,044
D24	+	,150	,190	,241	,105	,025
D25	+	,087	,096	,229	-,001	,077
D26	+	,146	,151	,124	,243	,063
D27	,026	,229	,203	,367	,372	,141
D28	,027	,062	,063	,248	,020	-,025
D31	,028	,202	,161	,263	,349	,193
D32	,034	,164	,169	,332	,222	,070
D33	<,001	,179	,173	,546	,063	,066
D34	,006	,161	,226	,106	,089	,021
D35	,025	,147	,180	,172	,095	,087
D36	+	,221	,241	,372	,168	,097
D37	+	,225	,274	,286	,137	,122
D40	,013	,148	,146	,175	,176	,131
D41	+	,102	,099	,065	,195	,080
D42	,002	,091	,138	-,032	,117	,013
Moduł F – charakterystyka społeczno-demograficzna						
F1	<,001	,024	,032	-,063	,017	,050
F41	,017	,163	,184	,216	,100	-,017
Moduł G – praca, rodzina, dobrobyt						
G4	+	,152	,161	,152	,190	,081
G5	<,001	,220	,196	,235	,267	,234
G6	,020	,171	,140	,440	,225	,090

Nr pytania	B1[i)]	A2[ii)]	~A2[iii)]			
		$\hat{\rho}$[iv)]	$\hat{\rho}_I$[v)]	$\hat{\rho}_{II}$	$\hat{\rho}_{III}$	$\hat{\rho}_{IV}$
Moduł H – skala ludzkich wartości						
H_a	+	,113	,105	,198	,111	,036
H_b	,027	,131	,158	,195	,036	,041
H_c	+	,168	,240	,057	,150	,053
H_d	+	,133	,191	-,071	,008	,108
H_e	,001	,160	,182	,054	,149	,149
H_f	+	,156	,158	,267	,188	,054
H_g	<,001	,147	,116	,163	,216	,193
H_h	<,001	,177	,204	,143	,123	,150
H_i	+	,087	,060	,126	,175	,087
H_j	+	,078	,109	,046	,101	-,047
H_k	+	,160	,206	,240	,088	-,030
H_l	+	,085	,082	,107	,193	,026
H_m	+	,134	,160	,257	,096	-,017
H_n	<,001	,156	,193	,013	,142	,160
H_o	+	,212	,254	,265	,051	,151
H_p	<,001	,203	,266	,206	,058	,128
H_q	+	,303	,389	,177	,100	,138
H_r	+	,261	,278	,147	,230	,275
H_s	+	,165	,178	,231	,136	,143
H_t	,001	,147	,160	,207	,057	,137
H_u	+	,114	,137	,184	-,004	,074

Źródło: Obliczenia własne

[i)] Kryterium B1 odnosi się do homogeniczności wariancji warstwowych. Hipotezę o równości wariancji zmiennych substantywnych weryfikowano w oparciu o test Levene'a, przyjmując poziom istotności $\alpha = 0{,}05$. W kolumnach zamieszczono wynik testu w postaci statystyki *p*-value dla tych zmiennych, dla których różnice były statystycznie istotne; oznaczenie + zastosowane zostało w odniesieniu do tych zmiennych, dla których nie było podstaw do odrzucenia hipotezy o homogeniczności wariancji.

[ii)] Kryterium A2 odnosi się do równości współczynników korelacji wewnątrzzespołowej we wszystkich warstwach populacji.

[iii)] Kryterium ~A2 oznacza odrzucenie warunku homogeniczności warstwowych współczynników korelacji wewnątrzzespołowej.

[iv)] $\hat{\rho}$ oznacza współczynnik korelacji wewnątrzzespołowej wyznaczony łącznie dla wiązkowanej części próby badawczej.

[v)] $\hat{\rho}_I$, $\hat{\rho}_{II}$, $\hat{\rho}_{III}$ oraz $\hat{\rho}_{IV}$ oznaczają współczynniki korelacji wewnątrzzespołowej wyznaczone w każdej z czterech warstw z wiązkowanej części próby badawczej.

Tabela A4. Oszacowanie mierników DEFF dla danych ESS5-PL (ed. 2010)

Nr pytania	$DEFF_s$ ii)	Wariant estymacji I i)				Wariant estymacji II				Wariant estymacji III			
		$DEFF_c^I$		$DEFF_{TOTAL}^I$		$DEFF_c^{II}$		$DEFF_{TOTAL}^{II}$		$DEFF_c^{III}$		$DEFF_{TOTAL}^{III}$	
		$\bar{l}$ iii)	$l^* = \bar{l}_w$ iv)	$\bar{l}$	$l^* = \bar{l}_w$	$\bar{l}$	$l^* = \bar{l}_w$	$\bar{l}$	$l^* = \bar{l}_w$	$\bar{l}$	$l^* = \bar{l}_w$	$\bar{l}$	$l^* = \bar{l}_w$
Moduł A – media, zaufanie do ludzi													
A1	1,001	1,082	1,101	1,098	1,116	1,105	1,127	1,111	1,132	1,104	1,124	1,105	1,126
A3	1,000	1,121	1,148	1,137	1,164	1,150	1,181	1,156	1,187	1,158	1,183	1,159	1,184
A5	,978	1,047	1,057	1,061	1,072	1,064	1,076	1,069	1,081	1,061	1,075	1,063	1,076
A8	,990	1,176	1,215	1,193	1,232	1,214	1,259	1,220	1,265	1,194	1,231	1,196	1,233
A9	,992	1,113	1,139	1,129	1,155	1,141	1,171	1,147	1,176	1,138	1,159	1,139	1,160
A10	1,008	1,206	1,251	1,222	1,269	1,248	1,301	1,254	1,307	1,255	1,305	1,256	1,306
Moduł B – polityka													
B1	,980	1,123	1,151	1,139	1,167	1,153	1,185	1,158	1,190	1,156	1,173	1,157	1,174
B4	1,007	1,141	1,172	1,157	1,188	1,173	1,209	1,178	1,215	1,183	1,225	1,185	1,227
B5	1,005	1,173	1,211	1,190	1,228	1,211	1,255	1,216	1,261	1,218	1,257	1,219	1,259
B6	1,000	1,090	1,110	1,105	1,125	1,114	1,137	1,119	1,142	1,120	1,143	1,121	1,145
B7	1,004	1,165	1,202	1,182	1,219	1,201	1,244	1,207	1,250	1,207	1,252	1,208	1,254
B8	1,004	1,171	1,208	1,187	1,225	1,208	1,251	1,213	1,257	1,214	1,260	1,215	1,261
B9	1,007	1,159	1,194	1,175	1,211	1,194	1,235	1,200	1,241	1,193	1,236	1,195	1,238
B10	1,001	1,116	1,142	1,132	1,158	1,144	1,174	1,150	1,180	1,144	1,173	1,146	1,175
B23	1,003	1,152	1,186	1,168	1,202	1,186	1,225	1,192	1,231	1,201	1,239	1,202	1,241
B24	1,002	1,092	1,112	1,107	1,128	1,117	1,140	1,122	1,146	1,131	1,153	1,132	1,155
B25	1,006	1,198	1,242	1,215	1,260	1,240	1,291	1,246	1,297	1,254	1,305	1,256	1,307
B26	1,009	1,171	1,209	1,188	1,226	1,208	1,252	1,214	1,258	1,212	1,257	1,214	1,259
B27	1,004	1,149	1,182	1,165	1,199	1,183	1,221	1,189	1,227	1,192	1,232	1,194	1,234
B28	,979	1,200	1,244	1,217	1,261	1,241	1,293	1,247	1,299	1,236	1,284	1,238	1,286

Nr pytania	$DEFF_S$ ii)	Wariant estymacji I i)				Wariant estymacji II				Wariant estymacji III			
		$DEFF_c^I$		$DEFF_{TOTAL}^I$		$DEFF_c^{II}$		$DEFF_{TOTAL}^{II}$		$DEFF_c^{III}$		$DEFF_{TOTAL}^{III}$	
		$\bar{l}$ iii)	$l^* = \bar{l}_w$ iv)	$\bar{l}$	$l^* = \bar{l}_w$	$\bar{l}$	$l^* = \bar{l}_w$	$\bar{l}$	$l^* = \bar{l}_w$	$\bar{l}$	$l^* = \bar{l}_w$	$\bar{l}$	$l^* = \bar{l}_w$
B29	,997	1,252	1,307	1,269	1,326	1,301	1,366	1,308	1,372	1,305	1,380	1,307	1,381
B30	,972	1,196	1,239	1,213	1,256	1,237	1,287	1,243	1,293	1,238	1,279	1,239	1,281
B31	,969	1,165	1,202	1,181	1,218	1,201	1,243	1,207	1,249	1,197	1,229	1,199	1,231
B32	1,003	1,162	1,197	1,178	1,214	1,197	1,239	1,203	1,245	1,215	1,252	1,216	1,254
B33	1,011	1,088	1,107	1,103	1,123	1,112	1,134	1,117	1,140	1,102	1,120	1,104	1,122
B38	1,011	1,238	1,291	1,255	1,309	1,286	1,347	1,292	1,353	1,288	1,350	1,290	1,351
B39	,997	1,273	1,334	1,291	1,353	1,327	1,397	1,333	1,403	1,332	1,406	1,334	1,407
B40	1,012	1,293	1,358	1,311	1,377	1,349	1,424	1,356	1,431	1,347	1,415	1,349	1,416
Moduł C – poczucie dobrobytu, wykluczenie społeczne, religia													
C1	1,005	1,087	1,106	1,102	1,121	1,110	1,132	1,116	1,138	1,122	1,147	1,123	1,149
C4	,999	1,112	1,137	1,128	1,153	1,140	1,169	1,146	1,175	1,147	1,176	1,148	1,177
C6	,966	1,182	1,223	1,199	1,240	1,221	1,268	1,227	1,274	1,210	1,253	1,211	1,255
C15	1,007	1,001	1,001	1,015	1,015	1,011	1,011	1,016	1,016	1,010	1,008	1,011	1,010
C21	,970	1,111	1,136	1,127	1,152	1,139	1,167	1,144	1,173	1,155	1,180	1,157	1,181
Moduł D – zaufanie do wymiaru sprawiedliwości													
D1	,990	1,186	1,227	1,202	1,244	1,225	1,273	1,231	1,279	1,229	1,276	1,230	1,278
D2	,999	1,116	1,142	1,132	1,158	1,145	1,174	1,150	1,180	1,148	1,178	1,150	1,180
D3	1,000	1,195	1,238	1,211	1,255	1,236	1,285	1,241	1,291	1,228	1,274	1,229	1,275
D4	,996	1,353	1,431	1,372	1,451	1,419	1,510	1,426	1,517	1,426	1,523	1,428	1,525
D5	,998	1,344	1,420	1,362	1,440	1,408	1,496	1,415	1,503	1,415	1,514	1,417	1,515
D6	1,002	1,356	1,435	1,375	1,455	1,423	1,514	1,429	1,521	1,435	1,530	1,437	1,532
D7	,993	1,173	1,211	1,189	1,228	1,210	1,255	1,216	1,261	1,229	1,274	1,231	1,275
D12	,985	1,244	1,298	1,261	1,316	1,292	1,355	1,298	1,361	1,305	1,367	1,307	1,368

Nr pytania	$DEFF_S$ ii)	Wariant estymacji I i)				Wariant estymacji II				Wariant estymacji III			
		$DEFF_c^I$		$DEFF_{TOTAL}^I$		$DEFF_c^{II}$		$DEFF_{TOTAL}^{II}$		$DEFF_c^{III}$		$DEFF_{TOTAL}^{III}$	
		$\bar{l}$ iii)	$l^* = \bar{l}_w$ iv)	$\bar{l}$	$l^* = \bar{l}_w$	$\bar{l}$	$l^* = \bar{l}_w$	$\bar{l}$	$l^* = \bar{l}_w$	$\bar{l}$	$l^* = \bar{l}_w$	$\bar{l}$	$l^* = \bar{l}_w$
D13	,979	1,286	1,349	1,304	1,368	1,341	1,415	1,348	1,421	1,350	1,419	1,352	1,420
D14	1,031	1,463	1,566	1,484	1,588	1,547	1,665	1,554	1,673	1,556	1,674	1,558	1,676
D18	1,000	1,203	1,248	1,220	1,266	1,246	1,298	1,251	1,304	1,263	1,320	1,265	1,322
D19	1,006	1,232	1,283	1,249	1,301	1,278	1,338	1,284	1,344	1,295	1,352	1,297	1,354
D20	1,006	1,207	1,253	1,224	1,271	1,250	1,303	1,256	1,310	1,261	1,310	1,262	1,312
D21	1,001	1,219	1,267	1,236	1,285	1,263	1,319	1,269	1,326	1,269	1,327	1,271	1,329
D22	1,000	1,189	1,231	1,206	1,248	1,229	1,278	1,235	1,284	1,250	1,301	1,251	1,303
D23	,993	1,213	1,260	1,230	1,278	1,257	1,311	1,263	1,317	1,281	1,325	1,282	1,326
D24	,991	1,190	1,232	1,206	1,249	1,230	1,278	1,236	1,285	1,244	1,289	1,245	1,290
D25	1,006	1,109	1,134	1,125	1,150	1,137	1,165	1,142	1,170	1,151	1,178	1,153	1,180
D26	,997	1,184	1,225	1,201	1,242	1,223	1,271	1,229	1,277	1,225	1,274	1,226	1,276
D27	1,005	1,289	1,353	1,307	1,372	1,345	1,419	1,351	1,426	1,347	1,427	1,348	1,429
D28	,999	1,078	1,095	1,093	1,111	1,100	1,120	1,106	1,126	1,102	1,118	1,103	1,120
D31	,996	1,255	1,311	1,272	1,330	1,305	1,370	1,311	1,377	1,304	1,378	1,306	1,380
D32	1,008	1,207	1,253	1,224	1,271	1,250	1,303	1,256	1,310	1,268	1,325	1,270	1,327
D33	,983	1,226	1,276	1,243	1,294	1,272	1,329	1,278	1,336	1,273	1,325	1,274	1,326
D34	,997	1,203	1,248	1,220	1,266	1,246	1,298	1,252	1,304	1,254	1,299	1,256	1,300
D35	,993	1,186	1,227	1,202	1,244	1,225	1,273	1,231	1,279	1,239	1,284	1,240	1,286
D36	,994	1,279	1,341	1,297	1,360	1,334	1,405	1,340	1,412	1,336	1,403	1,338	1,405
D37	1,002	1,285	1,348	1,303	1,366	1,340	1,412	1,346	1,419	1,356	1,425	1,358	1,427
D40	,996	1,187	1,229	1,204	1,246	1,227	1,275	1,233	1,281	1,233	1,284	1,235	1,286
D41	1,000	1,129	1,157	1,145	1,174	1,159	1,192	1,165	1,198	1,166	1,203	1,167	1,205
D42	1,004	1,115	1,140	1,130	1,156	1,143	1,172	1,148	1,178	1,158	1,187	1,160	1,189

Nr pytania	$DEFF_s$ ii)	Wariant estymacji I i)				Wariant estymacji II				Wariant estymacji III			
		$DEFF_c^I$		$DEFF_{TOTAL}^I$		$DEFF_c^{II}$		$DEFF_{TOTAL}^{II}$		$DEFF_c^{III}$		$DEFF_{TOTAL}^{III}$	
		$\bar{l}$ iii)	$l^* = \bar{l}_w$ iv)	$\bar{l}$	$l^* = \bar{l}_w$	$\bar{l}$	$l^* = \bar{l}_w$	$\bar{l}$	$l^* = \bar{l}_w$	$\bar{l}$	$l^* = \bar{l}_w$	$\bar{l}$	$l^* = \bar{l}_w$
Moduł F – charakterystyka społeczno-demograficzna													
F1	,987	1,030	1,036	1,044	1,051	1,044	1,052	1,049	1,057	1,050	1,058	1,052	1,060
F41	,956	1,206	1,251	1,223	1,269	1,248	1,301	1,254	1,307	1,223	1,263	1,225	1,264
Moduł G – praca, rodzina, dobrobyt													
G4	,986	1,192	1,235	1,209	1,252	1,233	1,282	1,238	1,288	1,234	1,283	1,235	1,284
G5	,970	1,278	1,340	1,296	1,359	1,333	1,404	1,339	1,410	1,328	1,403	1,330	1,405
G6	1,003	1,216	1,263	1,233	1,281	1,260	1,315	1,266	1,321	1,262	1,321	1,264	1,322
Moduł H – skala ludzkich wartości													
H_a	,998	1,142	1,174	1,158	1,190	1,175	1,211	1,181	1,217	1,165	1,197	1,166	1,199
H_b	1,000	1,165	1,202	1,181	1,218	1,201	1,243	1,207	1,249	1,201	1,236	1,202	1,237
H_c	,991	1,212	1,259	1,229	1,277	1,256	1,310	1,262	1,316	1,279	1,331	1,280	1,333
H_d	1,006	1,168	1,205	1,184	1,222	1,204	1,247	1,210	1,253	1,206	1,241	1,208	1,243
H_e	,986	1,202	1,247	1,219	1,264	1,244	1,296	1,250	1,302	1,249	1,301	1,251	1,303
H_f	1,002	1,197	1,241	1,214	1,258	1,239	1,289	1,244	1,295	1,240	1,289	1,241	1,291
H_g	,996	1,186	1,227	1,203	1,244	1,225	1,273	1,231	1,279	1,226	1,280	1,228	1,282
H_h	,991	1,223	1,273	1,240	1,291	1,269	1,326	1,275	1,332	1,278	1,333	1,279	1,335
H_i	,989	1,110	1,134	1,126	1,150	1,137	1,166	1,143	1,171	1,137	1,171	1,139	1,172
H_j	1,006	1,099	1,120	1,114	1,136	1,124	1,149	1,130	1,155	1,125	1,146	1,127	1,148
H_k	1,002	1,202	1,246	1,218	1,264	1,244	1,295	1,249	1,301	1,242	1,283	1,243	1,285
H_l	1,004	1,107	1,131	1,122	1,147	1,134	1,161	1,139	1,167	1,143	1,175	1,144	1,177
H_m	1,007	1,169	1,206	1,185	1,223	1,205	1,249	1,211	1,255	1,207	1,244	1,209	1,246
H_n	,992	1,197	1,241	1,214	1,258	1,239	1,289	1,244	1,295	1,255	1,308	1,257	1,309

Nr pytania	$DEFF_s$ ii)	Wariant estymacji I i)				Wariant estymacji II				Wariant estymacji III			
		$DEFF_c^I$		$DEFF_{TOTAL}^I$		$DEFF_c^{II}$		$DEFF_{TOTAL}^{II}$		$DEFF_c^{III}$		$DEFF_{TOTAL}^{III}$	
		$\bar{l}$ iii)	$l^* = \bar{l}_w$ iv)	$\bar{l}$	$l^* = \bar{l}_w$	$\bar{l}$	$l^* = \bar{l}_w$	$\bar{l}$	$l^* = \bar{l}_w$	$\bar{l}$	$l^* = \bar{l}_w$	$\bar{l}$	$l^* = \bar{l}_w$
H_o	1,006	1,268	1,327	1,285	1,345	1,320	1,388	1,326	1,395	1,327	1,389	1,329	1,390
H_p	,993	1,256	1,313	1,274	1,332	1,307	1,373	1,313	1,379	1,326	1,386	1,327	1,387
H_q	,985	1,383	1,468	1,402	1,488	1,454	1,551	1,461	1,559	1,443	1,524	1,445	1,526
H_r	1,009	1,329	1,402	1,348	1,421	1,391	1,475	1,398	1,482	1,395	1,480	1,396	1,481
H_s	,999	1,208	1,254	1,225	1,271	1,251	1,304	1,257	1,310	1,266	1,321	1,268	1,323
H_t	,994	1,186	1,227	1,202	1,244	1,225	1,273	1,231	1,279	1,232	1,277	1,234	1,279
H_u	1,006	1,144	1,176	1,160	1,192	1,177	1,214	1,182	1,220	1,181	1,213	1,183	1,214

Źródło: obliczenia własne

i) Warianty I, II oraz III estymacji mierników $DEFF_c$ oraz $DEFF_{TOTAL}$ zgodne z opisem w tabeli IV.3.

iii) W oszacowaniu miernika $DEFF_s$ wykorzystano wzór (IV.6''.), przyjmując jednocześnie, że dla każdej warstwy $DEFF_j$=1.

iii) Por. wzór (IV.28.).

iv) Por. wzory (IV.28'.) oraz (IV.28''.).

Representativeness of a representative study: analysis of selected methodological and practical problems within the total survey error paradigm

Summary

The monograph titled *Reprezentatywność badań reprezentatywnych. Analiza wybranych problemów metodologicznych oraz praktycznych w paradygmacie całkowitego błędu pomiaru* [*Representativeness of a Representative Study: Analysis of Selected Methodological and Practical Problems within the Total Survey Error Paradigm*] is devoted to theoretical issues and practical complications associated with methodological assessment of the level of survey sample representativeness. The main focus is on analyzing the consequences of certain practical actions. This monograph provides tools for assessing the level of sample representativeness. Critical analysis of certain survey and post-survey procedures whose correctness has until recently been taken for granted is of special significance. In the monograph, it is demonstrated that knowledge about the mechanisms behind survey sample representativeness is not easily translatable into practice, or in other words, that the reality of research is much more complex than the theory of representative sampling presupposes. This task required presenting the existing state of knowledge on this topic as well as the causes and consequences of complications associated with putting this knowledge into research practice.

The scope of the analyses that were carried out was limited to survey techniques that involve obtaining data by using personal, i.e. traditional or computer-assisted, standardized questionnaire-based interviews. All those survey techniques, including telephone interview techniques, that do not involve direct contact between the interviewer and the respondent were not taken into account. The study on the representativeness of survey samples was also restricted to scientific surveys, not because the author has a reluctant attitude toward commercial surveys, but because academic surveys are conducted for different purposes and the criteria that are used to assess their quality are also different. Therefore, the issue of whether surveys are *cost-consuming* is of secondary importance here. This is an intentional move that is aimed at emphasizing the issue of the *quality* of a representative sample. The last limitation of this monograph is related to the type of analyzed research sample. It does not deal with substituted samples, including samples that allow for replacing non-respondents with other persons who are selected by purposive or quota sampling. Moreover, this monograph only analyzes those sampling schemes that involve selecting population units (persons) based on random or quasi-random procedures. Thus, purposive, quota and systematic sampling schemes are not dealt with here.

This monograph consists of five chapters that are somewhat arbitrarily arranged into two parts. The first part, which comprises the first and second chapters, introduces the paradigm of total survey error and describes the sources of errors that make up total error. Both these chapters are of primary importance to the monograph as they define the scope of the author's study of selected issues related to the representativeness of survey samples, which is presented in the second part of this monograph (the third, fourth and fifth chapters).

The first chapter begins with a description of the total survey error paradigm and a discussion of the controversy associated with determining the size of this error, i.e. the difference between an approximate value of a certain parameter and the "actual" value of that parameter for the whole population. As a reference to literature review, a *statistical* and *psychometric* approach to assessing the quality of surveys is introduced here, which results in a search for external and internal standards for assessing surveys. The literature review that is presented in this chapter makes it possible to identify a set of errors that are recognized by the vast majority of methodologists who refer to the total survey error paradigm in their studies. A distinction is made between a random and a systematic component of total error. The final part of the first chapter is devoted to a total error measure, i.e. the *mean squared error*. The author proposes a modified version of the classical total error estimator which is based on defining a random component of this error in a different manner, i.e. in which the variance of estimators (for a given sample) refers to the theoretical variance of the estimators of the same parameters for a simple random sample.

The second chapter presents the issue of errors in survey research in more detail. The structure of this chapter corresponds to the classification of error sources, which is introduced in the first chapter. Thus, the first part of this chapter defines statistical sampling error as well as systematic errors that influence the level of sample representativeness, the second part describes measurement-related errors and the third part deals with post-survey processing errors. Although the second and third parts of this chapter go beyond the subject of this monograph, they allow one to formally define all types of errors that are analyzed in the first chapter as well as to systematize the literature on the subject and discuss the main theoretical and practical dilemmas associated with assessing the quality of survey research within the total survey error paradigm.

The third chapter is devoted to survey sampling frames. By referring to the concept of an "ideal" sampling frame and by describing the relations between the target population and the frame population, the author introduces four major classes of errors arising from the deficiencies in the registers that are used to take representative samples, i.e. undercoverage errors, overcoverage errors as well as errors related to multiplying the selection probability and to grouping / clustering units. Procedures for improving the quality of sampling frames are also discussed. The literature review, which was conducted by referring to empirical analysis, showed that the benefits of reducing systematic undercoverage error are usually eliminated by measurement errors. The second part of this chapter presents the results of analyses that were carried out by the author with the aim of exemplifying selected practical problems. These analyses were conducted on data repositories that had been established for the purpose of the European Social Survey project. Also, the topic of within-household selection of individual units from cluster / address frames is discussed, as well as the issue of the within-unit undercoverage errors. Additionally, this chapter presents the possibilities of using Polish public administration registers (PESEL and TERYT) in representative sampling.

The fourth chapter presents a discussion of the consequences of using particular survey sampling schemes. A simple sampling scheme, together with its definitional characteristics, i.e. unrestricted sampling, individual sampling, equal probability sampling and one-stage sampling, was the reference point. While the distinction between individual and address samples that was made in the third chapter resulted from waiving the one-stage sampling condition, the fourth chapter focuses on the consequences of waiving the condition related to the first three characteristics of a simple sample. First, stratified sampling is discussed, and then the cluster sampling scheme as well as the unequal probability sampling scheme. As part of the study, factors that determine a specific effectiveness of given sampling schemes (i.e. lower or higher effectiveness relative to a simple sample) were also discussed. The second part of the fourth chapter is devoted to the practical complications associated with attempts that are made to empirically assess the effectiveness of a given sampling scheme.

The last chapter explores selected methodological problems that arise from an incomplete response rate. First, the basic issues related to the post-implementation classification of sampled units are discussed. Following the postulate of developing uniform standards for determining the values of sample implementation indicators, sampled individuals were divided into clusters of: (1) respondents, (2) non-respondents (including (2.1) individuals who were non-respondents due to the lack of contact, (2.2) refusal or (2.3) another reason for failure to conduct an interview), (3) units of undetermined status as far as the target population was concerned, and (4) units that did not belong to the population. The second part of the fifth chapter analyzes issues related to the field equivalence of address and individual samples. The issue of comparability of individual and cluster frame samples is also discussed in the third chapter. However, while the third chapter focuses on translating the process of selecting units within clusters into errors of undercoverage (or overcoverage) with regard to the units that belong to the target population, which is characteristic of cluster samples, the analyses that are presented in the fifth chapter are aimed to answer the question as to whether using different types of sampling frames has a significant influence on the differences between sample implementation indicators as well as on the post-survey structure of the sets of respondents and non-respondents. The final part of the fifth chapter focuses on assessing sample representativeness in light of an incomplete "in-the-field" response rate. It also describes the assumptions of two paradigms, i.e. deterministic and probabilistic ones, within which non-response errors are analyzed, as well as the idea and possibility of using the so-called indicator of the representativeness response in a practical manner. This chapter closes with an example demonstrating the benefits of analyzing sample representativeness in light of the probabilistic paradigm of error resulting from an incomplete response rate.

The conclusion presents a set of recommendations and guidelines which will make it easier for researchers to assess the level of survey sample representativeness, and it also indicates problem areas which require further theoretical and methodological research.