Uniwersytet im. A.Mickiewicza
Wydział Nauk Społecznych
Instytut Psychologii UAM

Paweł Kleka

Zastosowanie teorii odpowiadania na pozycje testowe (IRT) do tworzenia skróconych wersji testów i kwestionariuszy psychologicznych

Rozprawa doktorska
przygotowana pod kierunkiem
prof. dra hab. Jerzego Brzezińskiego

Poznań 2012

Wstęp **5**

Rozdział 1. Teoria odpowiadania na pozycje testu (IRT) **8**

1.1. Pomiar psychologiczny – podstawowe definicje **8**

1.2. Założenia teorii odpowiadania na pozycje testu **13**

1.3. Modele dwukategorialne IRT **18**

1.4. Modele wielokategorialne IRT **23**

1.5. Rzetelność i błąd standardowy pomiaru **25**

1.6. Ewaluacja modeli opartych na IRT **29**

1.6.1. Dopasowanie modelu IRT do formatu pozycji testowych 29

1.6.2. Kryteria dopasowania modelu IRT do danych 30

1.7. Tworzenie skróconych wersji testów - ujęcie w ramach Klasycznej Teorii Testów i IRT **33**

Rozdział 2. Tworzenie komputerowych wersji psychologicznych testów i kwestionariuszy **37**

2.1. Użycie komputerów i internetu w badaniach psychologicznych **37**

2.1.1. Zjawisko „przepaści cyfrowej" – analfabetyzm komputerowy 39

2.1.2. Technofobia – lęk przed komputerem i internetem 41

2.1.3. Klasyfikacja zachowania się osób uczestniczących w badaniach internetowych 44

2.2. Podejście tradycyjne – przełożenie metody „papier i ołówek" na wersję elektroniczną **47**

2.2.1. Elementy testu i/lub kwestionariusza w badaniach komputerowych 48

2.2.2. Wpływ formy kontaktu z osobami uczestniczącymi w badaniach na odsetek odpowiedzi 50

2.2.3. Ograniczenia badań przez internet związane z użyciem stron internetowych. 51

2.2.4. Wpływ kontekstu wizualnego na wyniki uzyskiwane w badaniach internetowych 52

2.3. Podejście dynamiczne – kwestionariusze i testy adaptacyjne **54**

2.4. Porównanie jakości danych uzyskiwanych w papierowych i elektronicznych wersjach testów i kwestionariuszy **57**

2.4.1. Metaanaliza różnic między wynikami papierowych i komputerowych wersji narzędzi badawczych 60

Rozdział 3. Modele powiązań zmiennych. Problemy i hipotezy badawcze ..65

3.1. Zmienne niezależne 65

3.1.1. Moc różnicująca pozycji testowych lub kwestionariuszowych 65

3.1.2. Poziom trudności pozycji testowych lub kwestionariuszowych 66

3.1.3. Wielkość próby 67

3.2. Zmienne zależne 68

3.2.1. Błąd standardowy pomiaru 68

3.2.2. Skośność rozkładu wyników 68

3.2.3. Kurtoza rozkładu wyników 68

3.3. Problemy badawcze 69

3.4. Hipotezy 70

3.4.1. Wpływ wielkości próby na estymację parametrów modeli IRT 71

3.4.2. Wpływ parametrów modeli IRT na estymację wyników kwestionariuszy i testów 71

3.4.3. Jakość wersji skróconych w oparciu o IRT versus inne metody skracania 71

3.4.4. Zróżnicowanie wyników w kwestionariuszowym badaniu psychologicznym ze względu na sposób przeprowadzenia badania 72

3.5. Założone modele IRT 72

3.5.1. Model dychotomiczny, model trójparametryczny (3PL) 72

3.5.2. Model politomiczny, model klasy odpowiedzi (GRM) 73

Rozdział 4. Przeprowadzone badania i ich wyniki 74

4.1. Organizacja badań i opis procedury badawczej 74

4.1.1. Symulacja wyników w modelu dychotomicznym 75

4.1.2. Symulacja wyników w modelu politomicznym 79

4.2. Operacjonalizacja zmiennych 82

4.3. Osoby uczestniczące w badaniach 84

4.4. Szacowanie wielkości próby kalibracyjnej 89

4.5. Opis statystyczny uzyskanych wyników 92

4.6. Testowanie hipotez 100

4.6.1. Wpływ wielkości próby na dokładność estymacji parametrów modeli IRT 100

4.6.2. Wpływ długości skróconej wersji testu na wyniki 102
4.6.3. Równoważność wyników skróconych wersji testu 105
4.6.4. Równoważność skróconych wersji kwestionariuszy konstruowana za pomocą różnych technik.....107
4.6.5. Wersje papier-i-ołówek oraz adaptacyjna a wyniki kwestionariuszy osobowości 111
4.6.6. Wpływ wersji narzędzia na zróżnicowanie wyników w teście i kwestionariuszu.....112
4.6.7. Zróżnicowanie czasu odpowiedzi w różnych typach testów i kwestionariuszy 113

Rozdział 5. Dyskusja.....116

5.1. Wnioski i znaczenie wyników 117
5.1.1. Implikacje psychometryczne 118
5.1.2. Implikacje wyników dla praktyki psychologicznej.....119

5.2. Ograniczenia i przyszłe obszary badań.....120

Słownik symboli.....122

Bibliografia 125

Załącznik 1: kod programu R wykorzystany w badaniu.....147

Załącznik 2: skrócone wersje kwestionariusza PTS 160

Załącznik 3.1: artykuły wykorzystane w meta-analizie 161

Załącznik 3.2: podsumowanie wielkości efektów wyników badań uwzględnionych w meta-analizie.....163

Załącznik 4: instrukcja do systemu badań internetowych.....164

Wstęp

W literaturze poświęconej zagadnieniom badań testowych i kwestionariuszowych można zaobserwować rosnące zainteresowanie metodami probabilistycznymi, w tym testowaniem adaptacyjnym. Testowanie adaptacyjne wymyślono już na początku XX wieku – procedura testowania polegała na użyciu wcześniej skalibrowanych pozycji testowych, prezentowaniu ich w określonej kolejności zaczynając od wybranej pozycji oraz zakończeniu testowania po spełnieniu pewnego warunku, co odpowiada w pełni procedurze współczesnych testów adaptacyjnych (więcej na ten temat w rozdziale 2.3). Mimo tak długiej historii, stosowanie testów adaptacyjnych nie było możliwe na większą skalę ze względu na fakt, że całą administracyjną pracę podczas testowania musiał wykonywać człowiek. Dopiero od kilku lat do administrowania testów adaptacyjnych można wykorzystywać współczesne komputery (Butcher, Perry i Hahn, 2004). Jednocześnie można zauważyć, że mimo rosnącego zainteresowania testowaniem adaptacyjnym, nie wzrasta liczba zastosowań metod opartych na dynamicznej konstrukcji narzędzi w praktyce badawczej i diagnostycznej. Jednym z powodów takiego stanu rzeczy jest złożona teoria leżąca u podstaw metod adaptacyjnych oraz nikła liczba opracowań przybliżających w sposób przystępny tę, bądź co bądź, złożoną problematykę. Większość zastosowań dotyczy testów wiedzy, gdzie obliczenia są najprostsze i skąd wywodzi swe korzenie teoria odpowiadania na pozycje testowe. Lecz rozwój tej teorii na przestrzeni ostaniach 50-60 lat skutkujący wprowadzaniem kolejnych modeli, o coraz wyższym stopniu skomplikowania aparatu matematycznego, dawno już pozwala stosować tę teorię w odniesieniu do kwestionariuszy psychologicznych. Z jednej strony brak empirycznych opracowań na gruncie psychologii, a z drugiej strony duży potencjał metod opartych na teorii odpowiadania na pozycje testowe skłonił mnie do napisania tej pracy.

Rozprawa niniejsza ma charakter eksploracyjny, a jej głównym zadaniem jest sprawdzenie warunków towarzyszących tworzeniu i używaniu skróconych (w oparciu o teorię odpowiadania na pozycje testowe) wersji testów i kwestionariuszy. Oprócz określenia jak na uzyskiwane w narzędziach badawczych wyniki wpływają parametry wybieranych doń pozycji testowych i kwestionariuszowych, uwagę poświęcę także wpływowi medium elektronicznego na jakość i rzetelność danych. Zrealizowanie tego dodatkowego celu powinno umożliwić częstsze i łatwiejsze stosowanie metod komputerowych w naukach społecznych. Oczywiście tylko pod warunkiem, że wyniki

uzyskiwane w tych specyficznych, bo elektronicznych wersjach, nie będą odbiegać od wyników oryginałów.

W pierwszym rozdziale pracy wprowadzę definicje podstawowych pojęć oraz omówię teorię odpowiadania na pozycje testowe. Modele jednowymiarowe zostaną szczegółowo scharakteryzowane zaczynając od najprostszych – dwukategorialnych – stosowanych przy testach wiedzy, a kończąc na złożonych – wielokategorialnych – pozwalających na rozpatrywanie pozycji testowych z kodowaniem odpowiedzi na skalach porządkowych. Z obszernej dziedziny modeli opisujących odpowiadanie na pozycje testowe pominięty zostanie stosunkowo słabo eksplorowany obszar modeli wielowymiarowych, który jest rozszerzeniem modeli jednowymiarowych.

W rozdziale drugim omówię uwarunkowania użycia komputerów do przeprowadzania badań. Wskażę newralgiczne obszary konstruowania narzędzi elektronicznych związane z cechami osób uczestniczących w badaniach. Omówię także cechy osobowości związane z używaniem i upowszechnieniem technologii cyfrowej, które powinny być brane pod uwagę w procesie tworzenia i wykorzystywania elektronicznych narzędzi badawczych. W dalszej części rozdziału drugiego skupię się na właściwościach samych narzędzi badawczych, rozpatrując kolejno: formę zapraszania do badań, kontekst wizualny, pozycje testowe lub kwestionariuszowe. Nie pominę też ograniczeń związanych z użyciem komputerów jako medium, problemu doboru ochotniczego, problemu randomizacji oraz anonimowości w próbach internetowych.

Na koniec drugiego rozdziału przeprowadzę porównanie wyników papierowych i elektronicznych wersji narzędzi badawczych oparte na szerokim przeglądzie literatury, z wykorzystaniem meta-analizy. Stwierdzone różnice (lub ich brak) pozwolą badaczom podejmować z większą świadomością decyzje, jaką formę badań wybierać oraz jak wpływa ona na uzyskane ostatecznie wyniki.

Rozdziały trzeci i czwarty mają charakter empiryczny i poświęcone są przedstawieniu problemów badawczych. Omawiam tutaj dwa wybrane narzędzia badawcze w postaci testu inteligencji Omnibus i kwestionariusza temperamentu PTS oraz sposób symulacji wyników dla dużych prób wykorzystany do sprawdzania części hipotez. Rozdział trzeci poświęcony jest ponadto przedstawieniu założonych modeli badawczych, zaś czwarty - oprócz wyników badań zawiera także opis próby oraz rozważania na temat jej optymalnej wielkości.

Pracę kończy dyskusja uzyskanych wyników zawarta w rozdziale piątym. Dowodzę w niej użyteczności wersji elektronicznych oraz ich równoważności z wersjami

papierowymi. Staram się ponadto wykazać, że wykorzystując nowoczesne technologie komunikacyjne, tradycyjne kwestionariusze przełożone na wersję elektroniczną mają szansę uzyskać nowy wymiar interakcyjności. Największy udział w zmianie poszerzającej możliwości testów i kwestionaiuszy należy do testowania adaptacyjnego, które pozwala w trakcie badania dostosowywać narzędzie badawcze do poziomu badanej cechy osoby uczestniczącej w badaniu, w oparciu o uzyskiwane odpowiedzi, nie tracąc jednocześnie kontroli nad jakością uzyskiwanych danych.

Rozdział 1. Teoria odpowiadania na pozycje testu (IRT)

1.1. Pomiar psychologiczny – podstawowe definicje

W psychologii, jak w każdej dyscyplinie empirycznej, dąży się do konstruowania modeli opisujących rzeczywistość. W oparciu o przyjęte teorie empiryczne, za pomocą stosunkowo prostych reguł, modeluje się związki ilościowe i jakościowe pomiędzy właściwościami mierzonych obiektów. Postępowanie takie, służące do tworzenia reprezentacji tych właściwości za pomocą liczb, czyli określenia wartości (natężenia, wielkości, częstotliwości) cech, zjawisk lub zachowań właściwych dla obszaru badawczego psychologii, nazywane jest pomiarem (por. Grobler, 2006, s. 152). Pomiarowi podlegają przede wszystkim cechy osobowości, style zachowania, postawy, przekonania, zainteresowania, zdolności, umiejętności, preferencje, itp..

Charakterystyczne cechy pomiaru psychologicznego można ująć w pięciu punktach (por. Crocker i Algina, 1986 za: Hornowska, 2001, s. 18):

1. Cechy psychologiczne powinny być definiowane w terminach związków z obserwowalnymi zjawiskami. Innymi słowy, aby wynik pomiaru miał użyteczną wartość musi być wyrażony w takich kategoriach, które mogą być mierzone – np.: częstość zachowań lub ich zakres.
2. Żadna realizacja pomiaru psychologicznego nie ma charakteru uniwersalnego. Ponieważ mierzy się obserwowalne zjawiska i wnioskuje na ich podstawie o cechach ukrytych (latentnych), definiując pomiar należy określić nie tylko sam konstrukt, ale również jego ramy teoretyczne. Wynik pomiaru powinien posiadać odniesienie teoretyczne, które wiąże go z innymi konstruktami i stanowi podstawę interpretacji znaczenia uzyskanego wyniku.
3. Pomiar psychologiczny jest najczęściej oparty na ograniczonej próbce zachowań, co oznacza konieczność wyboru z puli wszystkich możliwych manifestacji mierzonej cechy tych zachowań, które w ramach danej teorii będą najbardziej adekwatne. Równie ważne jest zdefiniowanie kryteriów tego wyboru.
4. Wyniki pomiaru psychologicznego są zawsze obarczone pewnym błędem. Pomiar dokonywany jest w ograniczonym czasie, na ograniczonej puli zachowań, często na podstawie introspekcji (samoopisu). Jest przez to podatny na różnego rodzaju dystraktory, co przekłada się na obciążenie nieuniknionym błędem. Błąd ten wynika na przykład z podatności na aprobatę społeczną, motywacji do określonej

autoprezentacji, symulacji lub dysymulacji, ale także np. ze stylów odpowiadania (por. Paluchowski, 1983; Zawadzki, 2000).

5. Wyniki pomiaru psychologicznego nie zawsze przekładają się na skale o dobrze zdefiniowanych jednostkach. Ze względu na specyficzny przedmiot pomiaru psychologicznego, skonstruowana skala teoretyczna nie zawsze dobrze oddaje różnice między badanymi osobami.

Pomiar psychologiczny łączy, z pewną dokładnością, obserwowalne zachowania z ich ukrytymi „motorami", dlatego też niezbędna staje się teoria wyjaśniająca uzyskiwane wyniki, która stanowi podstawę interpretacji. Na gruncie przyjętej teorii możliwe jest wyjaśnianie związków między wynikami a cechami psychologicznymi, czyli pewnymi parametrami „[…] rozkładu charakteryzującego częstości występowania określonych zachowań człowieka" (Nowakowska, 1975, s. 20).

Od początków XX wieku włożono wiele wysiłku w badania nad pomiarami, np.: inteligencji (Spearman, 1904) czy postaw (Thurstone, 1928), a także różnych zmiennych w obszarze psychologii społecznej (Lewin, 1936), czy psychologii uczenia się (Hull, 1943). Powstało w tym celu wiele testów i kwestionariuszy psychologicznych „pozwalających na uzyskanie takiej reprezentatywnej próbki zachowań, o których można przyjąć założenie (na podstawie założeń teoretycznych lub związków empirycznych), że są one wskaźnikami interesującej […] cechy psychologicznej (Hornowska, 2001, s. 22). Pojęcie „test" oznacza narzędzie, za pomocą którego ocenia się poprawność lub jakość odpowiedzi w odniesieniu do pewnego standardu (por. SdTSwPiP, 2007). W testach udzielane przez osoby badane odpowiedzi są oceniane pod kątem ich poprawności, co przekłada się na wzrost lub spadek sumarycznego wyniku osoby uczestniczącej w badaniu. Natomiast kwestionariusze stanową grupę narzędzi badawczych, w których odpowiedzi wskazują na poziom natężenia badanych cech.

Przedmiotem analizy będą zatem zarówno testy właściwości poznawczych (posiadające poprawną odpowiedź, wymagające wiedzy, umiejętności czy zdolności), jak i kwestionariusze właściwości afektywnych (oparte na autoekspresji, wymagające od osoby uczestniczącej w badaniu bazalnej samoświadomości w zakresie własnych uczuć, postaw, przekonań, emocji, itp.).

Sposoby konstruowania narzędzi do pomiaru właściwości psychicznych człowieka – choć rola teorii jest w nich decydująca – akcentują rolę procedur empirycznych przy doborze wskaźników mierzonej właściwości psychicznej (por. np. strategie konstrukcji kwestionariuszy – Zawadzki, 2006). Narzędzia, które wykorzystywane są w praktyce

psychologicznej, składają się z **pozycji** (testowych, kwestionariuszowych). Wynikiem uzyskiwanym przez osobę uczestniczącą w badaniu danym narzędziem jest wartość symbolizująca poziom badanej cechy, ustalana na podstawie liczby odpowiedzi zgodnych z kluczem. Odsetek poprawnych odpowiedzi w badanej populacji wskazuje na **poziom trudności** (*difficulty*) poszczególnych pozycji narzędzia. Inaczej mówiąc prawdopodobieństwo udzielenia poprawnej odpowiedzi na daną pozycję w teście, lub odpowiedzi zgodnej z kluczem w kwestionariuszu, jest większe u osób charakteryzujących się wyższym poziomem badanej cechy (Guilford, 1954, s. 344). W podejściu klasycznym poziom trudności danej pozycji inaczej definiowany jest dla testu, a inaczej dla kwestionariusza. Dla pozycji testowej oparty jest on o prawdopodobieństwo udzielenia poprawnej odpowiedzi – dana pozycja ma taką trudność (wyrażoną w poziomie badanej cechy), dla której istnieje 50% prawdopodobieństwo udzielenia poprawnej odpowiedzi (Gulliksen, 1950, s. 369, Guilford, 1954, s. 419). Dla pozycji kwestionariuszowej nie można określić poprawności odpowiedzi, ale nasilenie cechy. Dlatego poziom trudności określany jest osobno dla każdej z poszczególnych kategorii odpowiedzi. Oznacza to w tym przypadku, że osoba uczestnicząca w badaniu, która zaznaczy daną odpowiedź, ma z 50% prawdopodobieństwem wyższy poziom badanej cechy od tego właśnie poziomu zmiennej latentnej (*between category thresholds* – Embretson i Reise, 2000, s. 312).

Do reprezentowania zjawisk psychicznych, zarówno na etapie konstrukcji, jak i działania narzędzia (stosowania testu bądź kwestionariusza), wykorzystuje się modele matematyczne, które pozwalają opisywać badane właściwości. Orzekanie o aktualnych zachowaniach pełni funkcję diagnostyczną, zaś przewidywanie przyszłych zachowań – funkcję prognostyczną wyniku (Brzeziński, 2000, s. 402). Funkcje te realizowane są zarówno w badaniach naukowych, jak i diagnostycznych (Maloney i Ward, 1976). W tych pierwszych celem jest dostarczenie danych pozwalających budować lub weryfikować teorie stawiane przez badaczy. W tych drugich celem pomiaru jest dostarczenie danych „do podjęcia decyzji o działaniach zmierzających do zmiany aktualnego stanu (położenia) psychospołecznego ludzi" (Paluchowski, 1991, s. 32).

Z uwagi na duże znaczenie przyjętego modelu teoretycznego dla interpretacji i rozumienia wyników to procedura budowy **narzędzia do pomiaru** zaczyna się nie od konstruowania pozycji, ale od konceptualizacji tego, czego ma ono dotyczyć. Dopiero potem określany może być rodzaj bodźca oraz format odpowiedzi. Następnie można przystąpić do budowania narzędzia poprzez wygenerowanie pozycji poddawanych

analizom językowym, treściowym i statystycznym pozwalającym wybrać najlepsze z nich.

Badacz podejmujący się skonstruowania narzędzia do pomiaru cech psychicznych człowieka w oparciu o pewien model teoretyczny, musi zmierzyć się z kilkoma ważkimi problemami. Należą do nich: 1) kwestia istnienia reprezentacji, 2) jednoznaczność narzędzia, 3) jego trafność oraz 4) rzetelność (por. Hornowska, 2001, s. 158; Brzeziński, 2000, s. 401). Pokrótce zostaną one tutaj omówione.

Problem istnienia reprezentacji, to konieczność takiego zoperacjonalizowania konstruktów teoretycznych leżących u podstaw mierzonych zachowań, aby istniała relacja między tymi zachowaniami a teoretycznymi czynnikami je wywołującymi. Innymi słowy, proces operacjonalizacji to powiązanie „terminów teoretycznych (odnoszących się do nieobserwowalnych właściwości zdarzeń i obiektów) z terminami obserwacyjnymi (oznaczającymi obserwowalne właściwości i relacje)" (Hornowska, 2001, s. 161). W skrajnym przypadku możliwe jest zatem zbudowanie narzędzia pomiarowego, które nie będąc dobrze osadzone w teorii będzie mierzyło „coś", co nie ma odniesienia teoretycznego i znaczenia praktycznego – mimo istnienia narzędzia nie będzie istniał przedmiot pomiaru ani jego właściwa reprezentacja.

Następny **problem jednoznaczności narzędzia** – oznacza niebezpieczeństwo zbudowania takiego zbioru pozycji (zachowań), które będzie mierzył nie tylko konstrukt leżący u ich podstaw, ale też inne konstrukty, niekoniecznie ważne w danym badaniu. Innymi słowy mówiąc można zbudować narzędzie badawcze będące zestawem pozycji testowych nie powiązanych ze sobą żadnym wspólnym czynnikiem. W takiej sytuacji uzyskiwane wyniki nie będą jednoznacznie wskazywać na to, co badacz zamierzał poddać pomiarowi. W efekcie wartość praktyczna wyników będzie bardzo niska, a wnioski na ich podstawie wysnuwane – co najmniej niejednoznaczne.

Z kolei rozwiązanie **problemu trafności** i **rzetelności** narzędzia to udzielenie odpowiedzi na pytania o to „co narzędzie mierzy i jak dobrze to robi?" oraz określenie wielkości związku między wynikiem otrzymanym a prawdziwym. Trafność jako stopień poprawności operacjonalizacji wielkości psychicznej, informuje o stopniu, w jakim traktowane łącznie pozycje danego narzędzia reprezentują tę wielkość. Rzetelność w klasycznym modelu wyniku prawdziwego (Gulliksen, 1950, s. 222) definiowana jest jako stosunek wariancji wyników prawdziwych do wariancji wyników otrzymanych. Z kolei w teorii odpowiadania na pozycje testowe (Lord i Novick, 1968, s. 139) rzetelność jest odwrotnością błędu pomiarowego – wysoka rzetelność oznacza mniejszy udział błędu w otrzymanym wyniku.

Psychometria, jako dział psychologii zajmujący się „określaniem warunków, jakie powinny spełniać narzędzia wykorzystywane do pomiaru cech psychologicznych, oraz budowaniem modeli wiążących wyniki takich pomiarów z rzeczywistymi wartościami mierzonych cech" (Hornowska, 2001, s. 20) rozwija się w dwóch głównych nurtach: teorii losowego doboru próby (*random sampling theory* – RST[1] – Gulliksen, 1950) oraz teorii odpowiadania na pozycje testu (*item response theory* – IRT – Lord, 1953). W RST, dzisiaj znanej jako klasyczny model wyniku prawdziwego lub klasyczna teoria testów (KTT), głównym obszarem zainteresowań jest problem generalizacji wyników uzyskanych w próbie na wyniki przewidywane w populacji. Według KTT wynik osoby badanej (wynik otrzymany) jest powiązany z poziomem mierzonej cechy (wynik prawdziwy) w sposób zależny od charakterystyki próby standaryzacyjnej i parametrów pozycji narzędzia. W IRT z kolei zakłada się, że sposób odpowiadania na poszczególne pozycje pozwala określić rzeczywistą wartość wyniku dla danej osoby w sposób niezależny od próby, a jedynie w oparciu o parametry tychże pozycji. Z uwagi na fakt, że IRT odwołuje się do silniejszych założeń (więcej na ten temat: podrozdział 1.3), to wyniki otrzymywane w tym podejściu są mocniejsze, wykraczające często poza rezultaty dostępne w KTT. Dzięki temu, że trudność pozycji oraz poziom badanych cech wyrażone są na tej samej skali ilościowej, możliwe są precyzyjne porównania zarówno między osobami, jak i między pozycjami (*person fit* oraz *item fit statistics*). Ponadto parametry poszczególnych pozycji szacowane w oparciu o IRT są niezależne od próby i niezależne od narzędzia pomiarowego. Ta właściwość umożliwia stosowanie badania adaptacyjnego, w którym kolejność kolejno prezentowanych pozycji testowych lub kwestionariuszowych osobie uczestniczącej w badaniu zależy od udzielonych już przez nią odpowiedzi. Ponadto możliwe jest konstruowanie narzędzi pomiarowych i dobór do nich takich pozycji, które wynikają ze specyficznych potrzeb (np. selekcja, diagnoza), a otrzymane wyniki wciąż będą porównywalne między sobą. Między innymi dlatego IRT czasami nazywana jest silną teorią wyniku prawdziwego (*strong true score theory*), nie tyle będącą w opozycji do klasycznej teorii testów, co raczej rozszerzającą jej możliwości. Wspólną cechą charakterystyczną obu nurtów psychometrii (klasycznej i probabilistycznej) jest to, że właściwości świata realnego reprezentuje się w pewnym systemie abstrakcyjnym, a dokonując określonych predykcji według przyjętych założeń, porównuje się je z danymi empirycznymi. W ten sposób określa się stopień zgodności

[1] Ten skrót i wszystkie następne wraz z objaśnieniem trudniejszych terminów można znaleźć w słowniku na końcu pracy.

modelu z rzeczywistością, czyli np. z konfiguracją cech psychicznych, które są nieobserwowalne w sposób bezpośredni.

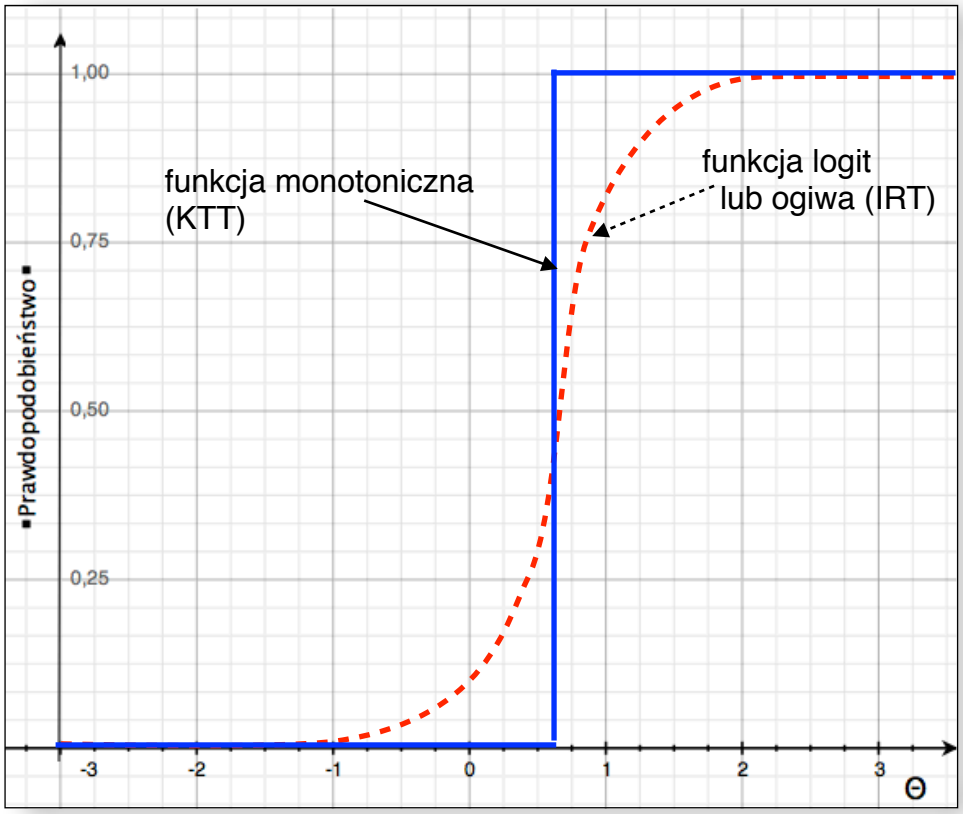
Ostatecznie narzędzie pomiarowe jest więc efektem wielu decyzji. Zarówno tych, jakie podjął badacz, mierząc się z wyżej wymienionymi problemami, a także decyzji związanych z wyborem procedur dotyczących konstrukcji testów i kwestionariuszy. Szczególnie tych decyzji, które służą określeniu obszaru zachowań identyfikowanych z mierzonym konstruktem (por. termin: „operacjonalizacja" – Hornowska, 2001).

1.2. Założenia teorii odpowiadania na pozycje testu

Początki IRT – „zbioru twierdzeń opisującego sposób, w jaki osoba badana odpowiada na pozycje testu" (Hornowska, 2001, s. 195) – należy datować na wczesne lata 40. od publikacji Derricka N. Lawley'a (1943) nt. krzywych charakterystycznych dla pozycji testowych (*ICC – Item Characteristic Curve*). Lawely postulował, iż wynik w teście jest spowodowany wartością nieobserwowalnej (latentnej) cechy, a każda pozycja testowa jest z nią związana w pewien niepowtarzalny sposób. W 1953 Fredric M. Lord sformalizował te postulaty budując modele matematyczne do ich opisu, a w 1968 opublikował wraz z Melvinem Novickiem monografię *Statistical theories of mental test scores* traktującą o praktycznym zastosowaniu teorii odpowiadania na pozycje testowe, w której rozdziały Allana Birnbaum'a wniosły statystyczne podstawy do praktycznego zastosowania IRT – od modelu jednoparametrycznego (uwzględniającego tylko stopień trudności pozycji testowych) do trójparametrycznego modelu (Lord i Novick, 1968).

Kluczowe dla modeli IRT jest traktowanie wyniku jako informacyjnego wskaźnika pewnej wartości nieobserwowalnej cechy (oznaczanej przez Θ), uzyskanego z pewnym prawdopodobieństwem przez osobę odpowiadającą na daną pozycję testową. Przyjmuje się, że wynik różnicuje osoby w punkcie b, będącym parametrem trudności danej pozycji testowej lub kwestionariuszowej. Osoby o poziomie mierzonej cechy wyższym niż wartość b konkretnej pozycji otrzymają odpowiedź kodowaną jako 1 – „sukces", zaś osoby o poziomie niższym niż wartość b tej pozycji – odpowiedź kodowaną jako 0 – „porażka". Dla pozycji kwestionariuszowych z wieloma kategoriami odpowiedzi, poziom b_i odpowiada progowi oddzielającemu osoby z cechą niższą i wyższą niż wartość b tej pozycji. Osoby z niższym poziomem cechy otrzymają odpowiedź kodowaną jako j-ta kategoria, zaś osoby z wyższym poziomem cechy jako kategoria j+1.

Dla modeli IRT kształt rozkładu prawdopodobieństwa opiera się o funkcję ogiwy rozkładu normalnego lub o funkcję logitową (obie krzywe mają zbliżony kształt – por. ryc. 1.1). W IRT zakłada się, iż poziom trudności b, nie różnicuje osób już tak jednoznacznie jak w KTT. Dla wartości cechy latentnej Θ zbliżonej do poziomu trudności danej pozycji, istnieje pewne niezerowe prawdopodobieństwo sukcesu, mimo niższego poziomu cechy ($\Theta < b$) oraz porażki mimo poziomu cechy wyższego niż trudność pozycji ($\Theta > b$).



Ryc. 1.1. Rozkład prawdopodobieństwa dla danej pozycji: dla skali idealnej Gutmana (linia niebieska) i dla skali według IRT (linia czerwona-przerywana). Źródło: opracowanie własne.

Modele IRT przyjmują dwa fundamentalne założenia, będące też pewnym ograniczeniem ich zastosowania. Po pierwsze, zakłada się, że **zmienna latentna leżąca u podstaw wyników jest jednowymiarowa**[2], co oznacza, że jedna i właśnie ta zmienna wystarcza do zupełnego wyjaśnienia zmienności obserwowanych wyników. A zatem to, jakie odpowiedzi są wybierane przez osoby uczestniczące w badaniu, ma źródło tylko w jednej ich cesze. Po drugie, poszczególne **pozycje testu lub kwestionariusza są lokalnie niezależne**, co oznacza z kolei niezależność wyników dla

[2] W tej pracy przez modele IRT rozumiane są modele jednowymiarowej zmiennej latentnej. Istnieje jednakże cała rodzina modeli dla wielowymiarowych zmiennych latentnych (oznaczana dla odróżnienia skrótem MIRT; zob. Reckase, 2009), dla których to założenie o jednowymiarowości zmiennej latentnej nie ma zastosowania.

pod-populacji homogenicznych pod względem wartości zmiennej latentnej Θ (Embretson i Riese, 2000, s. 187). Inaczej mówiąc, korelacje między pozycjami testu lub kwestionariusza, jeśli kontrolować wartość zmiennej leżącej u podstaw wyników, będą dążyły do zera.

Oznacza to, że budowanie narzędzia i dobór pozycji do niego w oparciu o IRT, musi uwzględniać stosowanie tylko niezależnych pozycji opisujących jedną zmienną latentną. Warto zaznaczyć, że nie wyklucza to zastosowania modeli IRT do teoretycznie wielowymiarowych konstruktów. Jeśli teoria zakłada kilka zmiennych, w podejściu opartym na IRT mogą być one mimo wszystko traktowane jako jedna zmienna latentna. Z taką sytuacją mamy do czynienia np. w przypadku testu, który ma mierzyć wiedzę i motywację, lub umiejętność czytania i rozumienia. Jeśli wszystkie pozycje testowe mierzyć będą obie zmienne w tej samej proporcji oznaczać to będzie, że matematycznie mierzą one jednowymiarową zmienną latentną.

Zatem zgodnie z założeniami, pierwsze etapy analiz muszą być poświęcone sprawdzeniu jednowymiarowości. Aby potwierdzić obecność jednego głównego czynnika, a tym samym spełnienie pierwszego założenia, używana jest eksploracyjna analiza czynnikowa. Reguła wyznaczania liczby czynników według ich wielkości własnych (*eigenvalue*) powyżej wartości 1 (Kaiser, 1960), może wskazywać na większą ich liczbę, lecz analiza wykresu osypiska (Cattell, 1966; por. Zakrzewska, 1994, s. 65-67) powinna wskazywać na jeden, dominujący czynnik. Statystycznym wskaźnikiem określającym jednowymiarowość zmiennej latentnej jest zaproponowana przez Gessaroli i De Champlain'a (1996) statystyka $X^2_{G/D}$ oparta na transformowanej korelacji reszt między pozycjami testu lub kwestionariusza.

Procedura obliczenia tego wskaźnika jest następująca: macierz reszt przekształcana jest w macierz korelacji reszt, następnie poddawana transformacji Fishera, podnoszona do kwadratu i sumowana poniżej przekątnej macierzy. Wynik ma rozkład statystyki X^2 i wyznaczony jest wzorem (Swaminathan, Hambleton i Rogers, 2007, s. 688):

$$\chi^2_{G/D} = (N-3)\sum_{i=2}^{n}\sum_{j=1}^{i-1} z_{ij}^2 \qquad (1.1a)$$

gdzie:

- z to wartość korelacji po transformacji Fishera,

np.: $z_{ij} = \frac{1}{tanh(r_{ij})}$ lub, $z_{ij} = \frac{1}{2} ln \frac{(1+r_{ij})}{(1-r_{ij})}$ (1.1b)

- i, j to poszczególne pozycje danego narzędzia pomiarowego,
- N jest wielkością próby.

Stopnie swobody dla powyższej statystyki wyznacza się z wzoru:

$$df = \frac{1}{2n}(n-5) \tag{1.1c}$$

gdzie n to liczba pozycji.

Innymi statystycznymi wskaźnikami jednowymiarowości są wskaźniki: 1) T_s oparty na kwadracie proporcji reszt (Maydeu-Olivares, 2001, za: DeMars, 2010, s. 45), charakteryzujący się rozkładem zbliżonym do rozkładu X^2; 2) wskaźnik T Stout'a oparty na sumie kowariancji wyników. Procedura obliczania T Stout'a jest następująca: wyniki grupowane są według wartości Θ i dla każdej z grup wyznaczana jest suma kowariancji między pozycjami analizowanego narzędzia. Jeśli te pozycje nie mają ukrytego drugiego czynnika, to dla osób o tej samej wartości Θ kowariancja między pozycjami powinna być zbliżona do 0 (Stout, 1987, za: DeMars, 2010, s. 43). Dodatkowo pozycje dzieli się na dwie części (w sposób losowy lub na podstawie np. analizy skupień) w celu wykrycia naruszeń założenia jednowymiarowości, które mogłoby być zależne od wybranych pozycji.

Porównując wyżej wymienione wskaźniki Finch i Habing (2007, za: DeMars, 2010, s. 47) wykazali, że dla modeli dwuparametrycznych wszystkie wskaźniki posiadają podobny poziom błędu i rodzaju, czyli skłonności do odrzucenia prawdziwej hipotezy zerowej o braku jednowymiarowości. Lecz dla bardziej skomplikowanych modeli, większych prób oraz dłuższych narzędzi (większej liczby pozycji) wskazane jest używanie wskaźników T_s i T Stout'a, mimo że są one trudniejsze w wyznaczeniu. Jeśli jednak poszczególne pozycje wykluczają zgadywanie odpowiedzi, decyzję można opierać na łatwiejszym do wyznaczenia wskaźniku $X^2_{G/D}$ Gessaroli'ego i De Champlain'a.

Drugie założenie przyjmowane w modelach IRT, dotyczące lokalnej niezależności zmiennych, sprawdza się analizując macierz korelacji cząstkowych między danymi pozycjami. Wartości bliskie zeru potwierdzają niezależność, natomiast wysoki wynik korelacji między parą pozycji zdominuje „definicję" zmiennej latentnej i spowoduje, że dla danych pozycji wartości b będą dużo wyższe, niż dla pozostałych. Należy podkreślić, że pozycje powinny być skorelowane w całej próbie, a jedynie przy kontrolowaniu wartości Θ korelacje te powinny zanikać. Innym sposobem badania niezależności zmiennych jest zaproponowany przez Glas'a i Falcon (2003), w oparciu

o analizy van den Wollenberga, wskaźnik S_{3ij} bazujący na różnicy między częstościami otrzymanymi i przewidywanymi, który jest wyrażony wzorem Swaminathan, Hambleton i Rogers, 2007, s. 690):

$$S_{3ij} = d_{ij}^2 \left\{ \frac{1}{E(N_{ij})} + \frac{1}{E(N_{\bar{i}j})} + \frac{1}{E(N_{i\bar{j}})} + \frac{1}{E(N_{\bar{i}\bar{j}})} \right\} \quad \text{(1.2a)},$$

gdzie:

• d_{ij}^2 to kwadrat różnicy między częstościami otrzymanymi (N_{ij}) i przewidywanymi $E(N_{ij})$,

• N_{ij} to liczba osób rozwiązujących test lub kwestionariusz, które poprawnie odpowiedziały na obie pozycje i oraz j w grupie osób, które otrzymały punktację pomiędzy $k = 2$ a $k = n - 2$.

Przewidywana liczba osób wyznaczona jest wzorem:

$$E(N_{ij}) = \sum_{k=2}^{n-2} N_k P(U_i = 1, U_j = 1 | X = k) \quad \text{(1.2b)},$$

przy czym $E(N_{\bar{i}j})$ dla wzoru 1.2a jest definiowane jak dla wzoru 1.2b dla osób, które odpowiedziały źle na pozycję i, a dobrze na pozycję j, analogicznie $E(N_{i\bar{j}})$ i wreszcie $E(N_{\bar{i}\bar{j}})$ dla osób, które odpowiedziały źle na obie pozycje i oraz j. Statystyka $S3_{ij}$ posiada rozkład zbliżony do rozkładu X^2 dla $df = 1$.

Podobnym, ale bardziej uniwersalnym wskaźnikiem (zarówno dla modeli dwukategorialnych, jak i wielokategorialnych) jest zaproponowany przez Yen'a (1984, za: DeMars, 2010, s. 48) wskaźnik Q_3. Jego obliczenie polega na tym, że po wyznaczeniu parametrów wszystkich pozycji oblicza się reszty dla każdej osoby i każdej pozycji między wynikiem otrzymanym i oczekiwanym. Q_3 jest definiowany jako współczynnik liniowej korelacji między resztami dla poszczególnych pozycji narzędzia. Wartość współczynnika powyżej 0,20 sygnalizuje parę pozycji, dla których nie spełniony jest warunek niezależności.

Najbardziej uniwersalną metodą, która jednocześnie sprawdza oba założenia (o jednowymiarowości i o lokalnej niezależności) jest wyznaczenie i analiza macierzy korelacji lub kowariancji między pozycjami. Powinna być ona wykonywana w grupach homogenicznych ze względu na wartość zmiennej latentnej, tak aby potwierdzić niezależność i jednowymiarowość dla całego jej kontinuum. Jednocześnie warto nadmienić, iż w heterogenicznej populacji, w której występują różne wartości cechy latentnej, odpowiedzi na pozycje będą skorelowane między sobą. Jeżeli oba omawiane

założenia są spełnione, współczynniki kowariancji lub korelacji poza przekątną macierzy będą bardzo niskie (Hambleton, Swaminathan i Rogers, 1991).

1.3. Modele dwukategorialne IRT

Jednym z kryteriów porządkujących modele probabilistyczne jest kryterium liczby możliwych kategorii odpowiedzi na daną pozycję narzędzia pomiarowego. Modele, w których osoba poddawana badaniu może uzyskać w każdym zadaniu (pytaniu, twierdzeniu) jedną z dwóch kategorii wyników (np.: prawda lub fałsz; dobrze lub źle, poprawnie lub niepoprawnie) nazywane są modelami dwukategorialnymi (*dichotomously scored items*).

Jak już zostało zauważone, teoria odpowiadania na pozycje testowe pozwala przyporządkować każdej osobie uczestniczącej w badaniu wynik odpowiadający wartości cechy latentnej Θ (poprzez wartość cechy latentnej rozumie się tu zarówno natężenie cechy badanej, jak i poziom umiejętności), reprezentujący określoną wartość na kontinuum (teoretycznie Θ może przyjmować wartości od -∞ do $+\infty$, lecz przeważnie ogranicza się zakres do przedziału -4;4, gdzie 0 oznacza przeciętną wartość danej cechy). Możliwe jest zatem określenie, z jakim prawdopodobieństwem osoba uczestnicząca w badaniu udzieli prawidłowej odpowiedzi (zgodnej z kluczem) na konkretne pytanie testu lub kwestionariusza dla każdej możliwej wartości cechy Θ. Kształt związku między wartością cechy latentnej a prawdopodobieństwem udzielenia poprawnej odpowiedzi na daną pozycję jest w IRT opisany krzywą (najczęściej) logistyczną – charakterystyczną dla każdej pozycji (ICC – *Item Characteristic Curve* – ryc. 1.2).

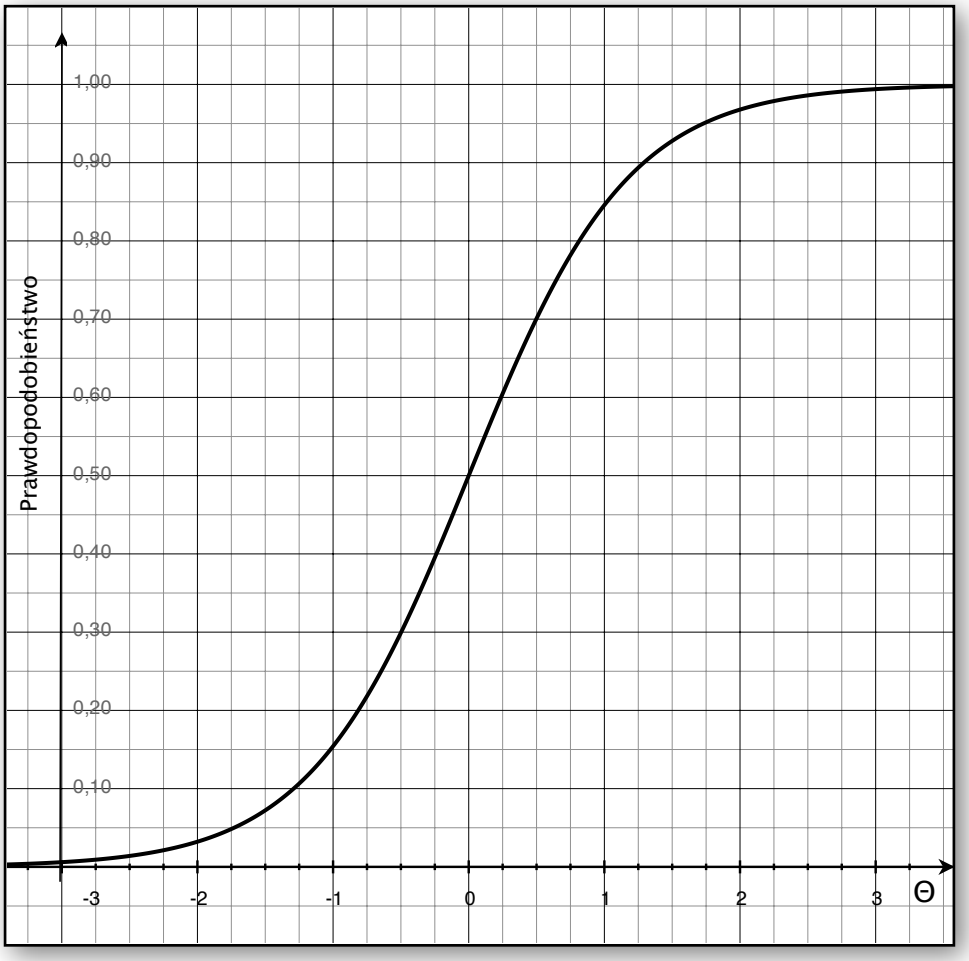
Model logistyczny dwukategorialny występuje w czterech wariantach w zależności od liczby uwzględnionych parametrów i jest opisywany funkcją w postaci:

$$p_i(\Theta) = \frac{1 - c_i - d_i}{1 + e^{-(Da_i(\theta - b_i))}} \qquad (1.3),$$

gdzie:

- $p_i(\Theta)$ – prawdopodobieństwo posiadania umiejętności, wiedzy, itp. na danym poziomie Θ,
- a_i – moc różnicująca danej pozycji,
- b_i – poziom trudności pozycji,
- c_i – parametr zgadywania,
- d_i – parametr niedbałości,

- e – podstawa logarytmu naturalnego – stała matematyczna,
- D – stały parametr maksymalizujący dopasowanie krzywej logistycznej do ogiwy rozkładu normalnego, która była pierwotnie używana w IRT. Podstawienie za D = 1,702 powoduje, że używana skala jest metryczna, dla D = 1 – skala jest logistyczna (Drasgow i Hulin, 1990; Baker, 2001). Zostaną teraz opisane znaczenia poszczególnych parametrów występujących w funkcji logitowej (por ryc 1.4 A – D).



Rycina 1.2. Związek między prawdopodobieństwem udzielenia odpowiedzi (oś y) a poziomem umiejętności lub natężeniem cechy (oś x) w teorii odpowiadania na pozycje testowe. Źródło: opracowanie własne.

Zakładając, że osoba uczestnicząca w badaniu posiada pewną wartość cechy latentnej – **poziom trudności** danej pozycji oznaczony przez b_i jest wartością cechy latentnej Θ, dla której prawdopodobieństwo prawidłowej odpowiedzi na daną pozycję jest równe 0,5, czyli medianie wartości cechy latentnej dla danej pozycji. Pozycja jest tym trudniejsza, im wyższą wartość przyjmuje parametr b, a to z kolei informuje o tym, w którym miejscu na skali Θ dana pozycja najlepiej różnicuje osoby badane. W odróżnieniu od pomiaru umiejętności, czy wiedzy, gdzie istnieją odpowiedzi

prawdziwe i fałszywe – parametr trudności dla pomiaru cechy, jak w przypadku temperamentu, osobowości itp., odpowiada miejscu osoby uczestniczącej w badaniu na kontinuum tejże cechy w ujęciu Goldberga (1972). Inaczej mówiąc, parametr b odpowiada wtedy natężeniu cechy, dla którego prawdopodobieństwo wybrania danej odpowiedzi wynosi dokładnie 0,5. Łatwe pozycje dobrze różnicują osoby o niskich umiejętnościach, czy też niskim poziomie danej cechy, zaś trudne pozycje – osoby o wysokich umiejętnościach, wysokim poziomie cechy. Parametr b przeważnie zawiera się w przedziale: $<-2; 2>$ (Hulin, Drasgow i Parsons, 1983).

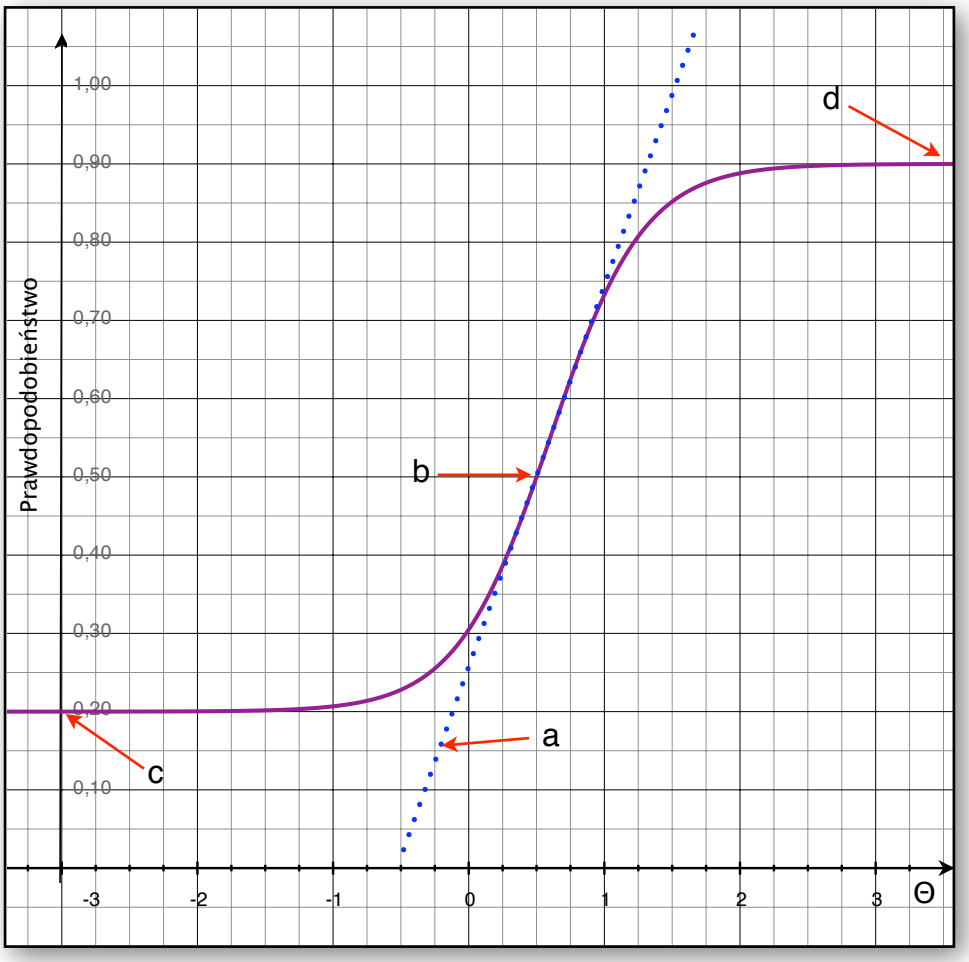
Moc różnicująca danej pozycji oznaczana za pomocą a_i jest proporcjonalna do kąta nachylenia krzywej charakterystycznej w punkcie przegięcia. Informuje ona, w jakim stopniu pozycja pozwala podzielić badanych na tych o poziomie cechy niższym i wyższym od poziomu trudności pozycji wyznaczonej przez parametr b. Moc różnicująca jest tym większa, im wyższą wartość przyjmuje parametr a – przeważnie przyjmuje on wartości z przedziału: $<0; 2>$, choć teoretycznie jest zdefiniowany w przedziale: $<-\infty; +\infty>$ (Hambelton i Swaminathan, 1985).

Parametr **zgadywania** oznaczony przez c obrazuje szansę na powodzenie osoby, która wskazała prawidłową odpowiedź, choć nie wynika to z jej wiedzy, lecz jest dziełem przypadku. Innymi słowy, c informuje, jakie jest prawdopodobieństwo uzyskania prawidłowej odpowiedzi dzięki zgadywaniu. Parametr ten ma stałą wartość, co oznacza, że osoby uczestniczące w badaniach posiadają tę samą szansę uzyskania niezerowego wyniku niezależnie od posiadanej przez nich wartości zmiennej latentnej Θ. Np. dla pozycji z dwiema odpowiedziami do wyboru współczynnik ten będzie miał wartość 0,5, dla czterech możliwych odpowiedzi, w tym jednej poprawnej 0,25; z kolei dla pozycji mającej formę pytania otwartego będzie wynosił 0. Należy nadmienić, iż im wyższy poziom zgadywalności, tym mniejsza efektywność danej pozycji wyrażona w jego mocy różnicującej (Reise i Waller, 2003).

Parametr **niedbałości** oznaczony przez d, będący niejako odwrotnością poprzedniego parametru, uwzględnia niepoprawne odpowiedzi udzielone przez osobę, którą cechuje wysoka wartość zmiennej latentnej. Odpowiedzi te wynikają na ogół z przypadkowych czynników. Inaczej mówiąc, jest to parametr określający błędy w odpowiedziach nie związane z poziomem wiedzy.

Rozważmy przykład, który pozwoli zilustrować graficznie znaczenie wymienionych wyżej parametrów. Przyjmując przykładowe wartości dla teoretycznej pozycji testowej lub kwestionariuszowej: trudność b = 0,6; moc różnicująca a = 1,7; parametr

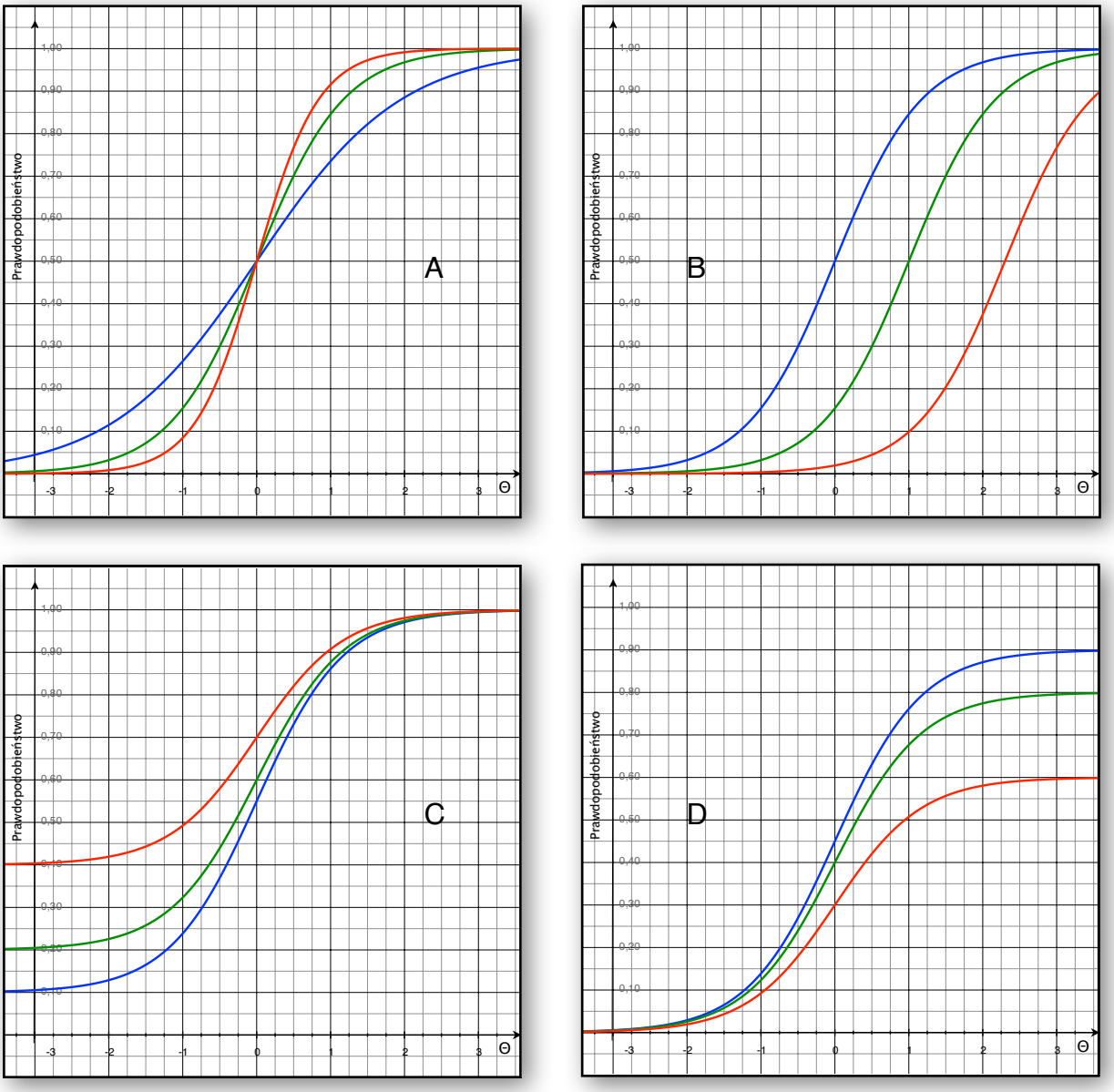
zgadywalności c = 0,2; parametr niedbałości d = 0,1 – otrzymamy krzywą logistyczną taką jak na rycinie 1.3.



Rycina 1.3. Krzywa logistyczna dla przykładowych wartości parametrów. Źródło: opracowanie własne.

Drugim z kryteriów podziału modeli logistycznych jest liczba przyjętych parametrów. Pełny model czteroparametryczny (według wzoru nr 1.3) jest bardzo rzadko stosowany, ponieważ wyznaczenie wszystkich parametrów wymaga bardzo dużych prób (powyżej 1000 osób – Hulin, Lissak i Drasgow, 1982).

Najczęściej przyjmuje się model trójparametryczny, zakładający zerowy poziom niedbałości d_i. *W* sytuacji, gdy można także założyć zerowy lub bliski zeru poziom zgadywalności c_i, (pozycje narzędzia pomiarowego są w postaci np. pytań otwartych, bez odpowiedzi do wyboru), otrzymuje się model dwuparametryczny.



Ryc 1.4. Ilustracja graficzna znaczenia parametrów modelu logistycznego.
A. Zwiększenie mocy różnicującej powoduje zwiększenie kąta nachylenia krzywej logistycznej, a = {0,6; 1; 1,4}. B. Zwiększenie trudności przesuwa w prawo krzywą logistyczną wzdłuż osi Θ, b = {0; 1; 2,3}. C. Zwiększenie parametru zgadywalności podnosi dolną asymptotę krzywej logistycznej, c = {0,1; 0,2; 0,4}. D. Zwiększenie parametru niedbałości obniża górną asymptotę krzywej logistycznej, d = {0,1; 0,2; 0,4}. Źródło: opracowanie własne.

Szczególnym przypadkiem jest przyjęcie założenia o jednakowej mocy różnicującej a_i dla wszystkich pozycji. Mamy wtedy do czynienia z logistycznym modelem jedno-parametrycznym, a w przypadku przyjęcia dla wszystkich pozycji parametru $a_i = 1$ model taki nazywany jest modelem Rascha od nazwiska duńskiego matematyka, który w 1960 roku opublikował pracę na gruncie teorii prawdopodobieństwa o danych uzyskiwanych w testach (Baker, 1987).

1.4. Modele wielokategorialne IRT

Zdarza się, że modele dwukategorialne są niewystarczające do opisu wyników. Dzieje się tak, gdy osoba poddawana badaniu w odpowiedzi na każdą pozycję narzędzia badawczego może wskazać jedną z kilku uporządkowanych kategorii odpowiedzi (*polytomously scored items*), np.: od „zupełnie mnie to nie dotyczy", przez „częściowo mnie to nie dotyczy", „trochę mnie to dotyczy", do „całkowicie mnie to dotyczy". W 1969 roku Samejima artykułem pt.: „*Estimation of latent ability using a response pattern of graded scores*" zapoczątkowała prace nad rodziną modeli wielokategorialnych. Dla stopniowanych pozycji testu lub kwestionariusza najczęściej mają zastosowanie dwa modele: uogólniony model punktów częściowych (GPCM – *Generalised Partial Credit Model* – Muraki, 1992; jako rozszerzenie modelu jednoparametrycznego PCM Mastersa, 1982) lub model klasy odpowiedzi (GRM – *Graded Response Model* – Samejima, 1969). Dla każdej pozycji testu lub kwestionariusza prawdopodobieństwo, że osoba osiągnie j-tą kategorię wzrasta wraz ze wzrostem natężenia cechy Θ do momentu, gdy prawdopodobieństwo osiągnięcia j+1 kategorii staje się wyższe, a prawdopodobieństwo osiągnięcia j-tej kategorii zaczyna maleć (por. ryc. 1.5).

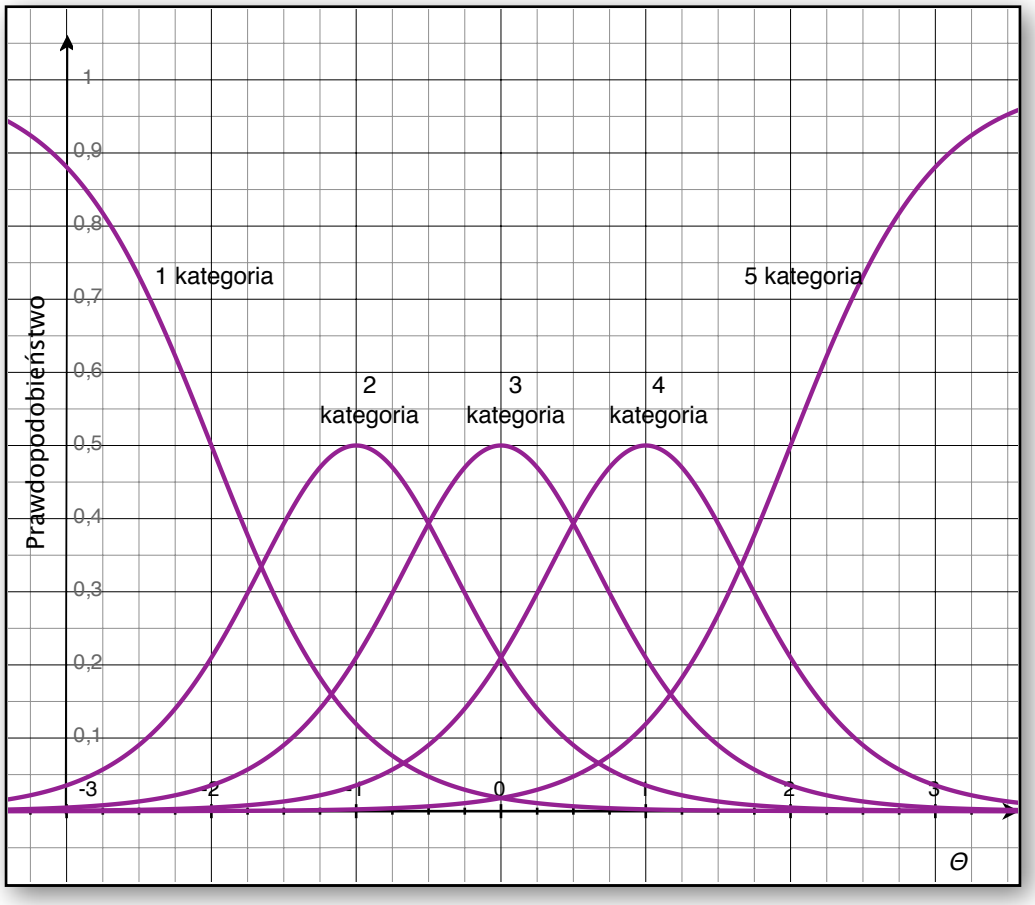
Model GRM zakłada, iż prawdopodobieństwo uzyskania odpowiedzi spośród j kategorii w danej pozycji kwestionariuszowej jest tożsame z prawdopodobieństwem uzyskania tej odpowiedzi lub odpowiedzi niższej. Pozwala to łączyć odpowiedzi w klasy i rozpatrywać za pomocą aparatu matematycznego modeli dwukategorialnych.

Model GPCM zakłada z kolei, że każda możliwa odpowiedź może być rozpatrywana z pewnym osobnym prawdopodobieństwem (patrz ryc. 1.5). Przy takim założeniu prawdopodobieństwo uzyskania odpowiedzi dla danej wartości Θ da się wyrazić wzorem (DeMars, 2010, s. 26):

$$p_i(\Theta) = \frac{e^{\left(\sum_{j=0}^{x_i}(\theta - b_{x_{ij}})\right)}}{\sum_{k=0}^{m_i} e^{\left(\sum_{j=0}^{k}(\theta - b_{x_{ij}})\right)}} \qquad (1.4),$$

gdzie:

- $b_{x_{ij}}$ jest parametrem trudności j-tej kategorii pozycji x_i,
- każda pozycja i ma m_i uporządkowanych możliwych odpowiedzi.



Rycina 1.5. Rozkład prawdopodobieństwa wskazania danej kategorii w zależności od wartości Θ w zadaniu z 5 kategoriami odpowiedzi. Moc różnicująca a wszystkich kategorii jest taka sama. Źródło: opracowanie własne.

Różnica między modelami GRM i GPCM polega głównie na sposobie obliczania parametrów pozycji testowych. Modele GRM kategoryzuje się często jako modele pośrednie (*indirect*), a GPCM jako bezpośrednie (*direct*), ponieważ proces obliczania prawdopodobieństwa poziomu Θ jest, w zależności od przyjętego modelu, dwu- lub jednostopniowy (Embretson i Reise, 2000).

W GRM dla każdej j-tej kategorii odpowiedzi najpierw wyznaczana jest krzywa charakterystyczna przy stałym poziomie mocy różnicującej a_i, gdzie parametr b_{ij} odpowiada wartości cechy latentnej Θ, dla której prawdopodobieństwo osiągnięcia j-tej kategorii wynosi 0,5. W drugim kroku dla poszczególnych kategorii obliczane jest prawdopodobieństwo uzyskania wyniku powyżej poziomu b i sumowane dla całej pozycji narzędzia badawczego. W GPCM prawdopodobieństwo pozycji obliczane jest w pojedynczym kroku według wzoru 1.4 i jest sumą prawdopodobieństw dla każdej z sąsiednich kategorii odpowiedzi.

Oba modele wielokategorialne stanowią uogólnienie modeli dwukategorialnych, które mogą być rozpatrywane jako szczególne przypadki GRM lub GPCM (De Ayala, Dodd i Koch, 1992).

1.5. Rzetelność i błąd standardowy pomiaru

W KTT rzetelność jest zdefiniowana jako kwadrat korelacji między wynikami prawdziwymi a wynikami otrzymanymi, czyli jako stosunek wariancji wyników prawdziwych do wariancji wyników otrzymanych. W tym kontekście błąd standardowy pomiaru (SEM) jest średnią wielkością różnic między wynikiem prawdziwym a wynikami otrzymanymi w nieskończenie wielu próbach. Przy wyznaczonym z próby odchyleniu standardowym (s_x) i znanej rzetelności narzędzia pomiarowego (r_{tt}), błąd standardowy pomiaru wyrażony jest wzorem:

$$SEM = S_X\sqrt{1 - r_{tt}} \tag{1.5}.$$

W IRT natomiast rzetelność wyznacza się w oparciu o funkcję informacyjną (Baker, 2001). Pojęcie informacji jest związane z błędem oszacowania i wyrażone jest wzorem:

$$I = \frac{1}{SEM(\theta)^2} \tag{1.6}.$$

Dla KTT zmiana pozycji testu lub kwestionariusza pociąga za sobą lawinowo konieczność wyznaczenia na nowo rzetelności całego narzędzia, podobnie dodanie lub usunięcie obserwacji w próbie. Wynika to z założenia KTT, według którego wynik uzyskany w teście jest sumą wyniku prawdziwego i błędu pomiarowego, a te właściwości są wyznaczane na określonej próbie badanej danym narzędziem. Co więcej, wyniki poszczególnych pozycji nie są bezpośrednio powiązane z cechą latentną (Hornowska, 2001, s. 201). W IRT usunięcie z próby wyniku osoby uczestniczącej w badaniu lub pozycji narzędzia badawczego pozwala szybko wyznaczyć nową wartość informacji (i błędu standardowego pomiaru) poprzez zwykłe odejmowanie, ponieważ każda pozycja jest powiązana z wartością cechy latentnej (Θ).

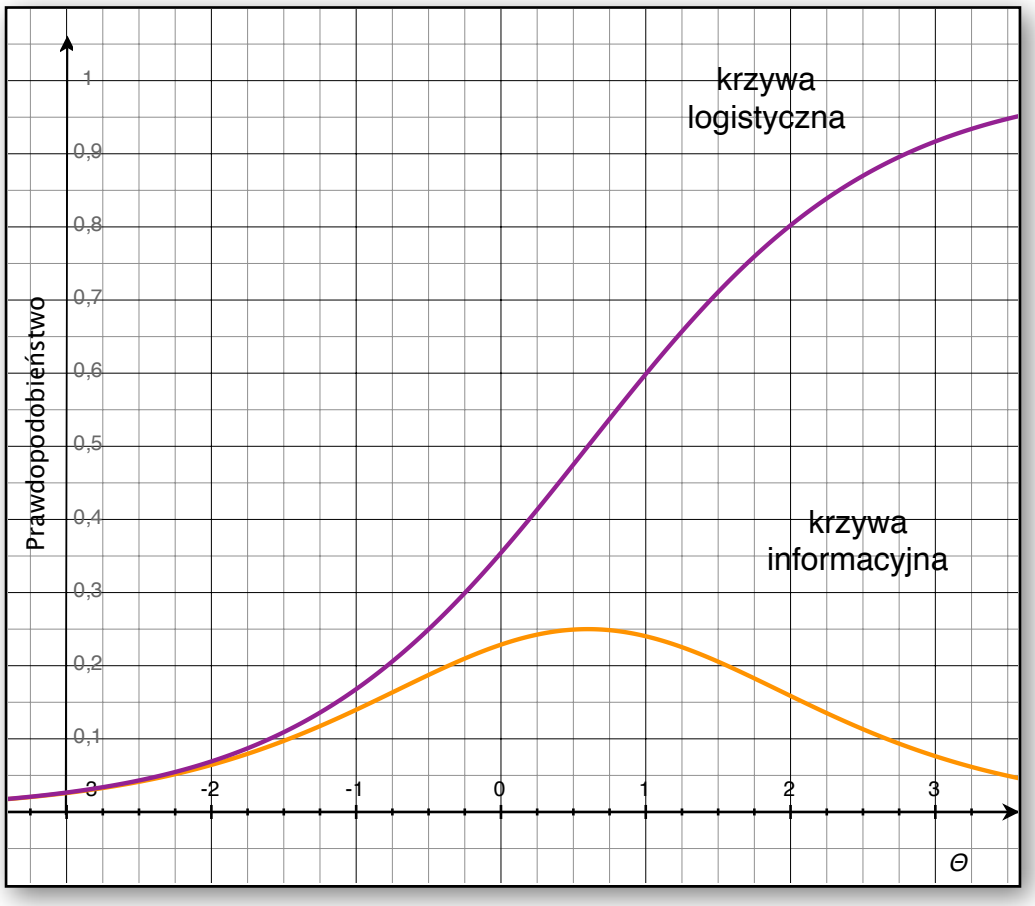
Dla modeli dwukategorialnych IRT najwięcej informacji o szacowanym wyniku Θ dostarczane jest przez daną pozycję wtedy, gdy prawdopodobieństwo odpowiedzi kodowanej jako „sukces” wynosi 0,5. Gdy prawdopodobieństwo jest bliskie 0 lub 1, wielkość informacji jaką możemy uzyskać bliska jest 0. Wzór na poziom informacji dla danej pozycji ma następującą postać:

$$I_i(\theta) = a_i^2 f_i(\theta)(1 - f_i(\theta)) \tag{1.7},$$

gdzie:

- f_i jest funkcją odpowiedzi na i-tą pozycję zgodną z przyjętym modelem według wzoru (1.3),
- a_i jest miarą mocy różnicującej dla i-tej pozycji.

Poniżej znajduje się wykres obrazujący relację między informacją a prawdopodobieństwem udzielenia odpowiedzi w zależności od wielkości cechy latentnej (ryc. 1.6).

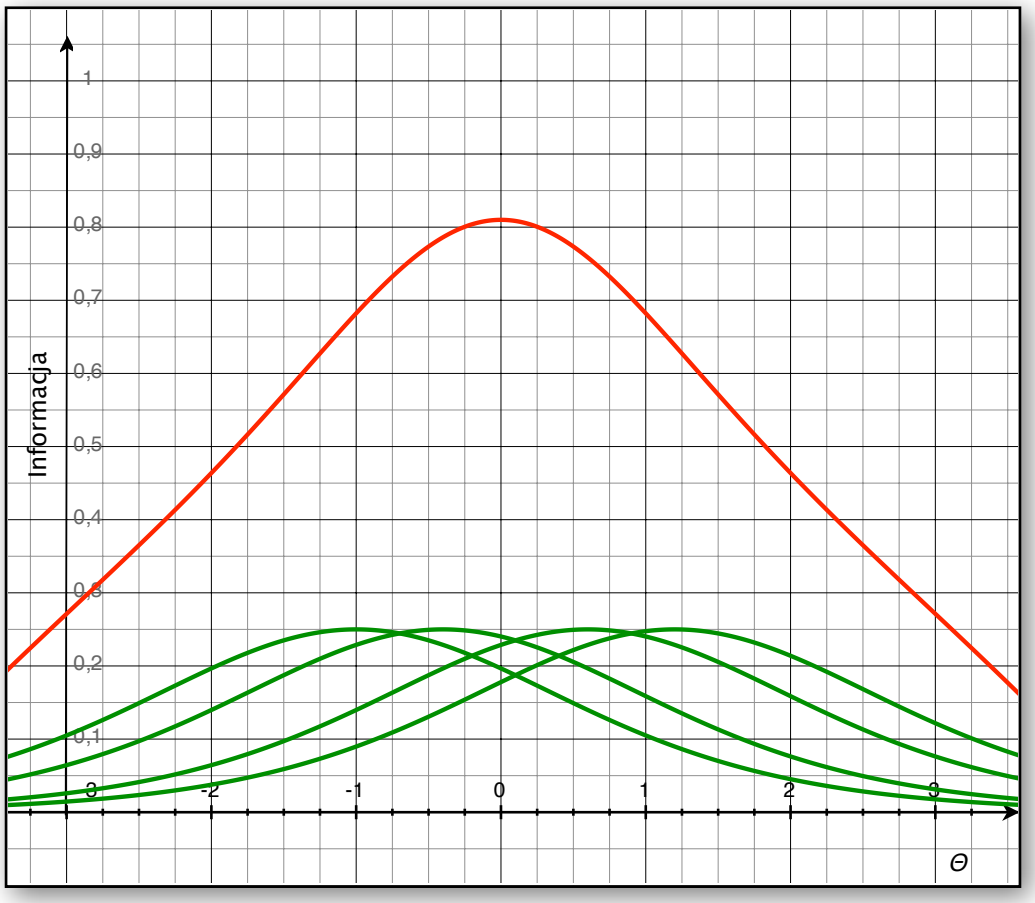


Rycina 1.6. Wykres krzywej informacyjnej (pomarańczowa) i krzywej logistycznej (fioletowa) danej pozycji dla przyjętych parametrów: a = 1, b = 0,6, c = 0, d = 0. Źródło: opracowanie własne.

Ze względu na to, że SEM definiuje się jako odwrotność pierwiastka kwadratowego z poziomu informacji, tam gdzie jest on wyższy – mniejszy jest błąd standardowy, a większa jest rzetelność. Z kolei z uwagi na fakt, że w podejściu IRT zakłada się statystyczną niezależność odpowiedzi na poszczególne pozycje, to funkcje informacyjne można dodawać do siebie. Funkcja informacyjna całego narzędzia może być zatem wyrażona wzorem:

$$I_{testu}(\theta) = \sum_{i=1}^{k} [a_i^2 f_i(\theta)(1 - f_i(\theta))] \quad (1.8).$$

Można to też przedstawić graficznie jak na przykładowym wykresie poniżej (ryc. 1.7):



Rycina 1.7. Funkcje informacyjne dla skali składającej się z 4 hipotetycznych zadań (zielone linie: a = 1; c = d = 0; b_1 = -1; b_2 = -0,4; b_3 = 0,6; b_4 = 1,2) oraz ich suma (czerwona linia). Źródło: opracowanie własne.

W modelach wielokategorialnych wyznaczenie funkcji informacji w zależności od Θ jest bardziej skomplikowane matematycznie ze względu na to, że dla każdej i-tej pozycji istnieje tyle krzywych logistycznych, ile dana pozycja zawiera opcji odpowiedzi (por. ryc. 1.5). Wielkość informacji jest wyznaczana poprzez kwadrat współczynnika nachylenia krzywej regresji wyników pozycji testowych lub kwestionariuszowych dla danej wartości Θ, podzielonej przez ich wariancję (Reckase, 2009, s. 47). Należy zauważyć, że dla modeli wielokategorialnych funkcja informacji nie jest monotoniczna i wyrażona jest wzorem:

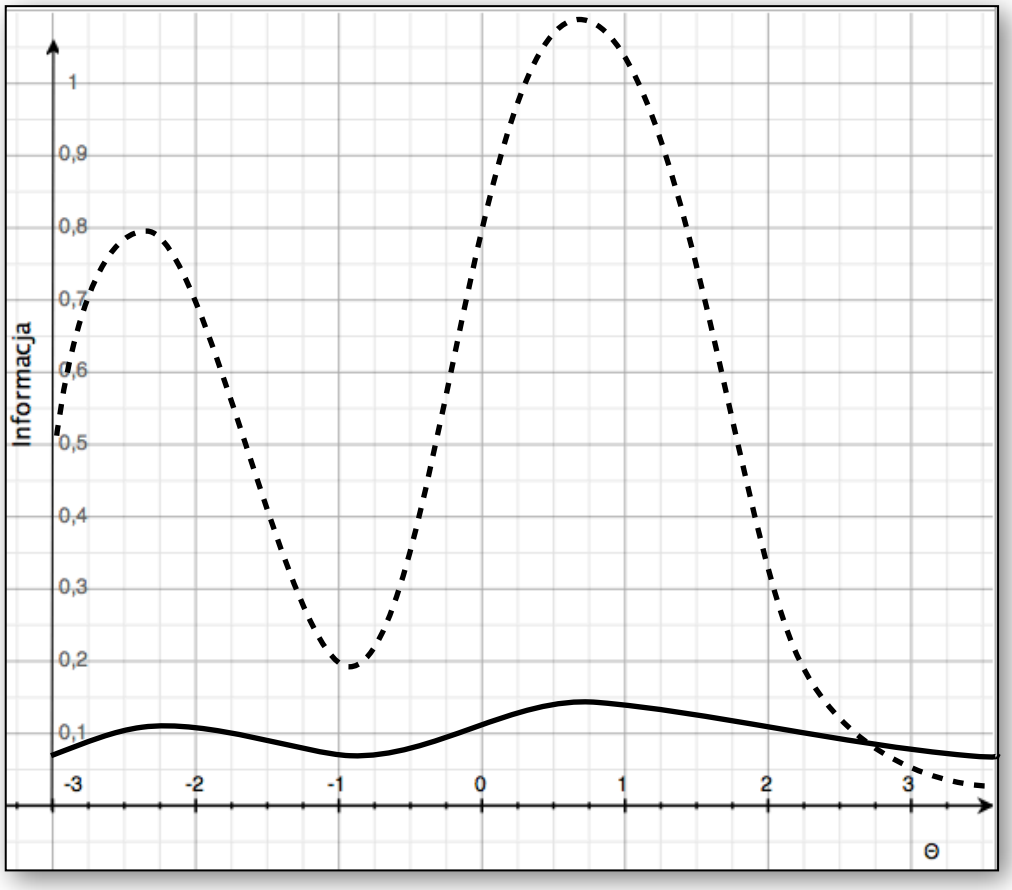
$$I(\theta_j, u_i) = \sum_{k=1}^{k+1} \frac{[Da_i P_{i,k-1}(\theta_j) Q_{i,k-1}(\theta_j) - Da_i P_{ik}(\theta_j) Q_{ik}(\theta_j)]^2}{P_{i,k-1}(\theta_j) - P_{ik}(\theta_j)}$$

gdzie:

- u_i jest wynikiem i-tej pozycji,

- $P_{ik}(\Theta_j)$ jest prawdopodobieństwem j-tej odpowiedzi w k-tej kategorii w i-tej pozycji,
- $Q = 1\text{-}P$.

Przykładowe wykresy funkcji informacyjnej dla wielokategorialnych pozycji przedstawiono na rycinie 1.8. Krzywe nie różnią się miejscami przegięcia, ponieważ to wynika z takich samych wartości b, natomiast różne wartości a wpływają na amplitudę poszczególnych krzywych informacyjnych.



Rycina 1.8. Krzywe funkcji informacyjnych dla dwóch hipotetycznych wielokategorialnych pozycji o parametrach b1 = -2, b2 = -1, b3 = -0,5, b4 = 0, b5 = 3 oraz a = 0,1 (linia ciągła) i a = 1,2 (linia przerywana). Źródło: opracowanie własne.

Powyższe wykresy funkcji obrazują wielkość wpływu parametru a na ilość informacji oraz wskazują na potencjalny problem przy wyborze pozycji o maksymalnym poziomie informacji przy estymowanej wartości Θ. Dobranie poszczególnych pozycji tak, aby uzyskać określony kształt funkcji informacyjnej dla całego narzędzia, jest zabiegiem bardziej skomplikowanym, niż dla modeli dwukategorialnych.

W literaturze przedmiotu można wyróżnić trzy podejścia obliczeniowe dla funkcji informacyjnej w modelach wielokategorialnych: 1) poprzez wyznaczanie pochodnych, 2) przez wykorzystanie warunkowych założeń lub 3) w oparciu o elementy składowe[3].

[3] Pełny opis tych metod wykracza poza zakres niniejszej pracy a można się z nim zapoznać w opracowaniu Ostini i Nering (2006) lub Reckase (2009).

Bez względu na zastosowany aparat matematyczny efektem obliczeń jest (analogicznie jak w modelach dwukategorialnych) krzywa funkcji informacyjnej, która reprezentuje najniższy poziom błędu SEM popełniany przy wyznaczaniu wartości Θ.

Niezależność pozycji i możliwość sumowania funkcji informacyjnej jest bardzo użyteczną cechą, ponieważ pozwala łączyć pozycje testowe lub kwestionariuszowe w różne zestawy dobierając krzywą informacyjną do konkretnych zastosowań konstruowanego narzędzia – np. do celów selekcyjnych można wybierać te pozycje, które mają wysoką wartość informacyjną dla przyjętej jako kryterium wartości Θ, zaś przy kwestionariuszach, których celem jest zbadanie różnic między osobami – wybrane pozycje powinny w sumie tworzyć w miarę płaską krzywą informacyjną.

1.6. Ewaluacja modeli opartych na IRT

Zastosowanie podejścia probabilistycznego do szacowania wartości zmiennej latentnej ma trzy podstawowe zalety. Po pierwsze, w modelach IRT wyniki zawsze są wyrażone na skali metrycznej jako miary wartości Θ. Jest to duża zaleta w stosunku do podejścia klasycznego, gdzie poziom pomiarowy w badaniach psychologicznych rzadko wykracza poza skalę porządkową. Po drugie, modele IRT pozwalają precyzyjniej szacować błąd standardowy pomiaru (SEM). Wyniki skrajne (niskie i wysokie) są obciążone większym błędem niż wyniki ze środka zakresu, czyli błąd standardowy pomiaru różni się w obrębie skali Θ, inaczej niż w KTT, gdzie zakładany jest jego stały poziom dla całego pomiaru. Przyjęcie zmienności SEM pozwala dokładniej określać rzetelność narzędzia badawczego, a także sporządzać różne zestawy z dostępnych pozycji, maksymalizując precyzyjność testu lub kwestionariusza dla zadanej wartości Θ. Trzecią zaletą modeli opartych na IRT jest obiektywność. Dysponując skalibrowanymi pozycjami, dostosowuje się narzędzie do potrzeb danej sytuacji badawczej oraz danej osoby uczestniczącej w badaniu. Nie traci się przy tym możliwości porównania otrzymanego wyniku z innymi. Nie ma znaczenia, w jakiej grupie znalazła się dana osoba badana ani z jakich pozycji składało się narzędzie pomiarowe – jej wynik jest obiektywny w stosunku do wyników wszystkich osób badanych narzędziem składającym się z pozycji dobranych z tego samego zbioru.

1.6.1. Dopasowanie modelu IRT do formatu pozycji testowych

Wybierając model, w oparciu o który będą szacowane parametry pozycji, należy znać odpowiedź na pytanie o liczbę kategorii odpowiedzi i ich rodzaj. Dla testów lub kwestionariuszy z pozycjami, na które są tylko dwie kategorie odpowiedzi: tak / nie,

prawda / fałsz itp., odpowiednie są modele dwukategorialne (*dichotomous models*): jedno-, dwu-, trzy- lub cztero- parametryczne (oznaczane w literaturze odpowiednio: 1, 2, 3 i 4PL). Jeśli można przyjąć dodatkowe założenie, iż poszczególne pozycje będą miały taką samą moc różnicującą (parametr a), można skorzystać z rodziny modeli Rasch'a (Rasch, 1960). Główną zaletą modeli Rasch'a jest ich prostota, ale rzadko mamy do czynienia z taką sytuacją, iż wszystkie pozycje charakteryzują się identycznym poziomem mocy różnicującej. Szacowanie różnych wartości parametrów dla każdej z pozycji lepiej oddaje rzeczywiste funkcjonowanie pozycji wchodzących w skład narzędzia badawczego. Z tego powodu model 3PL jest najczęściej wybierany w sytuacji badania za pomocą narzędzia składającego się z pozycji dwukategorialnych . W przypadku używania testów wiedzy wydaje się słuszne założenie pewnego niezerowego prawdopodobieństwa dla możliwości odgadnięcia przez osoby uczestniczące w badaniu prawidłowej odpowiedzi. Z kolei dla pozycji z kilkoma (> 2) uporządkowanymi kategoriami odpowiedzi (np.: tak / raczej tak / raczej nie / nie), najczęściej stosowane są wspomniane już modele wielokategorialne (*polytomous models*) np.: PCM – *Partial Credit Model* i jego rozszerzenie GPCM lub GRM. I wreszcie dla wielokrotnych odpowiedzi z nieokreślonym porządkiem, najbardziej odpowiednie są modele nominalne (*nominal polytomous*), zaś dla odpowiedzi w postaci rankingów odpowiednie są modele rankingowe (*rankings*) (Bock i Moustaki, 2007). Dla kwestionariuszy psychologicznych, które najczęściej opierają się na skali porządkowej, wielokategorialnej stosuje się przeważnie właśnie modele GPCM lub GRM.

1.6.2. Kryteria dopasowania modelu IRT do danych

Użyteczność wyników uzyskanych w analizach opartych o IRT zależy od stopnia, w jakim wybrany model odzwierciedla rzeczywiste dane. Ocena dobroci dopasowania przyjętego modelu do otrzymanych danych, polega głównie na sprawdzeniu rozkładu różnic między wynikami otrzymanymi a przewidzianymi przez przyjęty model. Dla modeli 1PL i modeli Rasch'a konstrukcja takiego wskaźnika prawdopodobieństwa LR (*likelihood ratio*) jest stosunkowo prosta. Obliczany jest on na podstawie dostępnej obserwacji proporcji osób z odpowiedziami zgodnymi i niezgodnymi z kluczem.

$$LR = \frac{1-\beta}{\alpha} \qquad (1.9),$$

gdzie α to wielkość błędu pierwszego rodzaju, β to wielkość błędu drugiego rodzaju. Stąd wskaźnik dopasowania *D* wyrażony jest wzorem 1.10:

$$D = -2ln\left(\frac{LR_{obs}}{LR_{obl}}\right) \quad (1.10),$$

gdzie istotność może być wyznaczona w oparciu o rozkład statystyki X^2 *ze stopniami swobody* df równymi liczbie parametrów dla przyjętego modelu.

Dla modeli bardziej złożonych, które szacują latentną wartość Θ, a więc wielkość z definicji nie podlegającą obserwacji, konstrukcja wskaźnika dobroci dopasowania jest trudniejsza (Rost i Davier, 1994; Glas, 1988; Wright i Mead, 1977; Wright i Panchapakesan, 1969). Najszerzej przyjętym sposobem (DeMars, 2010, s. 235) jest wykreślenie krzywych odpowiedzi (*item response curve*) według przyjętego modelu i wyznaczonych parametrów, a następnie porównanie ich z krzywymi dla otrzymanych odpowiedzi. Procedura wyznaczania krzywej dla otrzymanych odpowiedzi wygląda w ten sposób, iż po wyznaczeniu parametrów według przyjętego modelu i obliczeniu wartości zmiennej latentnej dla osób badanych sortuje się ich wyniki Θ i wyznacza g równolicznych grup. Następnie oblicza się dla każdej pozycji procent zgodnych odpowiedzi w obrębie każdej z g grup. Na podstawie mediany wartości Θ wewnątrz grupy (oś Y) oraz procentu zgodnych odpowiedzi (oś X) wyznacza się krzywą odpowiedzi. Różnice między obiema krzywymi: otrzymaną i wyznaczoną dla obliczonych parametrów, mogą być wskazówką występowania następujących problemów:

- niespełnienia założenia o jednowymiarowości zmiennej latentej,
- złego dopasowania modelu do danych,
- braku monotoniczności funkcji $f(\Theta) = p$,
- wrażliwości na próbę (wysoki wskaźnik DIF – *differential item functioning*),
- słabego różnicowania poszczególnych pozycji.

Jako statystyczny wskaźnik dopasowania wykorzystywany jest współczynnik Pearsona χ^2 (Swaminathan, Hambleton i Rogers, 2007, s. 699):

$$\chi^2 = \sum_{j=1}^{J}\sum_{k=1}^{K} N_{jk}\frac{(O_{jk} - E_{jk})^2}{E_{jk}} \quad (1.11a),$$

gdzie:

- k jest liczbą kategorii odpowiedzi,
- j jest indeksem pozycji testowych lub kwestionariuszowych,

- O_{jk} jest otrzymanym prawdopodobieństwem zgodnych pozycji dla pozycji j*-tej*,
- E_{jk} jest oczekiwanym prawdopodobieństwem zgodnych pozycji na podstawie wielkości parametrów obliczonych w oparciu o założony model IRT, dla mediany wartości Θ,
- N_{jk} jest liczbą obserwacji dla pozycji j-tej, w kategorii k*-tej*.

Wskaźnik ten ma rozkład zbliżony do rozkładu χ^2 ze stopniami swobody $df = j - k$,.

Dla modeli dwukategorialnych ($k = 1$) wzór ulega redukcji do postaci:

$$\chi^2 = \sum_{j=1}^{J} N_j \frac{(O_j - E_j)^2}{E_j(1 - E_j)}, \qquad \text{(1.11b).}$$

Wskaźnik ten podobnie jak inne oparte na dystrybuancie χ^2 jest wrażliwy na wielkość próby (Bock i Moustaki, 2007, Mair, Reise i Bentler, 2008). Z tego powodu we współczesnych programach komputerowych implementowane są bardziej zaawansowane algorytmy odpowiednie dla dużych prób i narzędzi badawczych z liczbą pozycji powyżej 20 (np. BILOG-MG – Zimowski, Muraki, Mislevy i Bock, 1996).

Porównując dopasowanie kilku modeli do danych, można zastosować procedurę zaproponowaną przez (Maydue-Olivares, Drasgow i Mead (1994) nazywaną „idealny obserwator". Polega ona na wyznaczeniu współczynnika prawdopodobieństwa LR między danymi oszacowanymi przez porównywane modele. Analizując wskaźnik dopasowania D dla każdego z modeli i otrzymując zbliżone wartości możemy założyć, że między modelami nie istnieją znaczące różnice wpływające na dokładność oszacowania. Zatem wybór każdego z modeli pozwoli otrzymać podobnie dopasowane dane.

Powyższa ocena modelu z punktu widzenia dopasowania do pozycji (*item fit*) może być uzupełniona o ocenę dopasowania odpowiedzi do osób badanych (*person fit*) (Swaminathan, Hambleton i Rogers, 2007; Emons, Sijtsma i Meijer, 2005). Celem tego rodzaju analizy jest zbadanie, czy istnieje taki zestaw odpowiedzi na poszczególne pozycje, który mógłby wskazywać na wpływ innej zmiennej latentnej, niż tej leżącej u podstaw narzędzia. Jako zestaw odpowiedzi używane są wszystkie możliwe wektory odpowiedzi, i tak dla zestawu liczącego 8 czterokategorialnych pozycji przykładowy wektor zakodowanych odpowiedzi mógłby wyglądać następująco: {4 4 3 4 4 1 2 1}. Czasami analizy tego rodzaju przeprowadza się aby wykryć oszustwa dokonywane

przez osoby badane podczas egzaminu, choć wynik analiz nie dowodzi tego wprost, a jedynie wskazuje na istnienie takiej możliwości. Podstawowym wskaźnikiem jest indeks ZL opracowany przez Drasgow'a, Levine'a i McLaughlina (1991). Jego wartości bliskie 0 wskazują na dobre dopasowanie danych do modelu z wyznaczonymi parametrami, wartości ujemne – na występowanie takich układów odpowiedzi, które nie są dopasowane do modelu, a wartości dodatnie – na istnienie takich odpowiedzi, które mają wyższe prawdopodobieństwo występowania, niż przewiduje model (por. Karabatsos, 2003).

Podsumowując: IRT może być bardzo użytecznym narzędziem do rozwoju, doskonalenia i oceny skal, wykraczającym poza możliwości analiz w KTT (takich jak analiza czynnikowa, regresyjna, czy analiza rzetelności), ale ich nie zastępującym. Wnioski wyciągane z obu rodzajów analiz powinny być spójne, a ich rozbieżność wskazuje np. na zły dobór modelu IRT lub nieuprawnione założenie o ciągłości mierzonej zmiennej. Warto także pamiętać, iż konstruując przydatne narzędzie należy zadbać nie tylko o zgodność wyników z punktu widzenia statystycznego, ale też o poziom znaczenia klinicznego, tak aby analizy miały wartość praktyczną.

1.7. Tworzenie skróconych wersji testów - ujęcie w ramach Klasycznej Teorii Testów i IRT

Gotowe narzędzia psychometryczne w konfrontacji z praktyką badawczą czasami (jeśli nie często) okazują się nie uwzględniać czasu potrzebnego na przeprowadzenie badania. Wysokie parametry trafności i rzetelności mogą być osiągnięte, lecz kosztem długiego czasu potrzebnego na pełny przebieg procedury badawczej. W tym kontekście nic dziwnego, iż psychologowie stosujący testy i kwestionariusze w praktyce, czasami dążą do uzyskania skróconej wersji narzędzia badawczego, opierając się na swojej intuicji i zaniedbując przy tym wymogi metodologiczne. Korzystanie ze skróconej wersji danego narzędzia pomiarowego jest szczególnie uzasadnione podczas badania osób, które mają problem z koncentracją, uwagą, są męczliwe lub ciężko chore, czy nawet w przypadku osób zdrowych, gdy plan badawczy przewiduje zastosowanie całej baterii testów. Bez względu na źródło potrzeby wykorzystania skróconej wersji narzędzia, wersje to powinny pozwalać określać poziom funkcjonowania osoby uczestniczącej w badaniu w takim samym stopniu, jak wersje pełne (lub nieznacznie gorszym – porównaj, np. Choynowski, 1971, s. 65–115).

Proces tworzenia skróconych wersji istniejących narzędzi badawczych można prześledzić na przykładzie testu inteligencji Wechslera WAIS. W przypadku gdy mamy do czynienia z baterią testów, procedura skracania może przyjąć dwojaki charakter. Pierwszy sposób zaproponowali Paul Satz i Steve Mogel – ich metoda (metoda Satz-Mogel'a – *item reduction*) opierała się na wybieraniu z danego testu co drugiej lub co trzeciej pozycji testowej (Mogel i Satz, 1962; Kaufman, 1972). Dzięki temu, zachowana została wieloczynnikowa struktura inteligencji, a długość testu skróciła się o połowę lub dwie trzecie. Wynik uzyskany w poszczególnych testach po wymnożeniu odpowiednio przez 2 lub 3 odpowiada wynikowi uzyskiwanemu w pełnej wersji testu WAIS. Drugi sposób zaproponowany po raz pierwszy w 1967 roku przez Artura B. Silversteina polegał na redukcji liczby testów wchodzących w skład skali inteligencji WAIS (*subtest reduction*). Autor (Silverstein, 1982) zaproponował wybranie z całej skali tylko dwóch, a potem tylko czterech testów[4] (konkretnie Słownika, Arytmetyki, Porządkowania Obrazków i Klocków). Wielokrotnie porównywane wyniki uzyskiwane w obu wersjach pokazują, że na podstawie skal skróconych możliwe jest przewidywanie wyniku skali pełnej w 90% (z 7 punktowym przedziałem ufności dla IQ przy p = 0,05) (Crawford, Alla i Jack, 1992; Clara i Huynh, 2003). W metodzie Silverstein'a nie zmienia się sposobu przeliczania wyniku – obowiązują tylko inne tabele norm. Ponadto cechuje się ona wyższą rzetelnością i korelacją wyniku z wynikiem skali pełnej. Natomiast wersja skrócona według metody Satz i Mogel'a wymaga nowego sposobu liczenia wyników, co wpływa na czas potrzebny na przeprowadzenie badania, a także może skutkować zwiększeniem liczby błędów mechanicznych związanych z przeliczaniem wyników (Boone, 1991).

Innym podejściem do skracania testów lub kwestionariuszy jest zastosowanie różnych metod analizy statystycznej. Należą do nich głównie analiza czynnikowa, korelacje między skróconą a pełną wersją, korelacje między pozycjami testowymi lub pytaniami kwestionariusza a wynikiem ogólnym, współczynnik rzetelności α-Cronbacha lub regresja krokowej (Coste i in., 1997). Wszystkie te procedury opierają się na wewnętrznej zgodności pozycji skracanego narzędzia i zakładają, iż wynik otrzymany jest sumą zakodowanych odpowiedzi na poszczególne pozycje (Nunnally i Bernstein, 1994). Skracając narzędzie na tej podstawie i posługując się tylko analizą wewnętrznej spójności, badacz jest narażony na ryzyko zmiany znaczenia wyniku uzyskanego za

[4] Istnieją różne warianty tej metody, w których wybór waha się od dwóch do siedmiu testów (Warrington, James i Maciejewski, 1986).

pomocą wersji skróconej w porównaniu do wyniku z narzędzia oryginalnego. Często narzędzie pomiarowe nie będąc w pełni homogeniczne „skazuje" na odrzucenie właśnie te pozycje, które z natury są słabiej skorelowane z pozycjami centralnymi, gdyż jego składowe opisują nie tylko centralny obraz danego konstruktu, ale także jego obszary brzegowe. W takiej sytuacji najmniej przesunięć treściowych generuje metoda regresyjna, która pozwala zapoznać się ze strukturą wewnętrzną i wymaga od badacza podjęcia świadomej, a nie tylko mechanicznej, decyzji dotyczącej usunięcia poszczególnych pozycji.

Powyższe metody – oparte na KTT – nie zakładają porządku w pozycjach oraz ich ewentualnej hierarchii, w przeciwieństwie do metod, które są oparte na IRT. Te drugie pozwalają uporządkować pozycje danego narzędzia, a także wskazać jak dobrze poszczególne z nich opisują osoby uczestniczące w badaniach. Jest to możliwe poprzez obliczenie dla każdej pozycji wskaźnika DIF (*Differential Item Functioning* – Gruijter i Kamp, 2008, s. 182)[5] oraz określenie jak dobrze do grupy dopasowane są poszczególne osoby poprzez obliczenie dla każdej osoby poziomu Θ (ibidem, s. 135 i dalsze). Dzięki temu decyzja badacza dotycząca włączenia lub wykluczenia danej pozycji opiera się na wiedzy o strukturze wewnętrznej narzędzia oraz uwarunkowaniach związanych ze zmiennymi mogącymi mieć wpływ na wyniki.

Wyniki uzyskiwane w nowej, skróconej wersji poddawane są (a przynajmniej powinny być) walidacji krzyżowej, która informuje w jakim stopniu skrócona wersja przygotowana na jednej próbie, będzie przydatna przy przewidywaniu wyników pełnego narzędzia na innej próbie. Z perspektywy korelacji wielozmiennowych taka walidacja, gdy wyniki z jednej próby generalizujemy na inną próbę, pozwala określić stopień kurczenia się mocy prognostycznej (*shrinkage*) narzędzia do predykcji (Kerlinger i Pedhazur, 1973, s. 282). Należy zauważyć, że ten sposób walidacji pozwala empirycznie stwierdzić „ile straciliśmy", lecz nie służy do przewidywania dobroci nowo utworzonej wersji. Sami autorzy twierdzą, że jest to bardzo konserwatywna metoda (*"the most rigorous approach to the validation of results from regression analysis in a predictive framework"* – ibidem, s. 284). Walidacja krzyżowa zyskała bardzo mało uwagi w literaturze dotyczącej skróconych form narzędzi pomiarowych, a tam gdzie jest

[5] Wskaźnik DIF jest definiowany jako różnica między odpowiedziami na daną pozycję wynikająca z przynależności osób odpowiadających do różnych grup. Przykładem takiej pozycji może być pytanie w teście językowym odwołujące się do wiedzy na temat np. piłki nożnej. Gorsze odpowiedzi kobiet nie będą spowodowane ich gorszymi umiejętnościami językowymi, a słabszą orientacją w dziedzinie sportu (por. Camilli i Shepard, 1994).

stosowana przynosi niskie wyniki korelacji i wysokie błędy standardowe (Woo-Sam i Zimmerman, 1973, s. 1121).

W obu przypadkach (tym opartym na KTT jak i na IRT) należy rozważyć potencjalne ograniczenia i straty właściwości pomiarowych skróconego narzędzia (Shrout i Yager, 1989). I nawet jeśli mierzony konstrukt teoretyczny nie ucierpi ze względu na skrócenie testu lub kwestionariusza, to liczba pozycji ma duży wpływ na jakość ostatecznego pomiaru, szczególnie w nawiązaniu do rzetelności pozycji i rozkładów odpowiedzi. Niewystarczająca liczba łatwych lub trudnych pozycji pozwalających na zróżnicowanie osób o skrajnym poziomie badanej cechy spowoduje skośność rozkładu wyników (efekt „podłogowy" lub „sufitowy", Anastasi i Urbina, 1997, s. 239).

Rozdział 2. Tworzenie komputerowych wersji psychologicznych testów i kwestionariuszy

W obszarze psychologii (i nie tylko) konstruowanie komputerowych wersji narzędzi, jakimi są testy i kwestionariusze przynosi wymierne efekty. Należą do nich: redukcja kosztów badania, zmniejszanie liczby błędów w punktowaniu, zwiększanie standaryzacji warunków testowania przy jednoczesnym zautomatyzowaniu administrowania narzędziem badawczym oraz znaczne przyśpieszenie procesu przetwarzania zebranych wyników surowych. Jednocześnie wykorzystanie komputerów w procesie badawczym poszerza możliwości badania poza zwykłe zebranie odpowiedzi.

2.1. Użycie komputerów i internetu w badaniach psychologicznych

Komputery, początkowo wykorzystywane tylko w analizach danych, zaczynają spełniać coraz ważniejszą rolę podczas samego badania. Znajdują coraz więcej zastosowań zarówno w psychologii, jak i w psychometrii. Poza podstawowym gromadzeniem i analizą danych, komputery wykorzystywane są do:

- administrowania testów i kwestionariuszy – w postaci tradycyjnej, tzn. przełożonej na wersje elektroniczną metody funkcjonującej dotąd w postaci „papier-i-ołówek", jak i w postaci adaptacyjnej – pytań dobieranych dynamicznie w zależności od przebiegu badania;
- ekspozycji bodźców multimedialnych;
- pomiaru wskaźników fizjologicznych i neurologicznych (pulsu, reakcji skórno-galwanicznej, rejestracji ruchów gałek ocznych, aktywności mózgu);
- prowadzenia dialogu – uczenie nowych treści, ale także przeprowadzanie wywiadu ustrukturalizowanego, czy też testowania kwestionariuszowego;
- symulacji – szczególnie fantomy osobowości służące kształceniu psychologów w diagnozowaniu oraz, jak optymistycznie sądzono, pozwalające na sprawdzenie założeń teoretycznych teorii osobowości (por. Paluchowski, 2007, s. 274).

Wprowadzenie komputerów do przeprowadzania badań psychologicznych ma niewątpliwe zalety: zapewnia wysoką standaryzację i obiektywizację sytuacji badawczej, jednocześnie automatyzuje monotonne i mechaniczne fragmenty tego procesu, takie jak wprowadzanie, przekształcanie i archiwizowanie danych. Powstają systemy eksperckie, które wyręczają diagnostę w zestawianiu często sporej ilości informacji i proponują diagnozę ułatwiając interpretację zebranych danych (*ibidem*). O tym, że komputery zostały potraktowane poważnie przez psychologów, świadczy fakt

powołania już w 1971 roku pierwszego stowarzyszenia *Society for Computers in Psychology* oraz powstanie szeregu czasopism poświęconych psychologicznym aspektom stosowaniu komputerów (*Behaviour Research Methods, Instruments and Computers*, 1968; *Social Science Computer Review*, 1982; *Computer in Human Behaviour,* 1985; *Human-Computer Interaction*, 1985; *Computer Mediated Communication,* 1994 i inne).

Jednocześnie zwrócono uwagę na to, że używanie komputerów w psychologii niesie za sobą także pewne problemy. Badacze podnieśli kwestię równoległości wersji papierowej i elektronicznej (por. Matusik, 2000, za: Paluchowski, 2007). Testy i kwestionariusze w wersji komputerowej ograniczają zachowanie osób uczestniczących w badaniu. Brak jest możliwości podkreślania tekstu, skreślania odpowiedzi na pewno nie pasujących czy robienia notatek na marginesie. Inna też jest prędkość czytania tekstu na monitorze jak na papierze (Bernt, Bugbee i Arceo, 1990). Raport amerykańskiej organizacji *National Center for Fair and Open Testing* w odpowiedzi na upowszechnianie się testowania komputerowego w szkołach, wskazuje także na różnice w przebiegu badania polegające na: 1) dostępności jednocześnie tylko do jednego zadania, 2) różnice kulturowe związane z różnym statusem społecznym i ekonomicznym oraz 3) dyskryminację kobiet, wśród których jest bardziej powszechny lęk przed komputerem (Legg i Buhr, 1992). Jednakże zaobserwowane różnice w wariancji wyników mogą wyjaśniane są lepszą kontrolą podczas użycia komputerów niż podczas badań z udziałem papierowych wersji. A ponad to, porównując wyniki uzyskiwane w obu rodzajach badań za pomocą meta-analiz doniesień badawczych, nie potwierdzono wpływu na wyniki tzw. *effect mode*, czyli użytego medium za pośrednictwem którego zbierane są dane (Shih, Fan, 2008; Wang, Jiao, Young, Brooks i Olson, 2007; Finger i Ones, 1999; Mead i Drasgow, 1993). Mimo to, często wskazuje się, że wersje komputerowe wymagają traktowania ich jako odrębnej formy narzędzia i sporządzenia dodatkowych norm, np. w testach mierzących czas odpowiedzi.

Nie można też pominąć faktu, iż ważnym źródłem zmienności wyników badań (szczególnie przeprowadzanych na osobach starszych), mogą być dwa czynniki: analfabetyzm komputerowy rozumiany głównie jako nieumiejętność posługiwania się komputerem i zasobami elektronicznymi, oraz lęk przed komputerem - z jednej strony wynikający z obawy przed uszkodzeniem urządzenia, a z drugiej z uwarunkowań osobowościowo-temperamentanych.

2.1.1. Zjawisko „przepaści cyfrowej" – analfabetyzm komputerowy

Pod koniec lat 90.-tych, wraz z upowszechnieniem się dostępu do internetu[6] oraz coraz częstszymi próbami przeprowadzania badań za pomocą tego medium, zwrócono uwagę na osoby, które w badaniach nie uczestniczyły. Pierwsze analizy wskazywały na istnienie zjawiska nazwanego przepaścią cyfrową (*digital divide*) i definiowały je w kontekście pozostawania niepodłączonym do internetu. Wyniki badań były alarmujące – osoby niekorzystające z internetu zarabiają mniej, są gorzej wykształcone, rzadziej mają stałego partnera (patrz raporty *National Telecommunication and Information Administration* – NTIA: 1995, 1998, 1999). Tłumacząc różnice, badacze postulowali, że *digital divide* to przejaw istniejących podziałów, znajdujących odzwierciedlenie w stosunku do innych mass mediów: oglądania TV, czytania wiadomości w gazetach i słuchania ich w radio (Robinson, Barth i Kohut, 1997). Pojawiały się też głosy przypisujące przepaści cyfrowej atrybut katalizatora istniejących podziałów (Pietrowicz, 2002), argumentując, że gdyby tylko *digital divide* była powodem różnic, to jej proste zniwelowanie (wystarczy przecież „podłączyć ludzi do internetu") powinno te różnice wyeliminować. W miarę badań nad zjawiskiem pojawiały się różne jego definicje. Wilson (2000) zaproponował wprowadzenie definicji „cyfrowej przepaści" jako problemów w dostępie do technologii w czterech obszarach: 1) finansowym, gdzie możliwości ekonomiczne decydują o dostępie do internetu, 2) poznawczym, gdzie użytkownicy wiedzą czego szukają i znajdują to, co z kolei wiąże się z 3) obszarem zasobów – istniejących treści, które mogą zaspokoić potrzeby użytkowników. I wreszcie obszar 4) polityczny, który oznacza respektowanie przez administrację elektronicznego sposobu dostępu do jej przedstawicieli, usług i tym podobne.

Inne, równie szerokie podejście do zjawiska zaproponował Norris (2001, za: Hargittai, 2003, s. 10) – wyróżnił on trzy poziomy przepaści cyfrowej: pierwszy, zwany globalnym, zdeterminowany przez rozwój techniczny i uprzemysłowienie danego kraju, wiążący wszelkie wątki ekonomiczne dostępu do technologii; drugi – socjalny, opisujący wszelkie nierówności demograficzne i trzeci – demokratyczny, zawierający podział na tych, którzy używają technologii i nie używają ich w życiu społeczno-politycznym.

Powyższe definicje „przepaści cyfrowej" są bardzo szerokie i bardziej skupiają się na kontekście, niż na podmiocie. Jak słusznie zauważyli Paul DiMaggio i Eszter Hargittai (2001; podobnie w Polsce: Batorski, 2006), sam sposób używania komputerów

[6] Pierwsza graficzna przeglądarka internetowa została wymyślona w 1992 roku (Sherman, 2003) – odtąd można datować rozwój i popularyzację stron www.

nie jest jednolity. Ludzie korzystają z nich różnorodnie, realizując odmienne potrzeby – stąd też bardziej powinno się mówić o cyfrowej nierówności (*digital inequality*), niż o cyfrowej przepaści (*digital divide*). W skład zaproponowanego przez autorów pojęcia nierówności weszły takie wymiary jak: 1) środki techniczne, 2) autonomia, 3) sieci społecznego wsparcia, 4) doświadczenie i 5) umiejętności.

Pod określeniem **środki techniczne** należy rozumieć jakość sprzętu, jakim dysponują użytkownicy (*hardware*), możliwości oprogramowania (*software*) oraz prędkość połączenia z internetem. Osoby dysponujące nowszym sprzętem, sprawnie działającym, z szerokopasmowym dostępem do internetu chętniej korzystają z jego zasobów, także w szerszym stopniu (Hargittai, 2003, s. 11).

Pod pojęciem **autonomii** kryje się rozróżnienie między tym, jak łatwy jest dostęp do internetu teoretycznie, a jak praktycznie ludzie z tego dostępu korzystają[7]. Na przykład to, że w bibliotekach uniwersyteckich są komputery z dostępem do internetu, nie oznacza, że wszyscy studenci o tym wiedzą i że mogą z nich korzystać (np. biblioteki są otwarte tylko w określonych godzinach).

Sieci społecznego wsparcia dostarczają wiedzy na temat sposobów korzystania z technologii, służą pomocą przy rozwiązywaniu ewentualnych problemów – im więcej osób wokoło korzysta z internetu, tym łatwiej zacząć samemu i tym łatwiej znaleźć innych, o podobnym poziomie zaawansowania, z którymi można dzielić swe doświadczenia. Osoby, które mają więcej znajomych, uczą się szybciej i korzystają z szerszych zasobów, niż osoby z mniejszą liczbą znajomych (Hargittai, 2003, s. 12).

Doświadczenie to z kolei wymiar nierówności cyfrowej obrazujący w jaki sposób ludzie inwestują czas w poznawanie technologii. Osoby, które spędzają więcej czasu w internecie mają większe doświadczenie w zakresie uzyskiwania pożądanych informacji, tym samym łatwiej potrafią ją zdobyć. Doświadczenie może też być zależne od rodzaju dominującej aktywności – osoby spędzające czas głównie na portalach społecznościowych, będą inaczej radziły sobie, np. z wyszukiwaniem informacji niż osoby wykorzystujące internet do pracy, czy też te, którym internet służy głównie do robienia zakupów.

Wymiar **umiejętności** według autorów powinien być rozpatrywany analogicznie do umiejętności czytania – mimo, że można je sprowadzić do binarnego podziału na osoby

---

[7] Przykład braku autonomii w miejscu pracy: w biurze jednej z większych firm architektonicznych w jednym z dużych miast polskich, tylko jeden z komputerów jest podłączony do internetu. Teoretycznie każdy może z niego w dowolnej chwili skorzystać, ale w praktyce ograniczone jest to do minimum, ponieważ wiąże się z naznaczeniem: „korzysta z internetu".

umiejące i nie umiejące korzystać z technologii, to bardziej zasadne jest postrzeganie tego wymiaru jako nabierania wprawy w umiejętnym i efektywnym korzystaniu z technologii. W wymiarze praktycznym jest to np. umiejętność zadawania zapytań wyszukiwarkom, tak aby maksymalizować pojawienia się pożądanych wyników, ale też korzystanie z możliwości, jakie istnieją w oprogramowaniu. Tutaj Hargittai podaje przykład bardzo rzadkiego wyszukiwania przez badanych informacji na stronie – tylko 1% badanych przez nią uczestników korzystało z tej funkcji, która jest wbudowana we wszystkie przeglądarki (ibidem, s. 14).

W miarę badań nad zjawiskiem nierówności cyfrowej po roku 2000 stwierdzono, iż wraz z upowszechniającym się dostępem do internetu zanikają różnice socjodemograficzne (Rice i Katz, 2003, s. 600). W związku z powyższym Ronald E. Rice i James E. Katz (2003) zaproponowali podział osób używających nowe technologie oparty na zakresie ich użytkowania i składający się z trzech warstw: 1) podział na użytkowników (*users*) i nie-użytkowników (*nonusers*), 2) użytkowników nowych (*recent*) i długotrwałych (*veteran*) oraz 3) używających kiedyś (*dropout*) i używających nadal (*current*). Badania przeprowadzone przez autorów, wykorzystujące taką kategoryzację użytkowników pokazały, że różnice socjodemograficzne dotyczą głównie wieku, zarobków i wykształcenia. Młodsze osoby to głównie grupa użytkujących nadal, nie-użytkownicy zarabiają mniej niż użytkownicy, a osoby, które zrezygnowały z używania technologii, są gorzej wyedukowane.

W 2003 roku rząd amerykański zaprzestał finansowania programów przeciwdziałających przepaści cyfrowej, uprawomocniając niejako wnioski wynikające z badań (Katz, Rice i Aspden, 2001) – użytkownicy komputerów i internetu są analogiczną grupą jak populacja kraju (przynajmniej w USA). Różnice do tej pory postrzegane między użytkownikami i nie-użytkownikami okazują się nieistotne, jeśli uwzględni się wykształcenie, wiek i zarobki.

2.1.2. Technofobia – lęk przed komputerem i internetem

Masowe upowszechnianie technologii powoduje konieczność przyjrzenia się zjawiskom psychologicznym, które mu towarzyszą: zarówno pozytywnym jak i negatywnym aspektom interakcji człowieka i technologii. Jednym z negatywnych efektów asymilacji technologi w życiu codziennym, mających wpływ na efektywność działań i postawy wobec technologii, jest lęk. Istnienie specyficznego rodzaju lęku przed technologią zostało potwierdzone w wielu badaniach (por. Chua, 1997). Może się on pojawić w następstwie konieczności, czy też możliwości interakcji z komputerem,

w szczególności przy korzystaniu z internetu (Simonson, Maurer, Montag-Torardi i Whitaker, 1987, Rosen, Sears i Weil, 1987). Należy nadmienić, iż jest to stan, który ulega zmianie, np. w skutek częstych kontaktów z technologią lub treningu (Kleka, 2011).

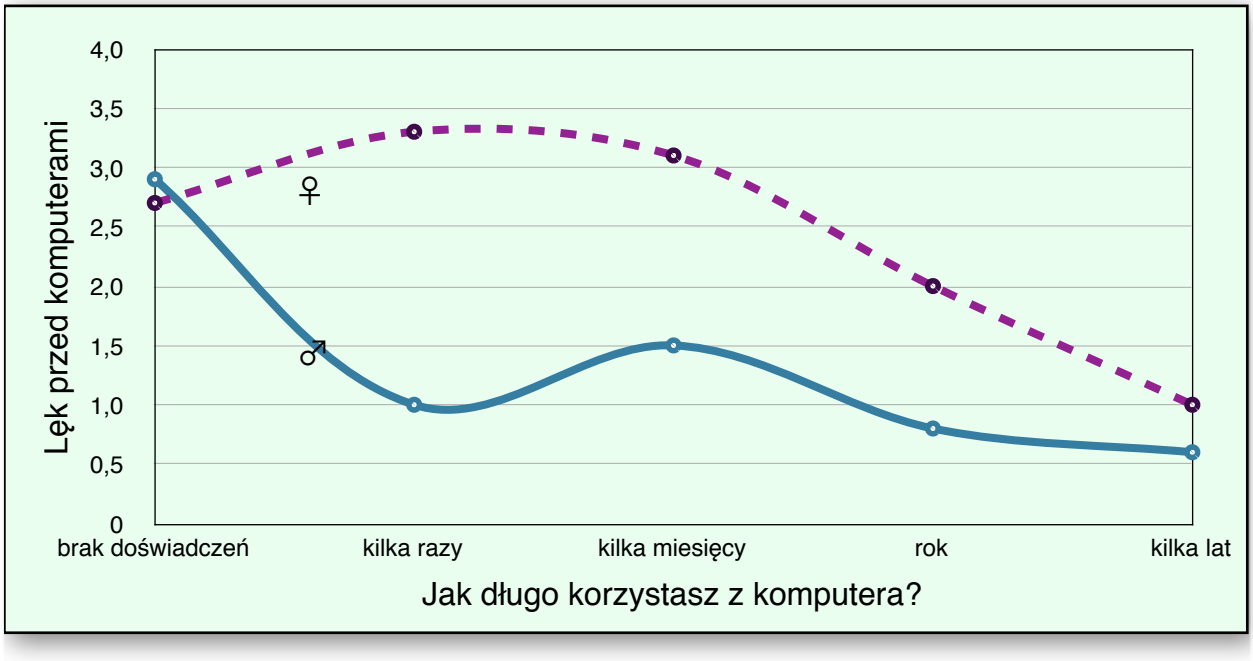
Do tej pory stworzono wiele narzędzi do pomiaru lęku przed technologią, np.: *Computer Anxiety Index* (Maurer, 1983 za: Marcoulides i in., 2004, s. 312); *Attitudes Toward Computer Scale and Computer Usage Checklist* (Raub, 1984, za: Marcoulides i in., 2004, s. 312); *Computer Attitude Scale* (Loyd, Gressard, 1984); *Computer Anxiety Scale* (Campbell, Dobson, 1987); *Computer Anxiety Scale* (Marcoulides, 1989); *Computer Anxiety Rating Scale* (Heissen, Glass, Knight, 1987); *Computer Anxiety Rating Scale* (Rosen, Sears, Weil, 1987). Badania nad korelatami lęku przed technologią (Harris, Neal, 1996; Brosnan, 1999; Zhang, 2005; Broos, 2005; Cooper, 2006) wskazują na istnienie trzech głównych zmiennych wpływających na jego odczuwanie: 1) wieku, 2) płci oraz 3) doświadczeń z komputerem.

W przypadku pierwszej z tych zmiennych, pogłębiona analiza (Broos, 2005) wykazała, że nie sam wiek jest istotny dla poziomu lęku, ale pokolenie, z którego pochodzą osoby uczestniczące w badaniach. Fakt późniejszej styczności z technologią wpływa na poziom lęku. Wiele, wydawać by się mogło, prostych umiejętności (jak obsługa klawiatury, myszki, percepcja „okienek" we współczesnych systemach operacyjnych) sprawia trudność ludziom, którzy nie mieli z nimi wcześniej kontaktu. Pomijając różnice neurologiczne między osobami młodszymi i starszymi, umiejętności te są łatwiejsze do przyswojenia w młodszym wieku, gdzie interakcja z komputerami ma często charakter zabawy, niż później, gdy może pojawiać się frustracja związana z dysonansem między oczekiwaną i spostrzeganą samoskutecznością.

Podobnie jak w przypadku wieku, różnice międzypłciowe wskazywane przez badaczy poddane szerszej analizie przynoszą zaskakujące rezultaty. Jak podaje Cooper (2006), stosunek do technologii jest różnicowany od najmłodszych lat. W społeczeństwie istnieje stereotyp, że komputery są dla mężczyzn – tzw. „*boy-toy*". Konsekwencją tego są np. gry edukacyjne uwzględniające potrzeby chłopców, a zupełnie niedopasowane do potrzeb dziewcząt. Przy okazji badań nad dziećmi uzyskano wyniki wskazujące na to, iż jednym z ważnych wyznaczników poziomu wykonania zadania, silnie związanym z płcią, jest ekspozycja społeczna. Dziewczęta rozwiązując zadania przygotowane dla chłopców są w stanie uzyskiwać nawet lepsze wyniki od nich, pod warunkiem, że rozwiązują je same, lub we własnym towarzystwie. W obecności chłopców dziewczęta osiągają gorsze wyniki, natomiast dla chłopców

obecność dziewcząt wpływa na poprawienie wyników (Light i in., 2000; porównaj także Nicholson i in., 1998).

Trzecią zmienną wpływającą na poziom lęku przed komputerami jest charakter pierwszych kontaktów z nimi. Pierwszy kontakt z czymś nieznanym przeważnie wywołuje reakcję lękową – to, czy zostanie ona utrwalona, czy przekształcona, zależy od przebiegu tej pierwszej, czy też kilku pierwszych relacji. Wyniki podłużnych badań Agnethy Broos (2005) wskazują, że kobiety i mężczyźni ostatecznie nie różnią się poziomem lęku – istotną różnicę zaobserwowano w przebiegu jego redukcji – mężczyznom wystarcza kilka kontaktów, aby osiągnąć końcowy, niski poziom lęku (patrz ryc. 2.1). Kobiety potrzebowały na ten sam proces ponad rok (także Chou, 2003).



Ryc 2.1. Zależność poziomu lęku przed komputerem od czasu doświadczeń i płci. Źródło: Bross (2005, s. 27), opracowanie własne.

Podsumowując kwestie analfabetyzmu i lęku komputerowego należy podkreślić, że użycie technologii informacyjnej w badaniach psychologicznych wymaga uwzględnienia dodatkowych zmiennych, aby móc kontrolować to potencjalne źródło błędów. W przypadku nierówności technologicznej uwzględnienie takich zmiennych jak wykształcenie, wiek i zarobki pozwala kontrolować to źródło wariancji. W przypadku lęku przed technologią najlepiej, gdyby był on mierzony w sposób bezpośredni, ponieważ doświadczenie jest gorszym, niż lęk predyktorem powodzenia w zadaniu, które wymaga użycia komputera (Harris, Neal, 1996). Jednocześnie należy pamiętać, że jest on w dużym stopniu pochodną doświadczeń osób uczestniczących w badaniach z komputerami, a jego zmiana w czasie jest modyfikowana przez płeć.

Brak na gruncie polskim znormalizowanego kwestionariusza lęku przed technologią może być zrekompensowany wprowadzeniem do planów badawczych takich zmiennych jak płeć i doświadczenie komputerowe, które są łatwiej dostępne w introspekcji dla osób uczestniczących w badaniach, niż samo doznanie lęku.

2.1.3. Klasyfikacja zachowania się osób uczestniczących w badaniach internetowych

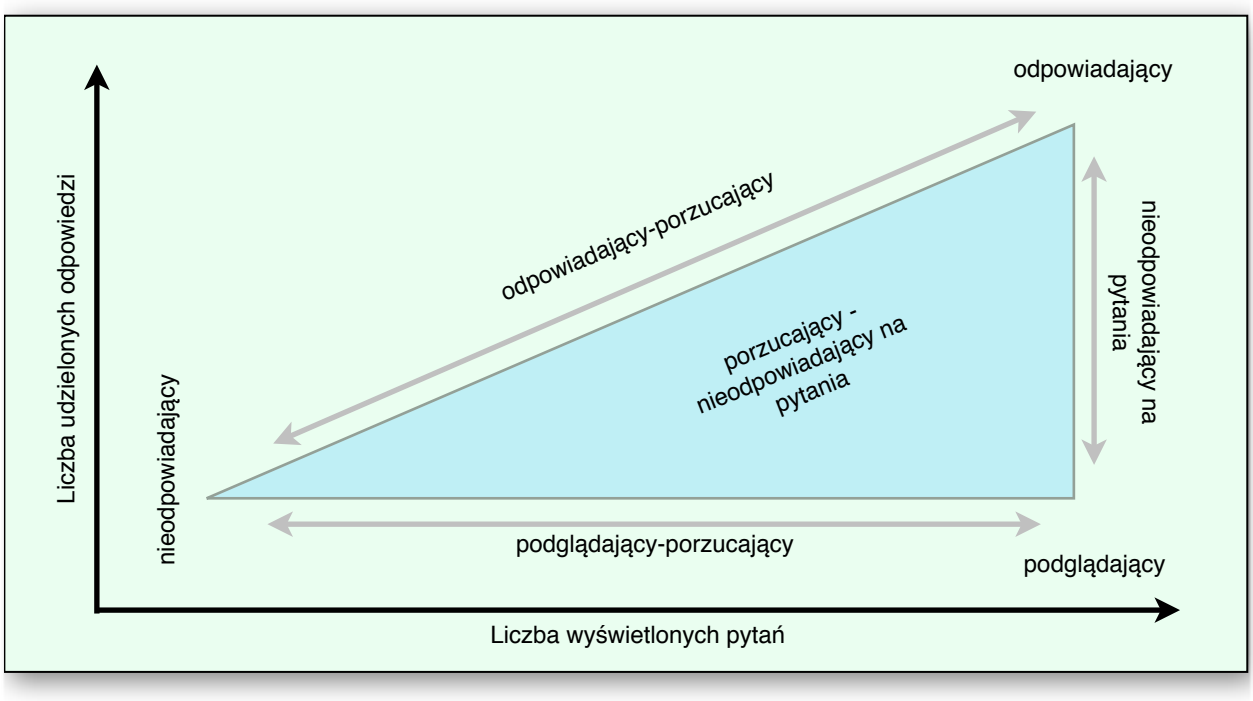
Decydując się na badania przeprowadzane za pomocą internetu napotyka się problem doboru osób do próby. Mimo, że populacja internautów jest dobrze opisana (por. Czapiński i Panek, 2011) dostęp do internetu w Polsce ma ponad 73% mieszkańców dużych miast i 51% mieszkańców wsi, co stawia pod znakiem zapytania możliwości generalizacji wyników (stan na czerwiec 2011; w niektórych krajach Europy wskaźnik ten zbliża się do 100% na przykład: Norwegia, Szwecja, Islandia, a w Stanach Zjednoczonych, które są pierwszym krajem pod względem bezwzględnej liczby użytkowników internetu odsetek ten wynosi ponad 77%). Ponadto użytkownicy internetu w porównaniu do całej populacji są młodsi, lepiej wykształceni, pochodzą z większych miast, a ze zmiennych psychologicznych charakteryzują się wyższym poziomem stresu i ekstrawersji (Batorski, 2005). Chcąc zbudować losową próbę napotykamy na dodatkowy problem z operatem losowania: tylko około 80% użytkowników internetu posiada adres poczty elektronicznej, a 67% osób sprawdza go tylko raz w tygodniu. Dodatkowo wiele osób posiada kilka adresów, co jeszcze bardziej utrudnia przeprowadzenie doboru losowego do prób badawczych. W tej sytuacji badacze polegają na rekrutacji poprzez ogłoszenia na stronach www[8], co zwiększa wpływ efektu ochotników. Należy dodać, że różne miejsca przeprowadzania rekrutacji ten efekt minimalizują (Batorski i Olcoń-Kubicka, 2006), ale nie usuwają.

Przyjmuje się za oczywiste, że stopień realizacji próby waha się od kilku do kilkunastu procent i w większości badań analizuje się dane, które udało się w nich zebrać. Badania poświęcone brakom danych i rezygnacji z udzielania odpowiedzi skupiają się na dwóch aspektach: 1) jak zwiększyć odsetek odpowiedzi, ewentualnie skorygować wyniki tak, aby uwzględnić osoby, które odmówiły udziału w badaniach (Dillman, 2000; Claycomb, Porter i Martin, 2000, Bickart i Schmittlein, 1999) oraz 2) jakie są przyczyny rezygnacji z udziału w badaniach (Couper i Rowe, 1996; Groves, Cialdini i Couper, 1992). Pozostaje jednak otwartym pytanie: Czy czynniki, które

[8] Czasami możliwe jest użycie jako operatu losowania adresów mejlowych zgromadzonych w różnych bazach – uczelni, list mejlowych, forów internetowych. To podejście nie rozwiązuje jednak w zupełności problemu generalizacji wyników.

spowodowały, rezygnację większości osób zaproszonych do badań, nie mają charakteru stałej cechy wpływającej na wyniki. Innymi słowy, pamiętając o obecności wpływu ochotniczego (szczególnie przy badaniach otwartych), należy także rozpatrzyć ewentualne powody, dla których osoby badane nie odpowiedziały lub odpowiedziały częściowo na pytania. Pomoc w tym zakresie stanowi klasyfikacja zachowań osób badanych, zaproponowana przez badaczy Michaela Bosnjaka i Tracy L. Tuten (2001). W artykule *Classifying response behaviors in web-based surveys* autorzy zaproponowali klasyfikację zachowań osób uczestniczących w badaniach za pośrednictwem internetu opierając się na dwóch wymiarach: liczbie zaprezentowanych pytań i liczbie udzielonych odpowiedzi[9]. W oparciu o to wyróżnili siedem podgrup osób uczestniczących w badaniach (patrz: ryc. 2.3). Z punktu widzenia badacza najbardziej pożądaną jest grupa osób, które zapoznały się z wszystkimi pytaniami oraz udzieliły na nie odpowiedzi (*Complete Responders*). Możliwa jest też sytuacja przeciwna, gdy osoby zrezygnują z badania zanim się ono rozpocznie (*Nonresponders*). Powodem mogą być przeszkody techniczne bądź decyzja o rezygnacji, na przykład po zapoznaniu się z planszą tytułową, bez kontaktu z pozycjami narzędzia. Trzecią grupę stanowią obserwatorzy – osoby, które chcą zapoznać się z metodą, ale nie biorą udziału w badaniu (*Lurkers*) – przeglądają one wszystkie pozycje, ale nie udzielają żadnych odpowiedzi. Osoby, które zapoznając się z kolejnymi pozycjami zrezygnowały z udziału w badaniu, w zależności od tego czy udzielały odpowiedzi, klasyfikowane są albo jako odpowiadający-porzucający (*Answering Drop-Outs),* albo jako obserwatorzy-porzucający (*Lurking Drop-Outs*). Ostatnie grupy stanowią osoby, które zapoznały się z częścią lub całością pozycji w badaniu i udzieliły odpowiedzi na niektóre z nich. Są to osoby, które zrezygnowały przed końcem badania – porzucający-nieodpowiadający na pytania (*Item Nonresponding Drop-Outs*) lub po zapoznaniu się z całą metodą – nieodpowiadający na pytania (*Item Nonresponders*).

[9] Przyjęli oni trzy założenia: osoba uczestnicząca w badaniu za pośrednictwem internetu może zobaczyć wszystkie pozycje, które pojedynczo wyświetlanie są na monitorze; może przejść do dalszych pozycji nie odpowiadając na poprzednie; jej kolejne odpowiedzi są zapisywane.



Rycina 2.3. Klasyfikacja typów zachowań osób uczestniczących w badaniu ze względu na relację między ilością pytań z którymi osoba się zapoznała a liczbą udzielonych odpowiedzi. Źródło: Bosniak i Tuten (2001).

Oczywiście należy dążyć do jak największego udziału osób odpowiadających na wszystkie pozycje, minimalizując grupy, w których osoby porzucają badanie. Według autorów umożliwiają to dwa kierunki działań: 1) dostosowywanie formy badania do oczekiwań osób uczestniczących – minimalizując ewentualną frustrację oraz 2) motywowanie do pozostawania w badaniu.

Ponieważ rezygnacja z badań w sytuacji wysokiego poczucia anonimowości wiąże się z niewielkimi kosztami psychologicznymi, oddziaływanie na motywację osób uczestniczących do pozostawania w badaniu przynosi dobre efekty. Ujęcie obejmujące prawie wszystkie czynniki motywacyjne zaproponował Ulf-Detrich Reips w artykule *Standards for Internet-based experimenting* (2002). Jego metoda wysokiego progu (*high-hurdle*) zakłada, że w badaniu należy zadbać o następujące czynniki wpływające na motywację do jego ukończenia:

- powaga – poprzez poinformowanie uczestników badania o ważności badań i potrzebie uzyskania rzetelnych wyników,
- personalizacja – poprzez zapytanie o adres mejlowy lub numer telefonu,
- przeświadczenie o kontroli – poprzez poinformowanie o identyfikowaniu komputera, z którego ankieta jest wypełniana,
- cierpliwość – najwięcej treści należy zamieścić na początku narzędzia, sukcesywnie zmniejszając ich ilość ze strony na stronę,

- czas trwania – należy podać orientacyjny czas trwania badania,
- prywatność – na początku badania poinformować osoby, które zechcą uczestniczyć w badaniach, że zostaną zadane pytania, np. o poziom dochodów,
- warunki – należy określić wymagane elementy oprogramowania i podać odnośniki do ich pobrania i zainstalowania,
- pre-test techniczny – przeprowadzenie testu sprawdzającego czy komputer osoby zaproszonej do wzięcia udziału w badaniach spełnia wymagania techniczne,
- nagroda – informacja o nagrodzie przyznanej / wylosowanej za pełny udział w badaniu.

Dodatkowo poprzedzając właściwą fazę badania kilkoma „pytaniami na rozgrzewkę" (*warm-up phase*) zapewniana jest wysoka jakość ostatecznie zebranych danych.

Stosowanie komputerów oprócz tego, że daje nowe możliwości przeprowadzania badań, wymusza też nowe sposoby kontrolowania warunków, w jakich badania się odbywają. Stosowanie tego rodzaju medium, ze świadomością ewentualnego wpływu na wyniki, przy uwzględnieniu interakcji cech osób i cech narzędzi badawczych, pozwoli wykorzystać zalety i uniknąć wad badania za pośrednictwem komputerów.

2.2. Podejście tradycyjne – przełożenie metody „papier i ołówek" na wersję elektroniczną

Najprostszym i także najczęstszym sposobem zastosowania technologii komputerowych do badania jest „przetłumaczenie" gotowego kwestionariusza lub testu w wersji „papier i ołówek" na wersję elektroniczną. Proces ten, odbywający się w tym samym obszarze językowym nie wymaga procesu adaptacji – zmiana medium jest bezproblemowa, pod warunkiem jednak, że zakłada się porównywalną sprawność posługiwania się technologią, co papierem i ołówkiem. O ile to założenie jest prawdziwe dla młodszych pokoleń, to wydaje się, że dla osób starszych, które z technologią stykają się w drugim, wtórnym procesie edukacji, wyniki uzyskiwane w obu wersjach mogą się różnić, ponieważ będą związane pośrednio ze „sprawnością technologiczną". Mimo, że problem wydawać się może marginalny lub marginalizujący się w miarę poszerzania dostępu do technologii, to nie może być bagatelizowany. Dla części badaczy wersja komputerowa ma status odrębnego narzędzia i wymaga podania odrębnych norm (por. Matusik, 2000, za: Paluchowski, 2007, s. 276). Jednakże, jak donoszą badacze zajmujący się różnymi dyscyplinami nauki, najczęściej nie stwierdza się różnicy w uzyskiwanych wynikach związanej ze sposobem prezentacji pozycji (Kim

i Huynh, 2008; Denscombe, 2006, 2008; Meade, Lawrence i Lautenschlanger, 2007; Whitaker, 2007; Carlbring, Brunt, Bohman, Austin, Richards, Ost i Andersson, 2007; Cole, Bedeian i Feild, 2006; Choi, Kim i Boo, 2003; Powella, Wilsonb i Hastya, 2002; Miller, Neal, Roberts, Baer, Cressler, Mertik i Marlatt, 2002; Wiechmann i Ryan, 2003). Według Nowakowskiej (za: Terelak i in., 1994, s. 380) wyniki są równoważne a testy równoległe, jeżeli dla każdej osoby badanej rozkład cechy posiada taką samą wartość oczekiwaną i wariancję. Oznacza to, że wprowadzenie nowego medium nie jest nowym źródłem systematycznej wariancji wyników oraz nie pojawiły sie nowe źródła wariancji błędu pomiaru.

Interaktywność środowiska komputerowego stanowiąca niespecyficzne źródło zmienności wyników badań odnosi się do: 1) interaktywności metody badawczej, w której mogą się pojawiać treści w zależności od zachowań osoby badanej (np. w ankiecie: pytania filtrujące ukrywające lub odsłaniające dalsze pozycje w zależności od udzielonej odpowiedzi) oraz przede wszystkim 2) interaktywność środowiska jako takiego: osoba uczestnicząca w badaniu potencjalnie ma w trakcie badania możliwość przełączenia się do innego programu, skorzystania np. z wyszukiwarki internetowej, słownika; może w jednej chwili analizować dane, by w następnej obejrzeć materiał wideo, posłuchać muzyki i tym podobne. Jak pisze Lev Manovich (2006, s. 324): „Współczesne interfejsy pozwalające na uruchamianie wielu programów w tym samych czasie oraz równoczesne otwieranie wielu okien na ekranie uznają wielozadaniowość za społeczną i poznawczą normę. Ta wielozadaniowość wymaga od użytkownika „poznawczej wielozadaniowości" – szybkiego przechodzenia między różnymi rodzajami uwagi i różnymi rodzajami umiejętności poznawczych. Współczesne komputery wymagają od użytkownika sprawnego rozwiązywania problemów, systematycznego eksperymentowania i szybkiego uczenia się nowych zadań". Interaktywność środowiska stanowi obok wymienionych wcześniej czynników (lęku przed komputerami i analfabetyzmu), ważną część uwarunkowań wariancji wyników w sytuacji badania z użyciem komputerów.

2.2.1. Elementy testu i/lub kwestionariusza w badaniach komputerowych

Kwestionariusz przeważnie składa się z dwóch głównych części – instrukcji i zestawu pozycji, które mogą być np. pogrupowane tematycznie (na część metryczkową i poszczególne skale). Instrukcja służy wprowadzeniu osoby uczestniczącej w badaniu w temat badania, wyjaśnieniu sposobu wypełniania, pełni też rolę motywującą, a w niektórych przypadkach także aktywizuje określone struktury

poznawcze odwołując się do konkretnych doświadczeń, sytuacji, czy też podając zakres czasowy jaki ma być brany pod uwagę przy wypełnianiu pozostałej części (np.: „Jak najszybciej odpowiedz na pytania…", „Pomyśl jak czułeś się przez ostatni tydzień i…"). Tradycyjny kwestionariusz składa się z pozycji, w których uzyskane wyniki są podstawą do obliczenia wyniku ogólnego, pozwalającego na pomiar natężenia badanej cechy (Zawadzki, 2006). Nie inaczej jest w przypadku kwestionariuszy używanych w badaniach internetowych. Pozycja może być:

- jednowyrazowym określeniem („smutny"),
- stwierdzeniem szczegółowym („Kiedy patrzę w niebo widzę mrugające gwiazdy"),
- stwierdzeniem ogólnym („Jestem osobą wrażliwą")
- pytaniem („Czy potrafiłbyś odebrać nieznajomą osobę z lotniska?").

To ostanie może: 1) mieć funkcję filtrującą, pomagającą osobie uczestniczącej w badaniu ominąć nie dotyczące jej pozycje lub 2) wymagać podania rozwiniętej, dowolnej wypowiedzi (pytanie otwarte) lub też 3) wybrania odpowiedzi z dostępnej puli przygotowanej przez badacza.

Jednakże istnieją pewne ograniczenia techniczne, sprowadzające zapędy twórcze przy konstruowaniu kwestionariuszy lub testów do konieczności wykorzystania kilku gotowych rozwiązań. Poszczególne pozycje mogą być zrealizowane w jednym z czterech sposobów interakcji z użytkownikiem:

- pobierania tekstu – tzw. pola tekstowe (*text fields*),
- wybierania jednej opcji – tzw. pola opcji (*radio buttons*),
- wybieranie z rozwijanej lista (*combo lists*),
- zaznaczania kilku opcji – tzw. pola wyboru (*check boxes*).

Za pomocą tych czterech elementów można zbudować większość pozycji potrzebnych do przedstawienia w formie elektronicznej kwestionariuszy i testów. I tak:

- pola tekstowe – służą jako miejsca zachęcające do wprowadzenia dowolnej odpowiedzi – funkcjonują samodzielnie jako pytania otwarte lub w zestawie z polami opcji jako pozycja „inne, proszę podać jakie". Pola te mogą być dowolnej wielkości, zachęcając do wprowadzenia większej lub mniejszej ilości informacji. Mogą być także wyposażane w mechanizmy ewaluacyjne, które sprawdzają poprawność wpisanych danych według ściśle określonego formatu (np.: data, adres e-mail);
- pola opcji stanowią podstawę pytań z jednokrotną odpowiedzią. Mogą mieć postać okrągłych przycisków umieszczonych przy pozycjach odpowiedzi lub też być prezentowane w postaci rozwijane listy, z której osoba uczestnicząca w badaniu wybiera jedną pozycję. Zmiana wybranej pozycji następuje poprzez wybór kolejnej;

- pola wyboru – pozwalające zaznaczyć dowolną lub określoną liczbę odpowiedzi. Wybieranie kolejnych pozycji przełącza je między stanem zaznaczenia i odznaczenia;
- skale – pola opcji ułożone narastająco, z charakterystyką (liczbową lub pojęciową) kolejnych pól, posiadające opis na skrajnych pozycjach lub tylko z jednej strony – pozwalają mierzyć natężenie, poziom lub wielkość zjawiska na skali ilościowej lub porządkowej. Skale odpowiedzi łączone ze sobą mogą tworzyć siatki odpowiedzi.

Korzystanie z narzędzi elektronicznych standaryzuje w pewnym stopniu ich budowę i ułatwia ich tworzenie. Nie uwalnia jednak od błędów odpowiedzi związanych z procesem szacowania: 1) błędu łagodności – tendencja do oceny znanych sobie osób w sposób bardziej korzystny; 2) błędu tendencji centralnej – czyli unikanie wartości skrajnych; 3) błędu efektu aureoli (*effect halo*) – czyli uleganie ogólnemu wrażeniu, np. początkowym określeniom i udzielanie pozostałych odpowiedzi zgodnie z tą tendencją; 4) błędu bliskości – jeśli cechy występują zbyt blisko siebie (np. skale oceniające podobne cechy) wzrasta ich interkorelacja; 5) błędu kontrastu – gdy osoby uczestniczące w badaniach podwyższają oceny cech posiadanych przez siebie, postrzegając innych jako podobnych do siebie (por. Brzezińska i Brzeziński, 2004, s. 299–303).

2.2.2. Wpływ formy kontaktu z osobami uczestniczącymi w badaniach na odsetek odpowiedzi

Zapraszając osoby do udziału w badaniach należy wybrać formę powiadomienia i wskazania „miejsca" badania. Zaproszenie to może być przesłane w sposób tradycyjny – papierowy – jeśli nie znane są adresy mejlowe osób wybranych do próby, lub jeśli chcemy kontrolować tożsamość osób w próbie. Częstą praktyką jest wysyłanie zaproszeń wykorzystując bazę adresów e-mail lub ogłoszenie na stronie www. Różne elementy zaproszenia mają wpływ na decyzję osób badanych o wzięciu udziału w badaniu. Jak podaje Alex R. Trouteaud (2004) samo podanie informacji o czasie potrzebnym na udział w badaniu zwiększa liczbę odpowiedzi. Z kolei Don A. Dillman (za: Heerwegh, Vanhove, Matthijs i Loosveldt, 2005) wykazał, że personalizacja zaproszeń zwiększa wskaźnik zwrotów (*response rate*) od 5 do 11% (podobny pozytywny wpływ personalizacji uzyskali Cook, Heat i Thompson w meta-analizie 68 badań internetowych – za: Porter i Whitcomb, 2007). Jednocześnie personalizacja zaproszeń zwiększa udział zmiennej aprobaty społecznej – badani mając poczucie bycia rozpoznawanym, są bardziej skłonni zmodyfikować swoje odpowiedzi, mimo zapewnienia o anonimowości wyników. Inni badacze – Stephen E. Porter

i Michael R. Whitcomb (2007) – analizowali wpływ powtórnych przypomnień różnej treści na odsetek odpowiedzi i mimo, że zwiększają one szanse na udział w badaniu większej liczby osób, to autorzy stwierdzają, że przysyłanie papierowych przypomnień o badaniu jest nieopłacalne z punktu widzenia ekonomicznego. Większy wpływ na wskaźnik odpowiedzi ma ujawnienie w zaproszeniu autora badań, nawet w postaci niejawnej (ankieta umieszczona pod adresem wskazującym na placówkę edukacyjną – .edu, lub rządową – .gov), niż takie czynniki, jak: informacja o końcu badania, identyfikowalny nadawca komunikatu, wielkość i znaczenie instytucji.

Analizując wyniki uzyskane przez badaczy można postawić tezę, że udział w badaniu zależy od konfiguracji trzech czynników: 1) zaufania, jakim obdarzony zostaje nadawca komunikatu, 2) subiektywnego kosztu, jaki trzeba ponieść biorąc udział w badaniu (w postaci czasu, zaangażowania, introspekcji) oraz 3) nagrody, jaką można otrzymać za swój udział (wymiernej lub też subiektywnej, np. w postaci zwiększenia wiedzy na swój temat w wyniku otrzymania informacji zwrotnej). Personalizacja kontaktu z osobami biorącymi udział w badaniach wpływa pozytywnie na dwa z wymienionych czynników: zwiększa zaufanie, ponieważ zwracamy się bezpośrednio do danej osoby oraz w przypadku spodziewanej wymiernej nagrody zwiększa subiektywne odczucie szansy na jej otrzymanie (już przecież jesteśmy wyróżnieni poprzez traktowanie osobowe).

2.2.3. Ograniczenia badań przez internet związane z użyciem stron internetowych.

W 1992 roku Maria E. Sanchez opublikowała ciekawy artykuł opisujący przypadek zmiany formy kwestionariuszy w trakcie badań. Okazało się, że proste zabiegi edycyjne polegające na użyciu tabel i w zamyśle twórców uproszczeniu wyglądu kwestionariusza, spowodowały istotne różnice miedzy wynikami uzyskiwanymi przez osoby uczestniczące w badaniach w obu wersjach. Jeśli wpływ na wyniki uzyskane w kwestionariuszach papierowych ma ich forma, to jaki wpływ będzie miało użycie elektronicznych wersji w miejsce papierowych? Jak różna forma kwestionariuszy elektronicznych będzie zwiększała wariancję wyników? Takie pytania należy sobie zawsze zadać podczas konstruowania elektronicznej wersji narzędzia. Próbą odpowiedzi na nie są np. badania na próbie 21 tysiącach studentów przeprowadzone w 2003 roku przez Andy Peytchev z zespołem (Peytchev, Couper, McCabe, Crawford, 2006, por także: Dillman i in., 1993). Zbadano różnicę między elektronicznymi wersjami kwestionariusza, z których jedna prezentowała na ekranie komputera pojedyncze pytania, a druga wersja wymagała przewijania, ponieważ pytania zaprezentowane były

na pojedynczej stronie www. Analizując wyniki badacze stwierdzili, że wersję przewijaną studenci wypełniali dłużej, niż wersję wielostronicową (krytycznie: Vehovar, Manfreda i Batagelj, 2000). Co więcej, w wersji, gdzie osoby uczestniczące w badaniu miały prezentowane pytania w swoim bezpośrednim sąsiedztwie stwierdzono wyższą korelację wewnętrzną skal. Ponadto subiektywna ocena ankiety częściej zawierała słowo „długa" dla wersji przewijanej niż wielostronicowej. Badanie to przyniosło także informację na temat zamieszczania linków w kwestionariuszu: podawanie informacji o liczbie ominiętych pytań w przypadku wyboru danej odpowiedzi zwiększało częstość jej wybierania. Podobnie umieszczanie linków w instrukcji wpływało na jej omijanie.

2.2.4. Wpływ kontekstu wizualnego na wyniki uzyskiwane w badaniach internetowych

Zaletą użycia grafiki jest często uproszczenie informacji – klaryfikacja zagadnienia poprzez jego ilustrację. Jak zatem wpłynie na wyniki użycie kolorowych elementów, obrazów czy też animacji? W 1991 roku Norbert Schwarz i współpracownicy wykazali w prostym eksperymencie na kwestionariuszach papierowych, iż samo użycie cyfr przy słownych opisach kategorii krańcowych danych skal zmienia znaczenie słów (Schwarz, Knauper, Hippler, Noelie-Neumann i Clark, 1991). Badania na tej samej dziesięciostopniowej skali z opisanymi krańcami: „zdecydowanie się zgadzam" i „zdecydowanie się nie zgadzam" podpisane dodatkowo w jednej wersji od -5 do +5, a w drugiej wersji od 0 do 10 przyniosły inne rozkłady wyników. Podobnie w 2003 roku Tourangeau, Couper i Conrad (2007) przeprowadzili serię eksperymentów sprawdzających wpływ na rozkład wyników elementów graficznych takich jak kolor, sposób rozmieszczenia odpowiedzi i podpisów w nich użytych. Przydzielając losowo osobom uczestniczącym w badaniu identyczne treściowo pytania na siedmiostopniowej skali, ale z innym tłem[10], otrzymano istotne różnice w rozkładach odpowiedzi. Wprowadzenie dwóch kolorów spowodowało, że osoby uczestniczące w badaniu traktowały skalę tak, jakby przedstawiała natężenie dwóch wymiarów, zaś w przypadku jednokolorowym – środek ciężkości wyników przesuwał się w stronę jednego z krańców. Co ciekawe, wprowadzenie opisów słownych pod skalą zlikwidowało efekt wywierany przez kolor (ibidem, s 101).

Umieszczanie ilustracji w kwestionariuszach elektronicznych zwiększa zaangażowanie osób w nich uczestniczących, ale jak donoszą badania poświęcone

[10] w jednej wersji kategorie pokolorowane były od ciemnego niebieskiego do jasnego niebieskiego, w drugiej wersji od ciemnego niebieskiego do ciemnego czerwonego, a środek był biały.

temu problemowi (Sargent, 2007; Prior, 2002, za: Couper i in., 2007) użycie grafiki w badaniach wpływa na wyniki w nich uzyskiwane. Ilustracje uaktywniają pamięć o danych wydarzeniach i na przykład umieszczenie obrazka przedstawiającego robienie zakupów, powodowało zwiększenie u respondentów liczby twierdzących odpowiedzi na pytanie o zakupy (Sergent, 2007). Nie stwierdzono jednocześnie różnicy w wielkości wpływu na wyniki związanego z umiejscowieniem ilustracji – z wyjątkiem umieszczania ilustracji na samym początku ankiety, w nagłówku. Couper, Conrad i Tourangeau (2007, s. 628) nazwali to zjawisko „ślepotą banerową" (*banner blindness*). Użytkownicy internetu bombardowani reklamami na stronach www, nauczyli się nie zwracać uwagi na grafikę w nagłówkach stron, ponieważ jest to zwyczajowe miejsce umieszczania reklam. Według badaczy obrazki w nagłówkach są pomijane jako nie związane z zadaniem (podobnie: Baneway i Lane, 1998). Pojęcie „banerowej ślepoty" uszczegółowili Bayles (2000) oraz Pagendarm i Schaumber (2006) zauważając, iż wystąpienie tego efektu jest związane z zadaniem stawianym przed osobami uczestniczącymi w badaniach – występuje on przy zadaniu polegającym na przeglądaniu internetu (*browsing mode*) a zanika, jeśli korzystamy z internetu w celu wyszukiwania informacji (*searching for information*). Częściowo potwierdzają tę tezę wyniki, które uzyskali Shrestha i Owens (2009) w badaniach nad fiksacją wzroku osób przeglądających strony internetowe – stwierdzili oni występowanie tzw. „wzoru F" (*F pattern*) – głównej fiksacji zaczynającej się nie od samej góry strony, ale poniżej linii nagłówka.

Analizując postrzeganie osób uczestniczących w badaniach Tourangeau, Couper i Conrad (2004) wyróżnili pięć heurystyk spostrzegania, które określają sposób i zakres wpływu konstrukcji wizualnej kwestionariusza lub testu na wyniki w nim uzyskane:

- element lewy i górny to pierwszy – w przypadku listy elementów pionowej lub poziomej, nadawana jest im hierarchia, osoby uczestniczące w badaniach traktują pozycje listy tak, jakby były uporządkowane,
- element środkowy to typowy – w listach elementów te, które zajmują centralną pozycję traktowane są jako symbolizujące przeciętne wartości,
- elementy występujące obok siebie są podobne – zarówno pytania, jak i pozycje odpowiedzi traktowane są tak, jak by były w związku między sobą,
- elementy umieszczone wyżej są lepsze – wtedy, gdy osoby uczestniczące w badaniach musiały interpretować wagę pozycji, te z nich, które umieszczono wyżej szacowano jako ważniejsze,

- elementy podobne z wyglądu są podobne treściowo – co wiąże się z traktowaniem wyglądu jako właściwości immanentnej dla obiektów i jeśli aspekty wyglądu pozwalają połączyć elementy ze sobą, to także treści są uwspólnione.

Heurystyki związane są z tendencją osób odpowiadających do doszukiwania się w bodźcach porządku i organizowania spostrzeganych treści tak, aby były łatwiejsze do poznawczego opracowania. Powyższe zasady są ważne przy konstruowaniu testów lub kwestionariuszy, np. umieszczanie odpowiedzi „nie wiem", „nie dotyczy" na końcu albo na początku listy elementów powoduje, że pozycja ta może być traktowana jako będąca poza skalą. Ponadto, jak już wcześniej wspomniałem, umieszczanie pytań w bezpośrednim sąsiedztwie, ale też podobnie wyglądających, zwiększa interkorelację wyników.

Gotowe narzędzie w wersji komputerowej może być wykorzystane w badaniu labolatoryjnym, gdzie kontrolowane są warunki zewnętrzne, ale też może być zaprezentowane z wykorzystaniem poczty elektronicznej lub stron www. W przypadku wykorzystania poczty elektronicznej może być ono: 1) częścią wiadomości, wymagającą wypełnienia i odesłania, 2) samodzielnym programem, który przeprowadzi badanie i prześle do badacza wstępnie sformatowane wyniki oraz 3) przekierowaniem na stronę www, która w sposób statyczny (na wzór wersji papier-i-ołówek) lub dynamiczny posłuży do przeprowadzenia badania.

2.3. Podejście dynamiczne – kwestionariusze i testy adaptacyjne

Na bazie coraz większej popularności IRT rozwija się podejście psychometryczne do testowania, wykorzystujące możliwości techniczne oferowane przez współczesną technologię. Pomysły na testy i kwestionariusze, które zmieniają się na bieżąco podczas badania, w zależności od kolejnych odpowiedzi udzielanych przez osoby uczestniczące w badaniach, pojawiały się od początku lat 70 (porównaj: Nowakowska, 1975, s. 174–183; Hornowska, 2007), ale dopiero współczesne komputery stworzyły wystarczające zaplecze techniczne, a internet – odpowiednie środowisko, aby wykorzystać w pełni zalety badania adaptacyjnego.

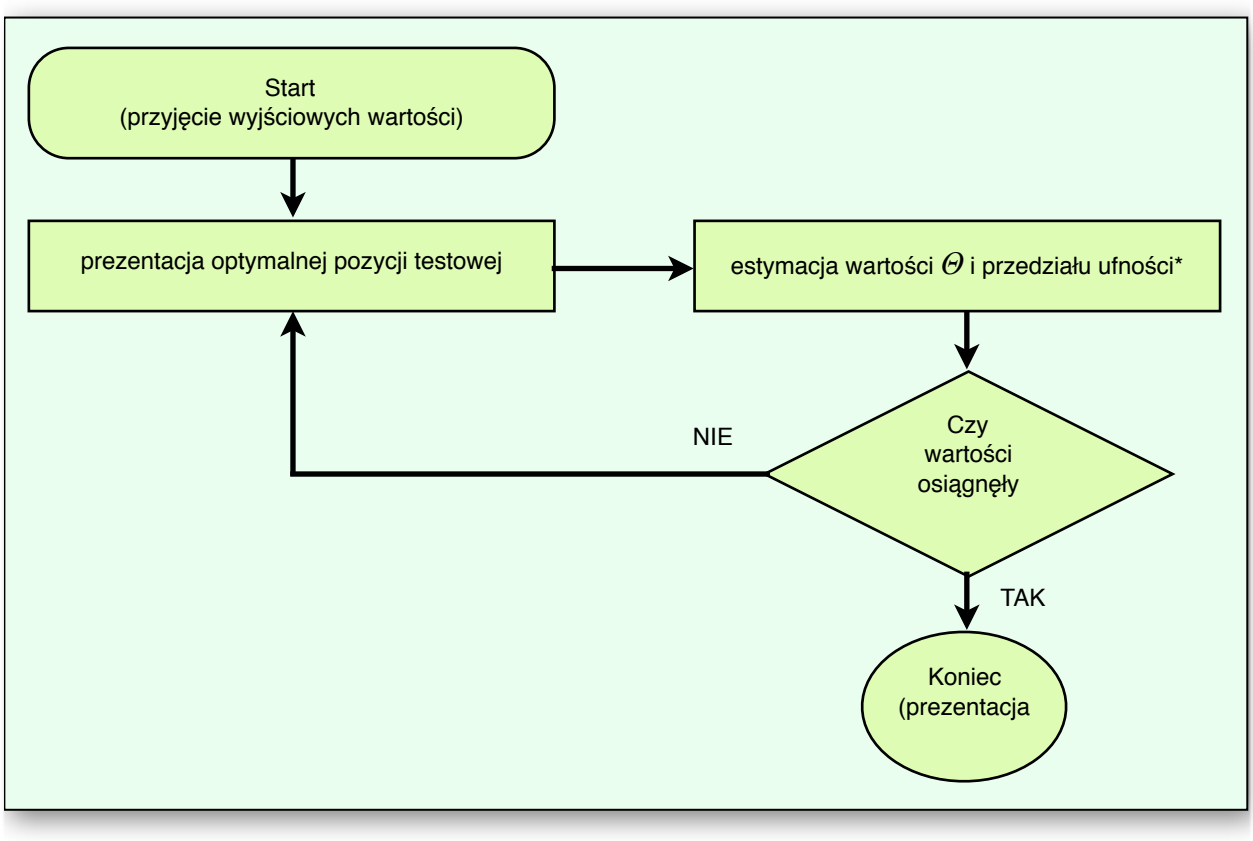
Wykorzystanie teorii odpowiadania na pozycje testowe do zbudowania testu lub kwestionariusza adaptacyjnego wymaga realizacji trzech zadań: 1) oszacowania parametrów poszczególnych pozycji – mocy różnicującej i trudności, 2) przyjęcia określonego modelu i sprawdzenia jego poprawności oraz 3) oszacowania wartości Θ na próbie kalibracyjnej. Powstały w ten sposób bank zadań może zostać wdrożony jako narzędzie badawcze na trzy sposoby (Vispoel, Rocklin i Wang, 1994):

1. FIT (*fixed-item test*) – osoby uczestniczące w badaniu rozwiązują taki sam zestaw pytań,

2. CAT (*computerized-adaptive test*) – osoby uczestniczące w badaniu otrzymują pytania ustalane w oparciu o dotychczasowe odpowiedzi i dobierane tak, aby zmaksymalizować ilość informacji i z jak najmniejszym błędem dokonać estymacji wartości Θ,

3. SAT (*self-adapted test*) – procedura dobierania pytań opiera się na wyborze przez osoby uczestniczącej w badaniu poziomu trudności następnej pozycji.

Dla sposobu CAT procedura badania może być opisana algorytmem przedstawionym na ryc. 2.4.



Rycina 2.4. Schemat algorytmu badania adaptacyjnego
* – jako regułę zatrzymującą procedurę stosuje się także estymację błędu standardowego.
Źródło: opracowanie własne.

Ponieważ w IRT estymowana wartość Θ badanej cechy u osoby uczestniczącej w badaniu (a także każdej pozycji) jest wyrażona na tej samej skali, możliwe jest łatwe szacowanie jednej wielkości przez drugą, poprzez dobieranie poszczególnych pozycji według oszacowanej na danym etapie wartości cechy. Badanie CAT zaczyna się od wyboru najbardziej optymalnej pozycji. Może to się odbywać w oparciu o deklarację poziomu trudności przez osobę uczestniczącą w badaniu albo przez wybór pozycji o średnim poziomie trudności. Po uzyskaniu odpowiedzi dokonuje się estymacji

wartości cechy Θ oraz określa poziom błędu standardowego pomiaru (lub poziom ufności). Dokonuje się tego za pomocą jednej z dwóch procedur: szacowanie współczynnika największej wiarygodności (*maximum likelihood estimation*) lub szacowanie Bayesowskiego (*Bayesian estimation*). Dysponując oszacowaną wartością cechy Θ, *w* następnym kroku badania prezentowana jest taka pozycja, która ma dla tej wartości cechy największą wartość funkcji informacyjnej. Procedura ta powtarzana jest do osiągnięcia zadanego kryterium – liczby zaprezentowanych pozycji lub określonego poziomu błędu standardowego.

Zaletą testowania CAT jest oszacowanie wartości cechy dla wszystkich osób uczestniczących w badaniu z tą samą precyzją, w przeciwieństwie do badań pełną wersją narzędzia (FIT), które słabiej estymują wyniki osób z krańców skal. Wartość cechy lub poziom umiejętności może być estymowany na podstawie różnych zestawów pozycji, co więcej, wyniki te są ze sobą porównywalne. Istotna jest także kwestia długości badania – w badaniach CAT jest możliwe oszacowanie wyniku z taką samą dokładności jak w badaniach typu FIT już przy 40% ich długości (Embretson i Riese, 2000). Dodatkową zaletą, np. w sytuacji egzaminu, jest prezentowanie dla różnych osób odmiennych zestawów zadań, co sprzyja obiektywizowaniu wyniku. Zważywszy, że dla każdej osoby dobierany jest indywidualny zestaw pytań przypomina to ustne udzielanie odpowiedzi, gdzie egzaminator dostosowuje parametry kolejnych pytań na podstawie tych już udzielonych, przy czym metoda ta jest automatyczna i nie generuje dodatkowych kosztów związanych z liczebnością próby.

Wady badania adaptacyjnego związane są głównie z procedurą przygotowania – kalibracji pozycji testowych lub kwestionariuszowych. Jest to proces wymagający dużej próby – według niektórych autorów nawet w granicach 1000 osób (porównaj rozdział 4.4), aby móc osiągnąć rzetelne parametry poszczególnych pozycji. Oznacza to także, iż niemożliwym jest skonstruowanie testu lub kwestionariusza z pozycji, które nie były nigdzie wcześniej prezentowane.

Kolejny problem związany jest z częstością ekspozycji danych pozycji. Dobrze zaplanowana procedura badania umożliwia kontrolę częstości pojawiania się poszczególnych pozycji oraz zapobiega ich zbyt częstej ekspozycji – szczególnie w warunkach testowania grupy ludzi, którzy mogą dzielić się informacjami na temat przebiegu testu i posiadają zbliżony poziom badanej cechy.

Problemem jest także umożliwienie przeglądania i poprawiania odpowiedzi podczas wypełniania testu bądź kwestionariusza – ponieważ całość jako zestaw pozycji

budowana jest na bieżąco, poprawienie jednego z początkowych zadań pociągnęłoby za sobą konieczność zmian w zestawie następujących po nim pozycji. Teoretycznie możliwe byłoby zatem odpowiadanie z premedytacją niepoprawnie, aby uzyskać zestaw łatwych pozycji, a następnie poprawienie ich, by otrzymać bardzo wysoką ocenę (Alfonseca, Rodriguez i Perez, 2007).

2.4. Porównanie jakości danych uzyskiwanych w papierowych i elektronicznych wersjach testów i kwestionariuszy

Analizując przytoczone wcześniej uwarunkowania podmiotowe oraz te związane z formą badań wykorzystujących komputery i/lub internet, pojawiają się wątpliwości co do jakości danych uzyskanych w takim badaniu. Jakość ta będzie dobra, jeśli badacz wykaże się wysoką świadomością metodologiczną podczas konstruowania narzędzia badawczego. Wpływ na późniejsze wyniki mają wszystkie decyzje podjęte podczas procesu konstruowania narzędzia. Przykładowo, jak donoszą Smyth, Dillman, Christian i Stern (2006), zmiana typu pytań z jednokrotnego na wielokrotnego wyboru istotnie zwiększa czas poświęcony na odpowiadanie oraz zmniejsza liczbę pytań bez odpowiedzi. Jak pokazują z kolei badania Krosnick z zespołem (2002) używanie pozycji „nie wiem", „nie dotyczy" wśród możliwych odpowiedzi do wyboru, mające zwiększyć komfort uczestniczących w badaniu, zniechęca ich jednak do podjęcia trudu poznawczego niezbędnego do wyrażenia prawdziwej opinii udostępniając „drogę na skróty poprzez kwestionariusz" i zmniejszając jego moc statystyczną. Problemy związane z budową pozycji testowych lub kwestionariuszowych pojawiają się jednak zawsze, bez względu na medium badania: monitor komputera czy też papier, natomiast użycie komputerów może poprawić jakość danych już na etapie ich zbierania.

Walidacja wyników może być przeprowadzana w tle, podczas badania, nadzorując wprowadzane wyniki i przyczyniając się znacznie do skrócenia etapu poprawiania uzyskanych danych. Błędy, jakim można zapobiegać w sposób automatyczny to:

- błędy wypełnienia,
- złej odpowiedzi,
- odpowiedzi na inne pytanie,
- odpowiedzi przypadkowych,
- braki odpowiedzi.

Błędy wypełnienia polegają najczęściej na podawaniu odpowiedzi spoza zakresu pytania lub zaznaczeniu innej liczby opcji niż wymagana. Błędy złej odpowiedzi pojawiają się wtedy, gdy osoby uczestniczące w badaniu podają formalnie poprawną

odpowiedź, lecz nie jest ona prawidłowa w kontekście pytania. Czasami, szczególnie gdy wiele pytań i odpowiedzi ma taką samą konstrukcję lub niejasne jest przyporządkowanie pozycji odpowiedzi do pytań, zdarza się przesunięcie między pytaniem a odpowiedziami i w kwestionariuszu pojawia się regularny błąd polegający na odpowiadaniu na pytanie w miejscu odpowiedzi na sąsiednie. Natomiast błędy odpowiedzi przypadkowych pojawiają się w sytuacji, gdy osoby uczestniczące w badaniu nie są zainteresowane udzielaniem prawdziwych odpowiedzi i zaznaczają je w sposób losowy. No i wreszcie ostatni rodzaj błędu związany z brakami odpowiedzi – zdarza się, iż są to nieintencjonalne ominięcia odpowiedzi wynikające z różnych powodów, np. losowej kolejności odpowiadania, pozostawienia na później pytania sprawiającego trudność, itp. Błędom tym nie da się zapobiec korzystając z narzędzi w wersji papier-i-ołówek, choć buduje się kwestionariusze i testy w ten sposób, aby zminimalizować ich liczbę. Tym nie mniej dane po badaniu, na etapie obróbki (która też może być źródłem błędów) muszą zostać sprawdzone pod względem integralności.

W wersji elektronicznej badacz może tak zaprojektować narzędzie badawcze, że walidacja danych odbywać się będzie automatycznie, lecz konsekwencją jest informowanie osoby uczestniczącej w badaniu o popełnianych błędach na bieżąco, co może wpływać na poziom frustracji i zwiększać odsetek osób porzucających badanie.

Kwestionariusz lub test w wersji internetowej może być wyposażony w mechanizm zgłaszania braków danych lub niepoprawności wprowadzonych danych, już podczas udzielania odpowiedzi. Zarówno dla wersji jednostronicowej, przewijanej, jak i dla takiej, gdzie pytania prezentowane są pojedynczo, osoba uczestnicząca w badaniach albo nie będzie miała możliwości udzielenia złej odpowiedzi, albo po jej udzieleniu będzie o tym fakcie natychmiast informowana. Wymaganie poprawnej odpowiedzi z jednej strony zmniejsza znacząco ilość pomyłek i błędów (Mooney, Rogers i Trunxo, 2003, za: Peytchev i Crawford, 2005, s. 240), ale z drugiej powoduje czasami wprowadzanie przez osoby badane losowego ciągu po to tylko, aby przejść do następnego pytania, lub zapobiec pojawianiu się informacji o błędzie (DeRouvray i Couper, 2002).

Kolejnym mechanizmem, który może poprawiać jakość danych zebranych w komputerowym narzędziu badawczym jest ograniczenie możliwości wypełniania pól odpowiedzi do konkretnych typów danych i/lub ich formatu. Szczególnie przydatne jest to przy pytaniach dotyczących dat (np. daty urodzin), ale nie tylko. Ograniczenie zawartości pól do danych o określonym formacie zmniejsza liczbę błędów złej odpowiedzi. Sama wizualna konstrukcja narzędzia też ma wpływ na poprawność wyników. Couper, Traugott i Lamias (2001, za: Peytchev i Couper, 2005) wykazali, że

nawet wielkość pola do wpisania odpowiedzi ma znaczenie: duże puste pole jest często interpretowane przez osoby uczestniczące w badaniach jako wskazówka do dłuższych wypowiedzi, nie zawsze zgodnie z intencjami badaczy.

Korzyścią zbierania danych od razu w postaci numerycznej jest możliwość wykorzystania wielu rodzajów automatycznej walidacji. Często możliwe jest określenie zakresu prawdopodobnych odpowiedzi, np. w pytaniu o wiek wartości ujemne i większe niż 150 będą błędne. Podobnie pozycja narzędzia badawczego z użyciem rankingu odpowiedzi, czy sumy (np. do 100%), może być wyposażona w automatyczną kontrolę poprawności.

Cennej informacji o wynikach dostarcza wiedza na temat czasu udzielania odpowiedzi – na podstawie rozkładu czasu w całej grupie osób uczestniczących w badaniach można wyciągać wnioski o udzielaniu przez osoby pobieżnych odpowiedzi (zbyt krótki czas odpowiedzi na tle grupy) i/lub dystraktorach (przerwy w odpowiadaniu).

Reasumując, walidację odpowiedzi można wprowadzić poprzez:

- identyfikację użytkownika,
- wymaganie odpowiedzi,
- określenie typu i formatu danych,
- określenie zakresu danych.

Drugą grupę stanowią zabiegi wpływające na jakość danych, takie jak:

- zbadanie niekonsekwencji (niezgodności w odpowiedziach na pytanie przedstawione co najmniej dwa razy, ale np. w zmienionej formie),
- przeprowadzanie obliczeń (czas, ranking, sumy odpowiedzi).

Sposoby walidacji można także podzielić na inne dwie kategorie: 1) zapobiegające powstawaniu błędów (np. format testu lub kwestionariusza uniemożliwiający udzielenie niepoprawnej odpowiedzi) i 2) poprawiające dane wpisywane przez osoby uczestniczące w badaniach (informujące o błędach lub niedopasowaniu do oczekiwań badaczy).

Mimo ciągłego rozwoju i postępu technologicznego, komputery w psychologii nie są w stanie w pełni zastąpić „żywego” psychologa z jego wiedzą, intuicją i empatią. Takie projekty jak wykorzystywanie gier czy wirtualnej rzeczywistości do oddziaływań psychologicznych są ciągle mało popularne i trudne w realizacji (Paluchowski, 2007). Mimo to za pomocą wirtualnej rzeczywistości próbuje się przeprowadzać badania nad zjawiskiem świadomości, a także leczyć fobie, zaburzenia lękowe, ból i podnosić samoocenę (Gerardi, Cukor, Difede i in., 2010). Korzystając z gier komputerowych

prowadzi się dynamiczną obserwację ludzi, służący lepszej diagnozie specyficznych zachowań (Ceranoglu, 2010).

W ciągu ostatnich kilkunastu lat obserwujemy narastającą akceptację komputerów i internetu w życiu codziennym ludzi (przyrost użytkowników internetu w Polsce w latach 2000–2010 wyniósł ponad 700%) a także w praktyce badawczej. Przyzwyczajenie się do nowego medium, a także jego rosnące możliwości powodują, że ludzie coraz częściej i coraz łatwiej wykorzystują internet aby prowadzić działalność biznesową, towarzyską, rozrywkową i naukową. Staje się on najpopularniejszym środkiem komunikacji, nie tracąc jednocześnie charakteru bycia pewnego rodzaju środowiskiem, w którym co prawda działalność ludzi określa cele, ale także konstytuuje środki do ich osiągania. Większa obecność technologii w postaci komputerów i innych urządzeń, a także powszechniejszy dostęp do internetu, powoduje rosnące zainteresowanie kręgów naukowych wykorzystaniem tych możliwości, co z kolei prowadzi do rosnącego zapotrzebowania na krótkie, zajmujące niewiele czasu, a jednocześnie rzetelne i trafne narzędzia badawcze do pomiaru psychologicznego. Narzędzia, które będąc ekonomicznymi (z punktu widzenia osoby biorącej udział w badaniu) nie przestaną być etycznymi i dostarczą badaczom tak pożądanych danych.

2.4.1. Metaanaliza różnic między wynikami papierowych i komputerowych wersji narzędzi badawczych

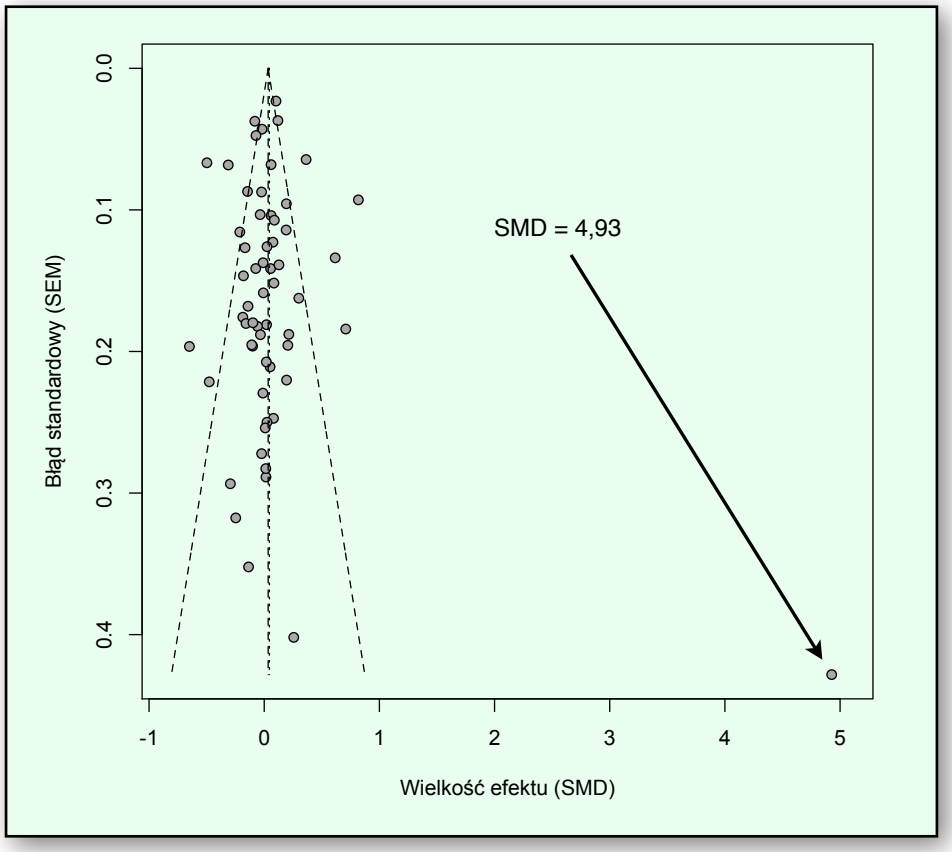
Na potwierdzenie tezy o równoważność wyników komputerowych wersji testów i kwestionariuszy psychologicznych skonstruowanych wcześniej w wersji papier-i-ołówek, sporządziłem przegląd literatury za pomocą meta-analizy. Przeszukałem 7 baz anglojęzycznych, pod kątem artykułów dotyczących porównań wyników uzyskiwanych za pomocą tych samych narzędzi, ale za pośrednictwem różnego medium. Bazy te to: Springerink, Ebsco (PsycARTICLES, ERIC, PsycINFO oraz Academic Search Complete), Scholar.google.com, Elsevier, ISI Web of Knowledge, Scopus i ScienceDirect. Wyszukiwałem zarówno w tytułach, jak i w treści następującą frazę: *(equivalence OR compare) AND (paper-and-pencil or traditional) AND (online OR internet OR web based)* i otrzymałem 29834 wyniki. Ponieważ algorytmy wyszukiwawcze w bazach danych zwracają rekordy posortowane według malejącego podobieństwa do zapytania, daną bazę przestawałem przeglądać, jeśli w kolejnych artykułach nie było żadnej wzmianki na temat *mode effect* przez kolejne 30 rekordów (dla Elsevier i ScienceDirect zastosowano inne reguły – por. tabela 2.1). Wyniki z trzech baz bibliograficznych przejrzałem w całości.

Tabela 2.1. Liczba rekordów w bazach związanych z porównaniem wersji elektronicznych i papierowych narzędzi kwestionaruszowych

Baza	liczba zwróconych rekordów	liczba przejrzanych rekordów	reguła zatrzymania
SpringerLink	409	80	30 bez trafienia
Ebsco (PsycArticles, PsycInfo, ERIC, Academic Search Complete)	59	59	–
Scholar.google.com	16000	90	30 bez trafienia
Elsevier	1319	140	podobieństwo <3
ISI Web of Knowledge	40	40	–
Scopus	20	20	–
ScienceDirect	11987	550	100 bez trafienia
Razem	**29834**	**979**	

Źródło: badania własne.

Z 979 wyszukanych do analiz rekordów zakwalifikowałem 102 artykuły ze względu na zawartość merytoryczną. Spośród tych artykułów 6 wykluczyłem ponieważ były to meta-analizy dotyczące różnych obszarów. Opierały się one kolejno na publikacjach z lat: 1977–1992 – meta-analiza 29 artykułów dotyczących zdolności poznawczych, 1976–1996 – meta-analiza 51 artykułów dotyczący zdolności, 1991–2006 – meta-analiza 73 artykułów dotyczących jakości życia pacjentów, 1998–2006 – meta-analiza 39 artykułów dotyczących liczby zwrotów ankiet, 1989–2005 – meta-analiza 44 artykułów dotyczących zdolności matematycznych, 1974–1996 – meta-analiza 12 artykułów dotyczących wyników w MMPI. Kolejne 11 artykułów zostało wykluczonych, ponieważ były to rozważania teoretyczne, a kolejne 26 artykułów nie zawierało danych niezbędnych do włączenia do meta-analizy. Ostatecznie wybrałem 59 artykułów, które zawierały dane na temat 480 porównań i były opublikowane w latach 1988–2010. Ich charakterystykę zawiera tabela zamieszczona w załączniku nr 3 (Artykuły użyte w meta-analizie), zaś w tabeli 2.2 przedstawiono charakterystykę skumulowanej próby i zebranych wielkości efektów.



Ryc. 2.1. Podsumowanie wielkości efektów analizowanych wyników badań. Źródło: badania własne.

Statystyka Q testu homogeniczności łącznie dla wszystkich porównań wynosi 6250 ($p < 0{,}001$, $df = 479$), co wskazuje na heterogeniczność wyników i sugeruje użycie modelu losowego lub mieszanego. Średnia wielkość efektu wynosi 0,045 w 95% przedziale ufności od -0,029 do 0,118, co pozwala stwierdzić brak różnic w wynikach uzyskanych za pośrednictwem komputerów oraz papieru i ołówka (por. ryc. 2.1 i załącznik 3.2).

Tabela 2.2. Charakterystyka próby (n=59) i porównań (n=480)

Charakterystyka	Liczba badań		Liczba porównań	
	N	%	N	%
Rok publikacji				
• 1988 ~ 1999	16	27,2	90	18,8
• 2000 ~ 2005	15	25,5	126	26,2
• 2006 ~ 2007	15	25,5	96	20,0
• 2008 ~ 2010	13	22,1	168	35,0
Źródło				
• artykuły opublikowane	55	93,2	463	96,5
• wystąpienia konferencyjne	2	3,4	14	2,9
• raporty nieopublikowane	2	3,4	3	0,6
Próby				
• uczniowie	1	1,7	3	0,6
• studenci	37	62,7	315	65,6
• dorośli	5	8,5	53	11,0
• ochotnicy	3	5,1	22	4,6
• weterani	1	1,7	18	3,8
• więźniowie	1	1,7	8	1,7
Wielkość próby				
• <40			155	32,3
• 40–80			95	19,8
• 80–150			58	12,1
• >150			172	35,8
Typ testu lub kwestionariusza				
• power / speed	17 / 1	28,9 / 1,7		
• CAT / CBT	2 / 41	3,4 / 69,7		
• własny / standaryzowany	9 / 53	15,3 / 90,1		
Plan badawczy				
• powtarzane pomiary (średnio po 26 dniach)			177	36,9
• losowy dobór do grup			127	26,5
• równoważenie kolejności			28	5,8
• kontrola kolejności			175	36,5
Obszar badawczy				
• depresja	6	10,2		
• inteligencja	3	5,1		
• jakość życia	16	27,1		
• językowy	4	6,8		
• lęk / strach	6	10,2		
• osobowość	12	20,3		
• wiedza / umiejętności	10	16,9		

Źródło: badania własne.

Jako moderatory wielkości efektu uwzględniono następujące zmienne: badanie, rok badania (jeśli nie podano w artykule daty badania, przyjmowany był rok publikacji), płeć osób uczestniczących w badaniach, rodzaj planu badawczego, kolejność dla powtarzanych pomiarów oraz obszar tematyczny badania. Na podstawie

jednoczynnikowej analizy wariancji w planie dla grup niezależnych uzyskano istotny statystycznie wynik na poziomie $p < 0,05$ dla czynnika: rok badania (por tabela 2.6).

Tabela 2.6. Wyniki analizy wpływu moderatorów na średnią wielkość efektów

	SMD	SEM	Z	p	95%CI	
					dolny	górny
badanie	0.0015	0.0013	1.0946	0.2737	-0.0012	0.0041
rok	**0.0112**	**0.0045**	**2.4744**	**0.0133**	**0.0023**	**0.0200**
kobiety	0.0968	0.0729	1.3286	0.1840	-0.0460	0.2397
mężczyźni	0.0661	0.1013	0.6523	0.5142	-0.1325	0.2647
plan badawczy (1 = losowy)	0.0709	0.0517	1.3705	0.1705	-0.0305	0.1723
kolejność (1 = pierwszy PP)	-0.0095	0.0576	-0.1641	0.8696	-0.1223	0.1034
kolejność (1 = pierwszy OL)	-0.0620	0.0611	-1.0148	0.3102	-0.1819	0.0578
obszar badania	0.0049	0.0112	0.4330	0.6650	-0.0171	0.0268

PP – papier i ołówek, OL – wersja komputerowa, SMD – standaryzowana średnia różnica, SEM – błąd standardowy, Z – wystandaryzowana wartość statystyki testu różnic, p – poziom istotności, 95%CI – granice przedziału ufności. Źródło: badania własne.

Przeprowadzone porównania *post hoc* za pomocą testu HSD Tukey'a wykazały, iż powyższa różnica spowodowana jest silnym odstawaniem wyników przedstawionych tylko w jednym artykule z 2003 roku – wielkość efektu SMD wynosiła w tym jednym przypadku 4,93 (por. ryc. 2.1). Po wykluczeniu tych doniesień, jako nietypowych uogólniona wielkość efektu dla 58 badań wyniosła 0,02 (95% przedział ufności: -0,05 – +0,09), co ostatecznie pozwala przyjąć ekwiwalentność narzędzi badawczych używanych równolegle w wersjach papier-i-ołówek oraz elektronicznej.

W świetle wykazanego braku różnic wynikających z *mode effect* (porównaj także Mead i Drasgow, 1993; Finger i Ones , 1999) uzasadnione jest obecnie projektowanie badań kwestionariuszowych i testowych z użyciem komputerów i internetu. Wniosek ten, moim zdaniem, dotyczy nie tylko badań psychologicznych, ale jest aktualny wszędzie tam, gdzie korzysta się z badań za pomocą testu lub kwestionariusza. Pozwala to wykorzystywać wszystkie zalety badań z użyciem nowych technologii zwiększając ich trafność w oparciu o teorię odpowiadania na pozycje testowe (IRT), jednocześnie upraszczając badanie z punktu widzenia osoby w nim uczestniczącej.

Opierając się na tych wynikach, można pójść dalej i postawić tezę o prawdopodobnej równoległości wyników uzyskiwanych także w wersjach adaptacyjnych, co będzie weryfikowane w części empirycznej tej pracy.

Rozdział 3. Modele powiązań zmiennych. Problemy i hipotezy badawcze

Część zawartych w pracy analiz oparta jest na metodzie symulacji łańcuchów Markowa Monte Carlo (*Markow Chain Monte Carlo* – MCMC), która wymaga określenia zmiennych zależnych i niezależnych (Harwell i inni, 1996). Zmiennymi niezależnymi są w przypadku tej pracy parametry użyte podczas symulacji: a – moc różnicująca danej pozycji, b – jej poziom trudności oraz n – wielkość próby.

Jako zmienne zależne przyjęto wielkość błędu SEM oraz parametry rozkładów: skośność i kurtozę.

W rozdziale tym przedstawię definicje przyjętych właściwości statystycznych oraz hipotezy badawcze.

3.1. Zmienne niezależne

3.1.1. Moc różnicująca pozycji testowych lub kwestionariuszowych

Zarówno w modelach dychotomicznych, jak i politomicznych, występuje pojęcie mocy różnicującej dla danej pozycji testowej lub kwestionariuszowej. Określa ono zdolność danej pozycji do rozróżniania osób o różnej wartości Θ poprzez przypisywanie im innych poziomów odpowiedzi. W symulacyjnych analizach modeli IRT można wyróżnić dwa podejścia. Pierwsze polega na ustaleniu pewnej stałej wartości tego współczynnika dla całej symulacji opartej na empirycznych lub teoretycznych przesłankach. Drugie podejście rezygnuje z tego uproszczenia. Zważywszy, że podstawowym celem przeprowadzonych symulacji jest m.in. osiągnięcie jak najwyższego poziomu trafności zewnętrznej – lepszym rozwiązaniem niż przyjmowanie stałych wartości wydaje się losowanie ich z określonego zakresu. Baker (2001) zaproponował kategoryzację poziomu mocy różnicującej według następujących progów: dla wartości z przedziału 0,01–0,24 – bardzo niska; 0,25–0,64 –niska; 0,65–1,34 – średnia; 1,35–1,69 – wysoka i > 1,70 – bardzo wysoka. Do analiz sporządzonych w niniejszej pracy zdecydowano się przyjąć trzy zakresy reprezentujące poziomy niski, średni oraz wysoki, w których to zakresach będą losowo wyznaczane wartości współczynnika mocy różnicującej dla danych pozycji określane jako łatwe, przeciętnie trudne oraz trudne. Zrezygnowano ze skrajnych przedziałów z kategoryzacji Bakera, ponieważ pozycje o bardzo niskiej mocy różnicującej są nieinteresujące z punktu widzenia konstruowania narzędzi badawczych, zaś pozycje o bardzo wysokiej mocy różnicującej nie zdarzają się w nich zbyt często (por. Reise i Waller, 2003).

3.1.2. Poziom trudności pozycji testowych lub kwestionariuszowych

Dychotomiczne modele IRT posiadają pojedynczy parametr odzwierciedlający poziom trudności, który to parametr związany jest z prawdopodobieństwem udzielenia poprawnej odpowiedzi przez osobę uczestniczącą w badaniu w oparciu o wartość jej cechy latentnej. Parametr b przyjmuje wartość Θ w tym punkcie krzywej logitu, gdzie prawdopodobieństwo sukcesu (poprawnej odpowiedzi) wynosi dokładnie 50% (przy zerowym poziomie zgadywalności c i niedbałości d).

Dla modeli politomicznych IRT istnieje j*-1* parametrów b_j (j – liczba kategorii odpowiedzi), gdzie osoba z 50% prawdopodobieństwem może udzielić danej odpowiedzi (kategoria 1, 2, 3, … j) przy danej wartości Θ. Poziomy trudności dla poszczególnych kategorii uporządkowane są narastająco, odpowiadając rosnącemu porządkowi prawdopodobieństwa uzyskiwania odpowiedzi wraz ze wzrostem natężenia cechy latentnej.

Opierając się na badaniach przeprowadzonych przez Kang i Waller (2005) założono trzy poziomy trudności dla symulowanych pozycji: łatwy, średni i trudny. Dla zmaksymalizowania trafności zewnętrznej dla każdej pozycji losowano na podstawie wyniku generatora liczb pseudolosowych wartość b_1 z przedziałów łatwego: <-1,5; 0>, średniego: <-0,5; 1> lub trudnego: <1; 2,5>. Dla modelu politomicznego do wylosowanej dla danej pozycji pierwszej wartości b_1 dodawano sukcesywnie 0,7 otrzymując uporządkowane b_2 i b_3.

Zakładane parametry dla symulacji przedstawia tabela 3.1:

Tabela 3.1. Zestawy parametrów używanych w symulacjach

	model dychotomiczny (2PL)								
a	niski: 0,25–0,64			przeciętny: 0,65–1,34			wysoki: 1,35–1,69		
b	niski: (-1.5;0)	średni: (0;1)	wysoki: (1;2.5)	niski: (-1.5;0)	średni: (0;1)	wysoki: (1;2.5)	niski: (-1.5;0)	średni: (0;1)	wysoki: (1;2.5)
	model politomiczny (GRM)								
a	niski: 0,25–0,64			przeciętny: 0,65–1,34			wysoki: 1,35–1,69		
b_1	niski: (-1.5;0)	średni: (0;1)	wysoki: (1;2.5)	niski: (-1.5;0)	średni: (0;1)	wysoki: (1;2.5)	niski: (-1.5;0)	średni: (0;1)	wysoki: (1;2.5)

Źródło: badania własne.

3.1.3. Wielkość próby

W literaturze przedmiotu można spotkać wiele definicji odpowiedniej wielkości próby dla modelowania w oparciu o IRT. Często założenia co do wielkości próby są przyjmowane arbitralnie, np. co najmniej 100 osób (Kline, 1979), lub 200 (Guilford, 1954), lub 250 (Cattell, 1978). Comrey i Lee (1992) zaproponowali nawet skalę, gdzie zdefiniowali pojęcie próby słabej (*week* – 100 osób), przyzwoitej (*decent* – 200 osób), dobrej (*good* – 300 osób), bardzo dobrej (*very good* – 500 osób) oraz doskonałej (*exellent* –1000 osób). Pierwszą analizę wpływu wielkości próby na dokładność parametrów można spotkać w pracy Lorda i Novicka (1968). Dla danych empirycznych ustalili oni (z dużym marginesem niepewności), iż dla testu składającego się z co najmniej 50 pozycji błąd maleje do akceptowanego poziomu dla prób powyżej 1000 osób. Pierwsze podejście z wykorzystaniem symulacji MCMC dla modelu 3PL zastosowali Hulin, Lissak i Drasgow (1982) – wykazali oni, że dla testu składającego się z 60 pozycji i prób w wielkości 200, 500, 1000 oraz 2000 osób wielkość $RMSE$ (*rooted mean squared error*) wynosi odpowiednio 0,06; 0,05; 0,04 i 0,03. Podobnie w jednej z częściej cytowanych prac Gao i Chen (2005), w której dla estymacji parametrów IRT wielkość próby potraktowana została jako czynnik, autorzy przyjęli próby w wielkości 100, 500 oraz 2000 osób. Dla testu składającego się z 60 pozycji i największej próby, $RMSE$ został przez nich oszacowany na poziomie 0,12; dla porównania w statystykach dopasowania modelu za akceptowalny uznaje się poziom $RMSE < 0{,}05$.

Dla modelu GRM podobne symulacje przeprowadzili Reise oraz Yu (1990). Próby liczące 250, 500, 1000 i 2000 osób pozwoliły na zdefiniowanie zalecanej minimalnej wielkości próby na poziomie 500 osób ($RMSE$ = 0,08), rekomendując 1000 osób jako próbę odpowiednią do dokładnego oszacowania parametrów modelu ($RMSE < 0{,}05$).

Nadmienić należy, iż wielkość próby jest ściśle zależna od wielkości efektu, który ma zostać wykryty. Zbyt wielka próba prowadzi do przeszacowania estymowanych parametrów, co rodzi niebezpieczeństwo stwierdzenia trywialnych zależności – zbyt mała próba nie pozwala wykryć zależności, które być może istnieją w populacji (Hays, 1973, s. 422-424). Wiadomo też, że modele z mniejszą liczbą parametrów wymagają mniejszych prób. Dla modeli Rasch'a wystarczające są takie o liczebności 100 osób (Wright i Linacre, 1994). Natomiast Ostini i Nering (2006) wykazali, że stabilne wartości parametrów IRT można uzyskać już przy próbach 250 osobowych. I mimo, że według Tsutakawy i Johnsona (1990) dla kalibracji pozycji testowych dla modeli wielo-parametrycznych należy użyć próby o wielkości około 500 osób, to wielu autorów

wykazuje, iż do tego celu wystarczające są próby około 200-stu osobowe (Orlando i Marshall, 2002; Thissen, Steinberg i Gerard, 1986).

Oczywiście, im większe próby, tym mniejsze błędy standardowe oszacowanych parametrów, jednakże w kalibracji pozycji według IRT równie ważne co wielkość próby jest rozkład wyników osób uczestniczących w badaniach w sposób równomierny wzdłuż wartości zmiennej latentnej. Duża próba, ale o wartości cechy skupionej wokół jednego wyniku dostarczy bardzo dobrych oszacowań dla tej właśnie wartości, a słabo obsadzone krańce przedziałów będą obciążone dużym błędem standardowym.

3.2. Zmienne zależne

W przypadku porównywania efektów „działania" testów bądź kwestionariuszy samo porównanie wyników średnich może nie dać pełnego obrazu. Z tego względu jako zmienne zależne wybrano popularne miary kształtu rozkładu wyników – skośność i kurtozę. Ponadto zbadano obciążenie wyniku poszczególnych wersji narzędzi błędem pomiarowym, obliczając w tym celu błąd standardowy pomiaru.

3.2.1. Błąd standardowy pomiaru

Dla obliczenia błędu rozkładu wyników w badanych wersjach przyjęto klasyczną definicję błędu standardowego pomiaru opartą o współczynnik rzetelności według wzoru 1.5 (Ferguson i Takane, 1997, s. 499):

$$SEM = S_X\sqrt{1 - r_{tt}}.$$

3.2.2. Skośność rozkładu wyników

Skośność (ibidem, s. 48) „...określa symetryczność bądź niesymetryczność rozkładu liczebności. Jeśli rozkład [wyników] jest niesymetryczny i istnieje tendencja do skupiania się większych liczebności w zakresie..." mniejszej lub większej wartości zmiennej, to mówimy, że rozkład jest skośny. Skośność jest też określana jako trzeci moment średniej i obliczana według wzoru:

$$g_1 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3 \qquad (3.2).$$

3.2.3. Kurtoza rozkładu wyników

Kurtoza (ibidem, s. 48) „...określa płaskość bądź stromość jednego rozkładu w stosunku do innego rozkładu", czyli inaczej mówiąc względną gęstość rozkładu

wyników. Kurtoza jest także definiowana jako czwarty moment średniej i wyrażona jest wzorem:

$$g_2 = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{N} \left(\frac{x_i - \bar{x}}{s} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)} \qquad (3.3).$$

3.3. Problemy badawcze

Mimo popularności i metodologicznej użyteczności symulacji opartych o metodę łańcuchów Markowa Monte Carlo sama symulacja nie wystarczy, aby wyniki były ważne i użyteczne. Harwell i inni (1996, s. 105) określili cztery niezbędne kroki warunkujące powodzenie w używaniu MCMC:

1. Sformułowanie problemu.

2. Zaplanowanie badania poprzez określenie zmiennych niezależnych i zależnych, planu badawczego, liczby powtórzeń oraz przyjętych modeli (w tym przypadku modeli IRT).

3. Napisanie lub wykorzystanie istniejących programów komputerowych do symulowania odpowiedzi i obliczania parametrów.

4. Analiza wyników.

Uwzględniając powyższe kroki sformułowałem cztery szczegółowe pytania badawcze, na które chcę znaleźć odpowiedź, a które odnoszą się do sfery przydatności probabilistycznej teorii odpowiadania na pozycje testowe do konstruowania skróconych wersji testów i kwestionariuszy psychologicznych.

Problem 1:

Jaki jest wpływ wielkości próby na szacowane parametry pozycji według modeli IRT?

Czy zwiększanie próby przynosi same korzyści w postaci zmniejszania się błędu standardowego, czy też wzrost wariancji będzie obniżał dokładność wyników? w którym momencie zwiększanie liczebności próby kalibracyjnej przestaje być uzasadnione – koszty związane z przebadaniem kolejnych osób przewyższają korzyści związane z uzyskiwaniem coraz dokładniejszych parametrów danych pozycji?

Problem 2:

Jaki jest wpływ doboru pozycji do skróconych wersji narzędzi badawczych na dokładność szacowanych wyników cechy latentnej?

Jest to problem opisujący sytuację badania tej samej osoby, różnymi wersjami testu lub kwestionariusza. Poszczególne pozycje narzędzia badaczego charakteryzuje

różny poziom trudności – tego powodu kolejne wersje będą się różnić ogólnym poziomem trudności. Dysponując próbą o znanym rozkładzie cechy chcę zbadać ewentualną różnicę między wynikami narzędzia badawczego o różnym stopniu trudności. Rozstrzygnięcie tego problemu i stwierdzenie braku różnic, pozwoliłoby stosować plany badawcze z powtarzanymi pomiarami, gdzie ekwiwalentne testy lub kwestionariusze z punktu widzenia badacza byłyby różne z punktu widzenia osoby uczestniczącej w badaniu.

Problem 3:

Na ile wyniki uzyskane za pomocą skróconych według IRT wersji narzędzi badawczych są porównywalne do wyników uzyskiwanych za pomocą wersji skróconych według np. konfirmacyjnej analizy czynnikowej (CFA) i wielo-czynnikowej analizy regresji (MR)?

Wiadomo, że skrócone wersje testów lub kwestionariuszy redukując koszt i oszczędzając czas zawsze były brane pod uwagę w praktyce psychologicznej (por. rozdział 1.2). Nawet skala *Stanford-Binet Intelligence Scale* – jedno z pierwszych narzędzi do pomiaru inteligencji – miała swoją wersję skróconą opublikowaną w 1917 roku (Silverstein, 1990, s. 3). Czy IRT stawiając przed badaczem większe wymagania w postaci wymogu dużych prób kalibracyjnych i skomplikowanego aparatu statystycznego jednocześnie pozwala uzyskać lepsze – bliższe wyniku prawdziwego – rezultaty, niż posłużenie się prostszą w zastosowaniu analizą regresji lub analizą czynnikową (ekwiwalentność parametrów IRT i CFA można prześledzić w pracy Konarskiego, 2004).

Problem 4:

Jaki jest wpływ formy narzędzi badawczych na wyniki?

W rozdziale drugim poruszona została kwestia wykorzystywania komputerów do przeprowadzania badań zarówno lokalnie (w laboratoriach), jak i przez internet. Rodzi się pytanie: Czy brak wpływu medium obserwowany dla pełnej wersji (por. rozdział 2.4.1) będzie również dotyczył wersji skróconych? Czy wersja adaptacyjna narzędzia dostarczy wyniki porównywalne z wersją pełną i/lub skróconą, ale np. wypełnianą w sposób tradycyjny?

3.4. Hipotezy

W odpowiedzi na wyłonione powyżej problemy sformułowano hipotezy dotyczące problematyki konstruowania skróconych wersji kwestionariuszy i testów w oparciu o IRT.

3.4.1. Wpływ wielkości próby na estymację parametrów modeli IRT

Jak zmieniają się parametry wraz ze wzrostem liczebności próby, a w szczególności jaka jest wielkość graniczna próby, powyżej której zwiększanie liczebności nie prowadzi do znaczącej poprawy estymacji parametrów w modelach dwu- i wielokategorialnych? Zakładam, że wraz ze wzrostem wielkości próby malał będzie błąd pomiaru, a parametry modeli IRT będą stałe powyżej pewnej, granicznej wielkości. Jednocześnie z uwagi na fakt, że skracanie testu nie pozostaje bez znaczenia dla osiąganego wyniku, zbadany zostanie wpływ długości wersji skróconej z kontrolą parametrów pozycji tej wersji. Aby wyeliminować maksymalnie wpływ czynników osobowych, analiza zostanie przeprowadzona z wykorzystaniem metody symulacji Monte Carlo. Zakładam, że stwierdzony zostanie brak związku między długością wersji skróconej a wynikami osób uczestniczących w badaniach, jednocześnie im dłuższa będzie wersja skrócona tym mniejszy błąd pomiaru powinien być obserwowany.

3.4.2. Wpływ parametrów modeli IRT na estymację wyników kwestionariuszy i testów

Czy konstruowanie skróconych wersji narzędzi badawczych w oparciu o model IRT wpływa na otrzymywane ostatecznie wyniki i jeśli tak, to w jaki sposób?

Stosując IRT do testowania adaptacyjnego osiąga się dużą oszczędność czasu związaną ze zmniejszoną liczbą wymaganych pozycji testowych, potrzebnych do estymowania poziomu mierzonej cechy osoby uczestniczącej w badaniu. Oszczędność jest największa w narzędziach klinicznych używanych do diagnozy i wynosi od 70 do 80% pozycji oryginalnej wersji (Walter i inni, 2007). Mniejsze wartości obserwuje się dla testów poznawczych - średnio 50% (Egberink i Veldkamp, 2007), a najmniejsze przy testach inteligencji - około 40% (Walter i Holling, 2008). Przyjmując zatem minimalny poziom skrócenia narzędzia badawczego na 40% zakładam zgodnie z teorią, że bez względu na wybrany zestaw pozycji i ich parametry, wyniki przeliczone powinny być stałe, natomiast wybranie pozycji skrajnie łatwych lub skrajnie trudnych powinno przynieść największe błędy oszacowania wyniku prawdziwego.

3.4.3. Jakość wersji skróconych w oparciu o IRT versus inne metody skracania

Tekane i deLeeuw oraz Bartholomev (za Konarski, 2004, s. 8) dowiedli, że w sensie formalnym parametry pozycji w IRT odpowiadają parametrom uzyskanym w analizie czynnikowej. Jeśli tak jest, to pojawia się pytanie: Czy skracanie testów

i kwestionariuszy w oparciu o IRT pozwala uzyskać taką samą wersję jak skracanie według CFA? Idąc dalej tym tropem: Czy wybierając pozycje w oparciu o jedną z trzech metod statystycznych (CFA, MR lub IRT) otrzymam różne, czy takie same zestawy pozycji? Zakładając, że te same dane powinny przynosić spójne rezultaty i wykazaną w literaturze przedmiotu ekwiwalentność parametrów analizy czynnikowej oraz mocy i trudności pozycji testowych, spodziewam się, że powinna wystąpić wysoka zgodność przy wyborze pozycji do wersji skróconej bez względu na sposób wyboru.

3.4.4. Zróżnicowanie wyników w kwestionariuszowym badaniu psychologicznym ze względu na sposób przeprowadzenia badania

Podstawową miarą jakości przy tworzeniu skróconej wersji narzędzia badawczego jest uzyskanie wyników wysoce skorelowanych z wynikami wersji pełnej. W kontekście możliwości używania zamiennie różnych skróconych wersji kwestionariuszy i testów, chcę uzyskać odpowiedź na pytanie, czy istnieją różnice w wynikach uzyskiwanych przez osoby uczestniczące w badaniach, związane ze sposobem oraz z medium badania. Jak wiadomo, sposoby badania są równoważne dla pełnych narzędzi badawczych (rozdział 2.4.1), dlatego chcę zbadać, czy taka ekwiwalentność dotyczy też wersji skróconych. Zakładam, że wyniki przeliczone uzyskiwane z różnych form nie będą się od siebie różnić w sposób istotny.

3.5. Założone modele IRT

Do analiz wybrano najpopularniejsze modele IRT. Modele dychotomiczne zakładają zero-jedynkowy wynik dla i-tej pozycji testowej, zaś modele politomiczne zakładają wielopoziomowość odpowiedzi jako wskaźnika natężenia cechy latentnej. Z tego powodu analizy zaplanowane w tej pracy muszą biec dwutorowo – osobno dla wybranego testu wiedzy (możliwe wyniki: prawda–fałsz) i osobno dla kwestionariusza cechy na przykładzie temperamentu (natężenie cechy latentnej na skali co najmniej porządkowej).

3.5.1. Model dychotomiczny, model trójparametryczny (3PL)

Model 3PL jest najbardziej ogólny w klasy modeli dwu-kategorialnych (Birnbaum, 1968), ponadto stwierdzono jego dużą użyteczność w wypadku analizy testów, w których do poszczególnych pozycji dołączono zbiór możliwych odpowiedzi do wyboru (Hulin, Drasgow i Parsons, 1983), a z taką sytuacją mamy do czynienia w przypadku testu wiedzy Omnibus (opis testu w rozdziale 4.2). Model ten zakłada, iż nawet osoby

o bardzo niskim poziomie inteligencji mierzonej tym narzędziem z pewnym niezerowym prawdopodobieństwem mogą udzielać odpowiedzi poprawnych (Lord, 1980).

3.5.2. Model politomiczny, model klasy odpowiedzi (GRM)

Model GRM jest najlepiej dopasowany teoretycznie do charakteru odpowiedzi udzielanych w kwestionariuszu temperamentu PTS (opis kwestionariusza w rozdziale 4.2) – zakłada on, iż udzielenie danej odpowiedzi jest równoznaczne z udzieleniem zawartych (podrzędnych) odpowiedzi (Reckase, 2009, s. 39). Pozycje kwestionariusza PTS mierzące poziom zgodności doświadczenia osoby uczestniczącej w badaniu z twierdzeniem skali kodowane są od 1 do 4 pkt, a wynikiem ogólnym jest suma punktów, co odpowiada właśnie takiemu założeniu.

Rozdział 4. Przeprowadzone badania i ich wyniki

4.1. Organizacja badań i opis procedury badawczej

Badania zostały zaplanowane w „dwóch równoległych światach" – wirtualnym i rzeczywistym. Pierwsza cześć badań przeprowadzona została za pomocą symulacji w środowisku matematycznym R-project (R Core Development Team, 2010) z użyciem funkcji i procedur zawartych w programie, jak i napisanych specjalnie na potrzeby tej pracy przez jej autora. Kod symulacji zawarty jest w załączniku nr 1. R-project jest otwartym środowiskiem matematycznym stworzonym na potrzeby obliczeń matematycznych i statystycznych, pozwalającym na swobodne stosowanie różnych technik obliczeniowych. Co ważne, środowisko R jest wyposażone w generator liczb pseudolosowych, co jest niezbędne do przeprowadzenia symulacji z użyciem metody Monte Carlo. Z uzyskanych za pomocą symulacji danych możliwe było obliczenie wartości zmiennych *quasi*-prawdziwych dla populacji (przyjmując jej wielkość na poziomie 10000 osób), z którymi porównywano wartości uzyskiwane w losowych próbach o wielkościach zbliżonych do rzeczywistych badań w psychologii rzędu od 50 do 300 osób. Gdyby planować uzyskanie wyników w realnych badaniach, po pierwsze, trudno byłoby osiągnąć tak duże próby, a po drugie, trzeba by kontrolować wiele zmiennych: doświadczenie komputerowe, płeć, ewentualnie poziom lęku. Co więcej, dzięki zastosowaniu symulacji możliwe stało się wyznaczenie funkcji wielkości próby, parametru trudności i mocy różnicującej poszczególnych pozycji względem błędu pomiarowego i rozkładów wyniku prawdziwego, a pondto wyeliminowano zakłócający wpływ powtarzania pomiarów oraz.

Jak wspomniałem wcześniej, w modelach IRT zakładana jest lokalna niezależność pozycji, tak więc dla dowolnego ich zestawu powinienem otrzymać identyczne wyniki przeliczone narzędzia badawczego. Aby zbadać, czy teza ta jest prawdziwa, przeprowadziłem symulację z wykorzystaniem metody Monte Carlo. Na stałej próbie obserwacji o znanych wartościach Θ, za pomocą odpowiedniego algorytmu uzyskałem teoretyczne odpowiedzi, które posłużyły z kolei do zbadania relacji między parametrami wyników otrzymanych i obserwowanych z pełnej wersji testu i kwestionariusza.

Aby odpowiedzieć na pytania dotyczące różnic między wersjami narzędzi badawczych, skróconymi za pomocą różnych technik statystycznych oraz różnic związanych z medium ich prezentacji, przeprowadzone zostały badania na próbach rzeczywistych. Wielkość prób określono we wcześniejszych symulacjach. Po obliczeniu

parametrów dla przyjętych modeli, sporządzone zostały skrócone względem oryginału wersje posiadające kontrolowane parametry mocy różnicującej i trudności poszczególnych zestawów pozycji. Porównane zostały wyniki uzyskane za pomocą różnych wersji tych samych narzędzi: pełnej (FIT – *Fixed-Item Test*), skróconej (SAT – *Self-Adapted Test*) oraz adaptacyjnej[11] (CAT – *Computerised-Adaptive Test*), zarówno w wersji papierowej jak i internetowej. Plan badawczy przewidywał porównania międzygrupowe za pomocą dwuczynnikowej analizy wariancji z uwzględnieniem formy narzędzia badawczego oraz środowiska badania w planie 3x2, według poniższego schematu:

Tabela 3.2. Plan badawczy dotyczący grup osób uczestniczących w badaniach

forma	środowisko badania	
	papier	internet
pełna	TAK	NIE
skrócona	TAK	TAK
adaptacyjna	NIE	TAK

Analiza zgodności między wersjami skróconymi w oparciu o różne metody statystyczne według podejścia klasycznego KTT (CFA oraz MR) i probabilistyczne (IRT) była przeprowadzona z wykorzystaniem współczynnika W Kendalla.

4.1.1. Symulacja wyników w modelu dychotomicznym

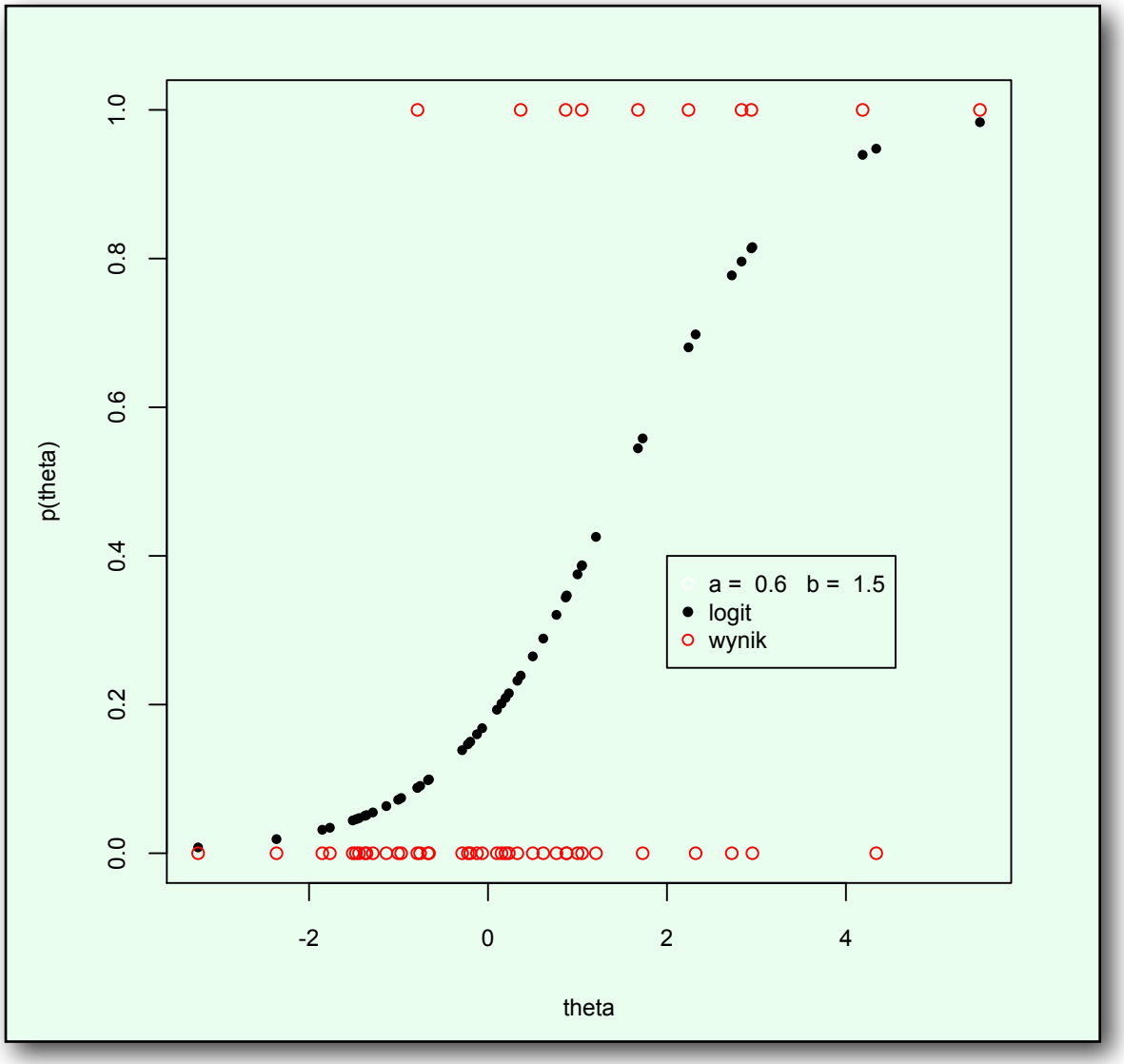
Jak już wcześniej wspomniałem, do analiz na *quasi*-populacji użyto metod symulacyjnych. Z uwagi na fakt, że analizowane modele IRT różnią się od siebie, utworzono dwa niezależne algorytmy: pierwszy, dla zbadania zależności w obrębie zmiennych w modelu dychotomicznym – symulujący test z odpowiedziami zaklasyfikowanymi jednoznacznie jako poprawne i niepoprawne, oraz drugi – dla modelu politomicznego – symulujący zależności między zmiennymi dla kwestionariusza ze skalą typu Likerta.

W pierwszym przypadku używając generatora liczb pseudolosowych w oparciu o rozkład normalny wyznaczono losowe wartości Θ. Następnie, korzystając z funkcji logit dla modelu dwuparametrycznego obliczono wartość prawdopodobieństwa $P(\Theta)$ dla każdej symulowanej osoby. Na koniec porównano losowe wartości z rozkładu jednostajnego (0 – 1) z wartościami obliczonego prawdopodobieństwa. Jeżeli wartości

[11] Do przeprowadzenia badania adaptacyjnego wykorzystano oprogramowanie powstałe w ramach grantu badawczego IP11/1 na Wydziale Nauk Społecznych Uniwersytetu im. A. Mickiewicza.

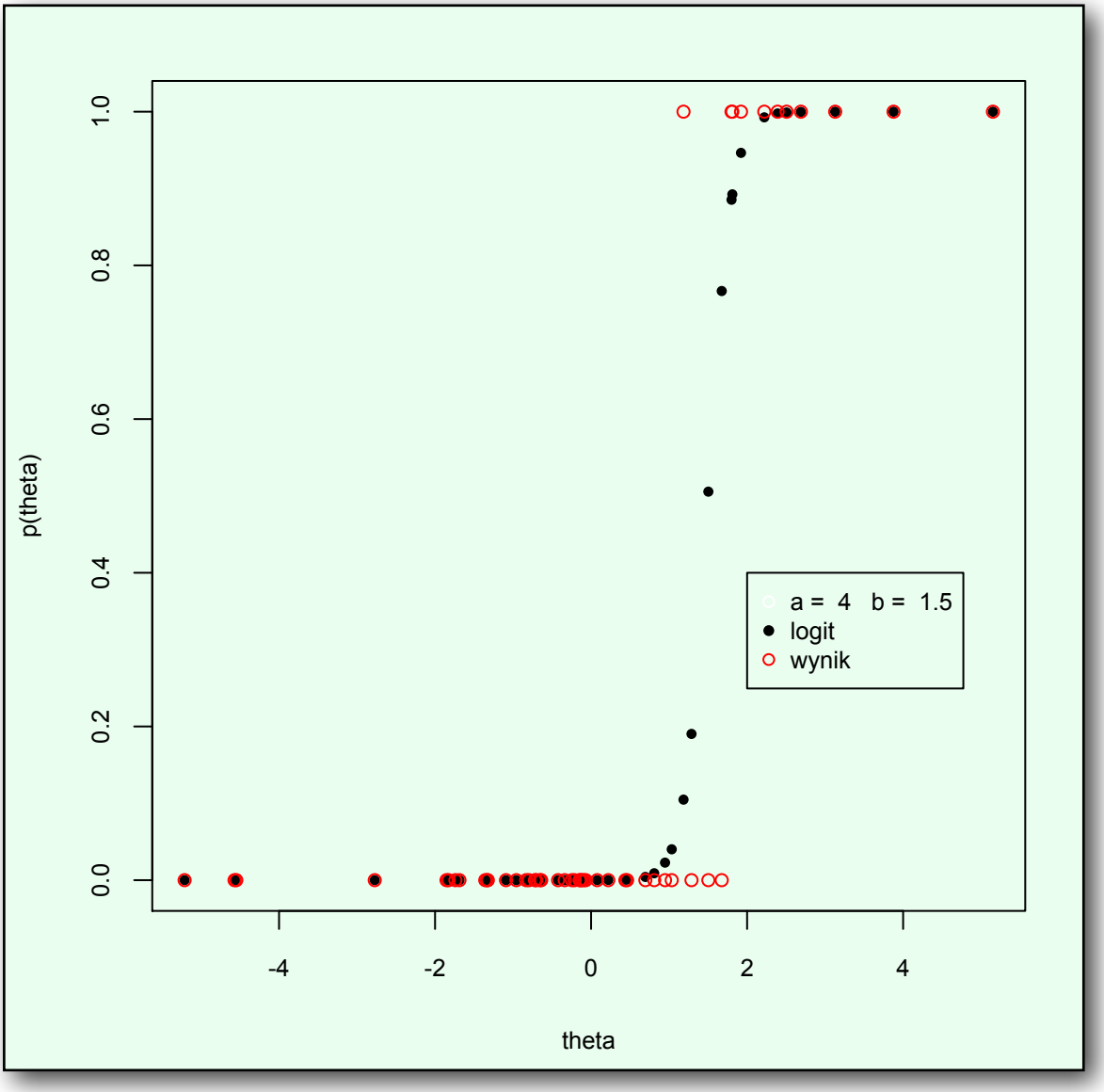
z rozkładu były wyższe niż obliczone prawdopodobieństwo dla danego poziomu Θ, wynikom przypisano wartość sukcesu (1), w przeciwnym wypadku przypisano wartość porażki (0).

Na rycinach od 4.1 do 4.3 przedstawiono trzy wykresy dla przykładowych wartości mocy różnicującej i trudności pozycji testowej. Ryciny te obrazują wyniki procedury symulacji (w celu uproszczenia wyników dla parametru zgadywania c przyjęto stałą wartość równą 0). Na rycinach czarne kropki przedstawiają rozkład prawdopodobieństwa dla losowo wybranych wartości Θ, zaś czerwone kółka obrazują hipotetyczną odpowiedź, jakiej mógłby udzielić uczestnik badania, przyczym górne wartości oznaczają odpowiedź poprawną, zaś dolne błędną.



Ryc 4.1. Wygenerowane wyniki dla pojedynczej pozycji testowej (czerwone) dla przeciętnej mocy różnicującej i wysokiej trudności pozycji testowej (czarne). N = 50. Źródło: opracowanie własne.

Pierwsza wybrana pozycja testowa (por. ryc. 4.1) ma niską moc różnicującą, co ilustruje niskie nachylenie krzywej logitu (czarny kolor). Znajduje to odzwierciedlenie w wynikach (czerwone okręgi), które pojawiają się zarówno po poprawnej (wysoko) i niepoprawnej (nisko) stronie wykresu dla prawie całego zakresu zmiennej latentnej. Mamy więc do czynienia z pozycją, na którą zarówno osoby o niskim poziomie Θ mogą odpowiedzieć poprawnie, jak i osoby o wysokim poziomie Θ mogą odpowiedzieć niepoprawnie.

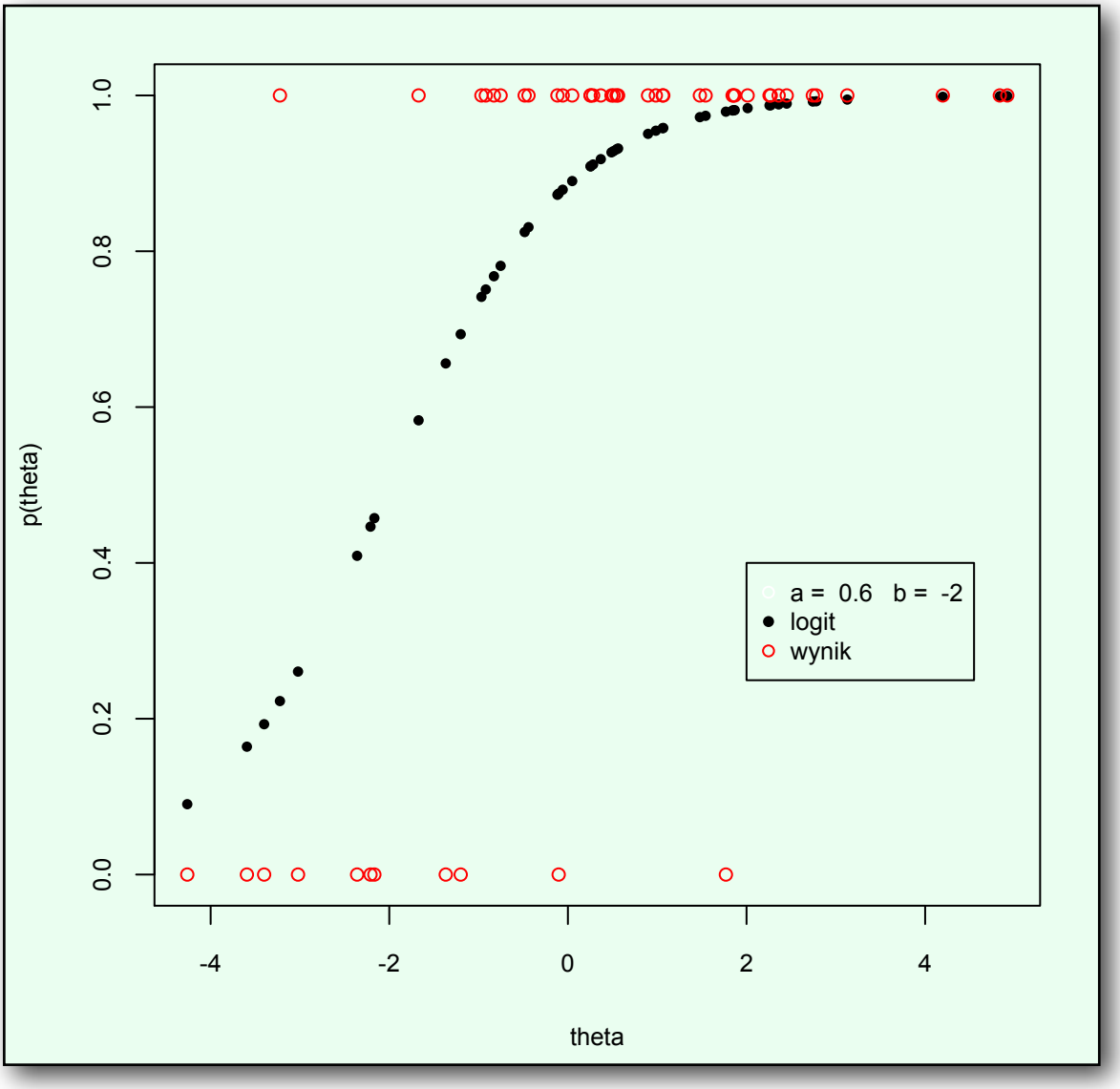


Ryc 4.2. Wygenerowane wyniki dla pojedynczej pozycji testowej (czerwone) dla wysokiej mocy różnicującej i wysokiej trudności pozycji testowej. N = 50. Źródło: opracowanie własne.

Kolejna pozycja testowa prezentowana na rycinie 4.2 obrazuje pytanie bardzo trudne, które mając wysoką moc dyskryminacyjną mocno rozróżnia odpowiedzi osób

badanych na poprawne i niepoprawne. Wyniki osób o poziomie mierzonej cechy poniżej $\Theta = 1{,}5$ są zawsze ocenione jako niepoprawne.

Na kolejnej rycinie (4.3) przedstawiono pozycję testową, która podobnie jak pierwsza – ma niską moc różnicującą i jednocześnie niski poziom trudności. Takie zestawienie parametrów skutkuje tym, że pytanie jest postrzegane jako bardzo łatwe – większość osób odpowiadałaby na nie poprawnie, a odpowiedzi niepoprawne zdarzałyby się osobom zarówno o niskim, jak i średnim poziomie cechy latentnej.



Ryc 4.3. Wygenerowane wyniki dla pojedynczej pozycji testowej (czerwone) dla przeciętnej mocy różnicującej i niskiej trudności pozycji testowej. N = 50. Źródło: opracowanie własne.

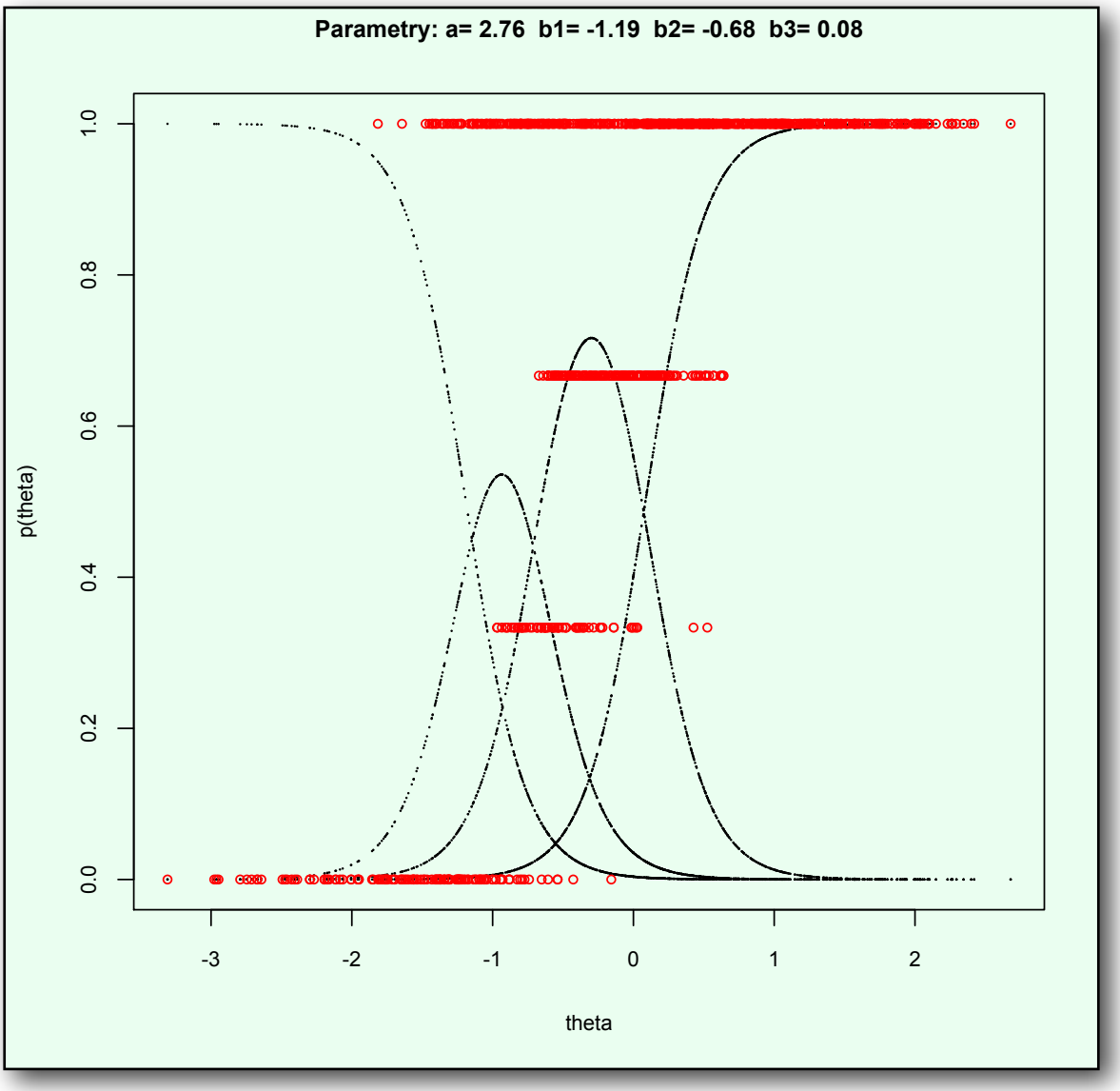
Zgodnie z teoretycznymi założeniami modelu, wraz ze spadkiem trudności pozycji testowej wzrasta liczba osób, które odnoszą sukces, natomiast wraz ze wzrostem mocy

różnicującej skraca się niejednoznaczny zakres Θ, w którym osoby badane mogą odnieść zarówno sukces jak i porażkę przy tej samej wartości cechy latentnej.

4.1.2. Symulacja wyników w modelu politomicznym

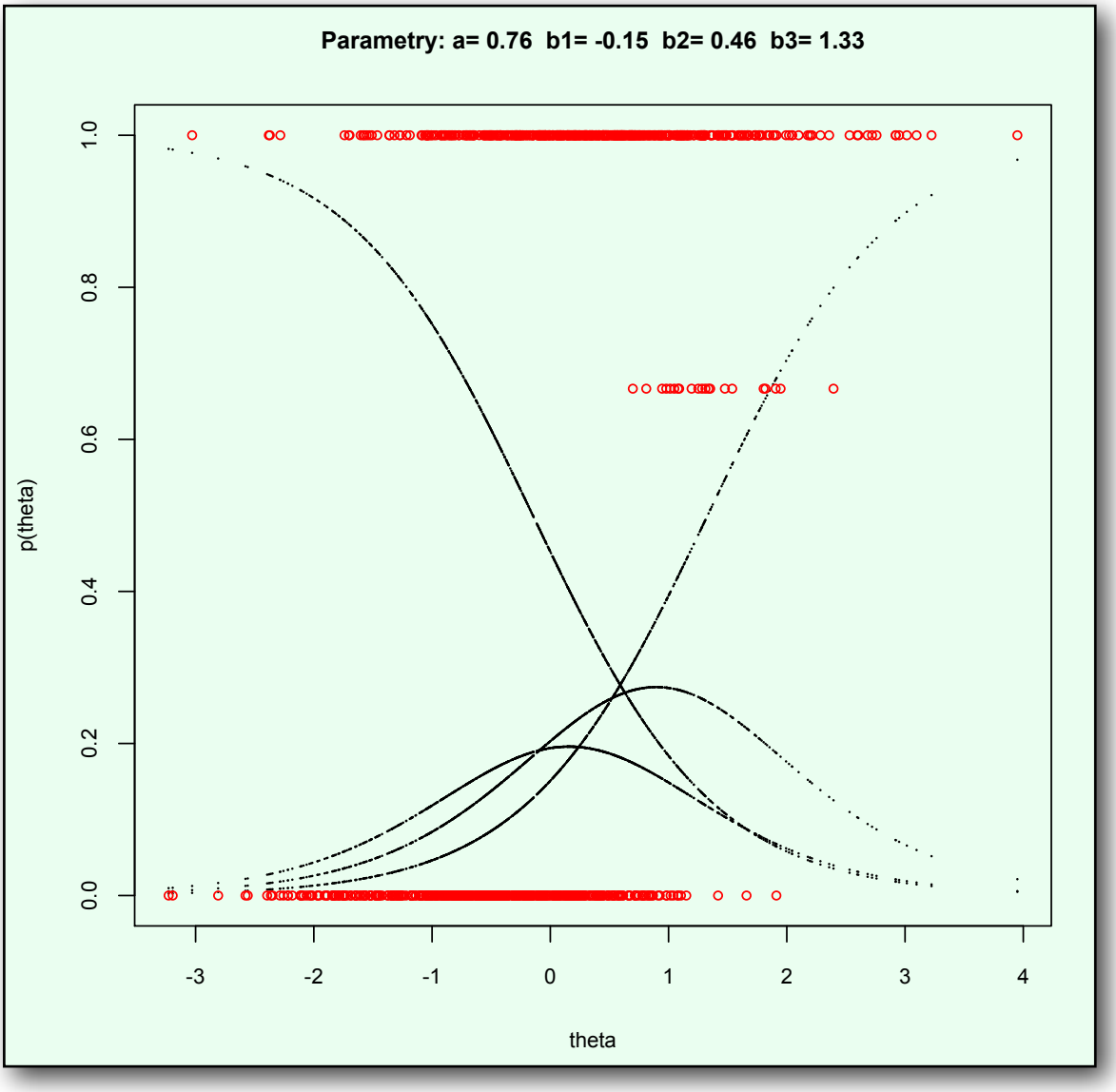
Analogicznie do przedstawionej wyżej symulacji dla modelu dychotomicznego przeprowadzono symulację w modelu dla zmiennych porządkowych – GRM. Różnica polegała na wyznaczeniu prawdopodobieństwa udzielenia odpowiedzi na jednym z czterech poziomów, a nie na określeniu prawdopodobieństwa udzielenia odpowiedzi poprawnej. Poziom wyjściowy przyjmował różne wartości, w zależności od przyjętych parametrów mocy różnicującej i trudności. Zrealizowane to następująco: dla wylosowanych wyników Θ obliczono ich prawdopodobieństwo, a następnie porównano z liczbami z rozkładu losowego 0 – 1. Jeśli prawdopodobieństwo udzielenia odpowiedzi było mniejsze niż wylosowana liczba ustalano odpowiedź na poziomie pierwszym, jeśli było większe – iteracyjnie sprawdzano następny próg.

Ryciny 4.4 oraz 4.5 przedstawiają dwie przykładowe pozycje kwestionariuszowe uzyskane za pomocą symulacji. Czarne punkty pokazują rozkład prawdopodobieństwa, czerwone okręgi – poziom uzyskanych odpowiedzi (1, 2, 3 lub 4).



Ryc 4.4. Wyniki symulacji w modelu GRM dla wysoce różnicującej pozycji kwestionariuszowej. Źródło: opracowanie własne.

Pierwsza rycina (ryc 4.4) ilustruje pozycję testową o wysokim poziomie parametru a oraz poziomie trudności dla poszczególnych kategorii odpowiedzi od b_1 = -1,19 do b_3 = 0,08. Na drugiej rycinie (ryc 4.5) poziom mocy różnicującej jest niższy, natomiast poziom trudności wyższy.



Ryc 4.5. Wyniki symulacji w modelu GRM dla słabo różnicującej pozycji testowej. Źródło: opracowanie własne.

Różnice między rycinami 4.4 oraz 4.5 obrazują związek między parametrami pozycji kwestionariuszowych a odpowiedziami zebranymi za ich pomocą – przy słabo różnicującej pozycji osoby uczestniczące w badaniu udzielają skrajnych odpowiedzi – mało jest wartości środkowych. Poziom trudności b odpowiada zaś za przesunięcie zbioru wyników wzdłuż osi Θ.

4.2. Operacjonalizacja zmiennych

Opisane w rozdziale 4.1 sposoby generowania danych w oparciu o metody symulacyjne stanowią jedno z dwóch źródeł danych wykorzystanych w tej pracy. Tym drugim są badania przeprowadzone za pomocą istniejących i wykorzystywanych w praktyce psychologicznej narzędzi badawczych.

Do badań zostały wybrane dwa narzędzia:

- test Omnibus (Jaworowska, Matczak, 2002) badający wiedzę ogólną. Jest to przykład testu z odpowiedziami typu prawda-fałsz, jednen z testów inteligencji zbudowanych w oparciu o pozycje kwestionariuszowe a nie zadania do rozwiązania;
- kwestionariusz PTS (*Pavlovian Temperament Survey* – Strelau, Angleitner i Newberry, 1999) badający temperament. Wynik w tym kwestionariuszu jest uzyskiwany na podstawie odpowiedzi na skali typu Likerta.

Narzędzia te zostały wybrane, ponieważ ich konstrukcja pozwala na zastosowanie testowania adaptacyjnego bez naruszania warunków przeprowadzania badania. Na gruncie klasycznej teorii testów charakteryzują się one wysokimi właściwościami pomiarowymi: rzetelnością i trafnością teoretyczną (PTS – Strelau i inni, 1995). Są też często wykorzystywane w praktyce psychologicznej do pomiaru inteligencji i temperamentu.

Test Omnibus składa się z 60 zamkniętych pozycji testowych, gdzie osoby uczestniczące w badaniu mają za zadanie wybrać jedną z pięciu[12] podanych odpowiedzi jako poprawną. W skład puli zadań wchodzą antonimy, analogie werbalne, szeregi liczb, wyrażenia frazeologiczne oraz sylogizmy. Każdy z pięciu typów zadań występuje tyle samo razy w teście w sposób naprzemienny, w zestawach po trzy zadania, które mają charakter zamknięty. Trudność poszczególnych zadań jest narastająca, a na wykonanie wszystkich zadań limit czasu wynosi 60 minut. Zadania mogą być wykonywane w dowolnej kolejności.

[12] Dla sylogizmów jedną z trzech odpowiedzi.

Przykładowe pozycje testu Omnibus

Przykładem pozycji testowej dotyczącej antonimów jest zadanie polegające na wybraniu przez osobę uczestniczącą w badaniu spośród pięciu wyrazów takiego, który ma przeciwstawne znaczenie do podanego, np.: do słowa ***pean*** należy wybrać jedno z następujących: *motto*, *blef*, *panegiryk*, *paszkwil*, *intryga*.

Przykładem pozycji testowej dotyczącej analogii werbalnych jest zadanie polegające na znalezieniu relacji łączącej elementy podaje pary i użycie tej relacji do uzupełnienia pary niekompletnej. Np. należy odkryć relację dla pary: ***ołówek - brudnopis***, a następnie wskazać pasujący rzeczownik do wyrazu ***igła*** z puli następujących: *szycie*, *ścieg*, *nić*, *materiał*, *fastryga*.

Kolejne zadanie – szeregi liczb – polega na znalezieniu reguły, według której zbudowany jest ciąg liczb i uzupełnienie brakującej poprzez wybranie z podanych. Np.: dla ciągu ***1, 4, 9, 16, ?*** należy wybrać jedną z podanej listy liczb: *21*, *23*, *24*, *25*, *26*.

Wyrażenia frazeologiczne to czwarty typ zadań, który wymaga od osób uczestniczących w badaniach wyjaśnienie sensu podanego wyrażenia za pomocą jednego z pięciu podanych twierdzeń. i tak dla ***tajemnicy poliszynela*** należy wybrać jedno z następujących: tajemnica publiczna, tajemnica faraonów, tajemnica lekarska, tajemnica państwowa, tajemnica spowiedzi.

I wreszcie sylogizmy to zadania polegające na ocenie prawdziwości wniosku wyciągniętego na podstawie podanych przesłanek. Osoba uczestnicząca w badaniu zapoznaje się z całym procesem wnioskowania: ***Niektórzy pracownicy fabryki nie są robotnikami*** oraz ***Niektórzy pracownicy fabryki są mieszkańcami osiedla fabrycznego*** to ***Niektórzy mieszkańcy osiedla fabrycznego nie są robotnikami***. Ocena trafności wnioskowania jest wyrażona poprzez wybór jednej z trzech odpowiedzi: tak, nie lub nie wiadomo.

Za wskazanie poprawnej odpowiedzi przydzielany jest 1 punkt, a wynikiem ogólnym w teście jest suma punktów. Na podstawie analizy czynnikowej wyróżnione są także dwa czynniki: rozumowania oraz wiedzy (po 25 pozycji) odpowiadające pojęciu inteligencji płynnej i skrystalizowanej. Rzetelność wyniku ogólnego i obu czynników waha się od 0,78 dla czynnika wiedzy w normalizacyjnej grupie studentów do 0,93 dla wyniku ogólnego w grupie osób dorosłych.

Kwestionariusz PTS w polskiej wersji składa się z 57 zamkniętych pozycji kwestionariuszowych, gdzie osoby uczestniczące w badaniu orzekają prawdziwość podanych twierdzeń w odniesieniu do siebie na czterostopniowej skali (od: „zdecydowanie nie zgadzam się” poprzez „raczej się nie zgadzam", „raczej się zgadzam” do „zdecydowanie zgadzam się"). Poszczególne pozycje mają charakter zdań twierdzących i tworzą w oparciu o koncepcję Pawłowa trzy czynniki pozwalając oszacować siłę procesów pobudzenia, siłę procesów hamowania oraz ruchliwość procesów nerwowych.

Przykładowe pozycje kwestionariusza PTS

Stwierdzenia zawarte w kwestionariuszu mają charakter zdań oznajmujących. Osoba badana jest proszona o ocenę na ile dane twierdzenie opisuje ją samą. Przykładowe zdania to: **Chętnie opowiadam dowcipy i anegdoty** lub **Od czasu do czasu chętnie wdaję się w pogawędkę.**

Za zgodę z danym twierdzeniem osoba uczestnicząca w badaniu uzyskuje zgodnie z kluczem od 1 punktu do 4 punktów. Wynik sumaryczny przeliczany jest według norm stenowych w grupach wiekowych, zróżnicowanych ze względu na płeć. Na podstawie konfiguracji wyników uzyskanych w 3 czynnikach osoby badane mogą mieć przypisany w celach opisowych jeden z 4 typów temperamentu: melancholiczny (niska siła procesów pobudzenia), choleryczny (wysoka siła procesów pobudzenia oraz niska siła procesów hamowania), sangwiniczny (wysoka siła procesów pobudzenia, wysoka siła procesów hamowania i wysoka ruchliwość procesów nerwowych) oraz flegmatyczny (w odróżnieniu od poprzedniego niska ruchliwość procesów nerwowych). Często też w procesie diagnozy pozostaje się na poziomie opisu wyników na trzech skalach wymienionych wyżej, dostarczając informacji na temat sposobu funkcjonowania osoby badanej w sytuacjach dużej stymulacji, dużej zmienności oraz sytuacjach trudnych, gdzie przydatna jest informacja o sposobie kontroli zachowania.

4.3. Osoby uczestniczące w badaniach

Badania papierowe przeprowadzono w latach 2009 – 2011 roku. W badaniu wzięło łącznie udział 601 osób – 293 wypełniły kwestionariusz temperamentu PTS, a 308 test Omnibus, badający inteligencję. W obu przypadkach były to grupy homogeniczne wiekowo: dla kwestionariusza PTS byli to uczniowie szkół średnich, a dla testu Omnibus studenci poznańskich uczelni. Charakterystykę płci i wieku osób uczestniczących w badaniach papierowych przedstawia tabela 3.3.

Tabela 3.3. Opis grupy badawczej

Płeć	**Omnibus**		**PTS**	
Kobiety	231 (74,8%)		174 (59,4%)	
Mężczyźni	67 (21,7%)		119 (40,6%)	
B.D.	10 (3,6%)			
Wiek	kobiety	mężczyźni	kobiety	mężczyźni
średnia ± SD	21,16 ± 2,26	21,37 ± 2,71	15,54 ± 1,75	15,08 ± 1,72
skośność	0,79	1,21	0,18	0,55
kurtoza	0,31	0,75	-0,92	-0,74
mediana	20	20	15	15

Źródło: badania własne.

Badanie z użyciem pełnej wersji kwestionariusza temperamentu PTS przeprowadzono wśród młodzieży uczestniczącej w obozach wypoczynkowych w Zespole Uzdrowisk Kłodzkich w 2009 roku. Z 442 osób uczestniczących w obozach, które pochodziły z miejscowości o różnej wielkości (por. tabela 3.4), 293 osoby wyraziły zgodę i wzięły udział w badaniu. Badanie przeprowadzono podczas kilku zebrań grupowych. Najpierw udzielano informacji na temat badania i rozdawano kwestionariusze, a następnie każdy uczestnik w badania wypełniał kwestionariusz indywidualnie. Wypełnione arkusze zbierano w sposób zapewniający poufność danych.

Tabela 3.4. Rozkład częstości wielkości miejsca zamieszkania dla osób badanych kwestionariuszem PTS

Zamieszkanie	Częstość	Procent	Procent ważnych
wieś i miasteczko	18	6,1	7,7
miasto	143	48,8	60,9
duże miasto	74	25,3	31,5
Ogółem	235	80,2	100,0
Braki danych	58	19,8	
Ogółem	293	100,0	

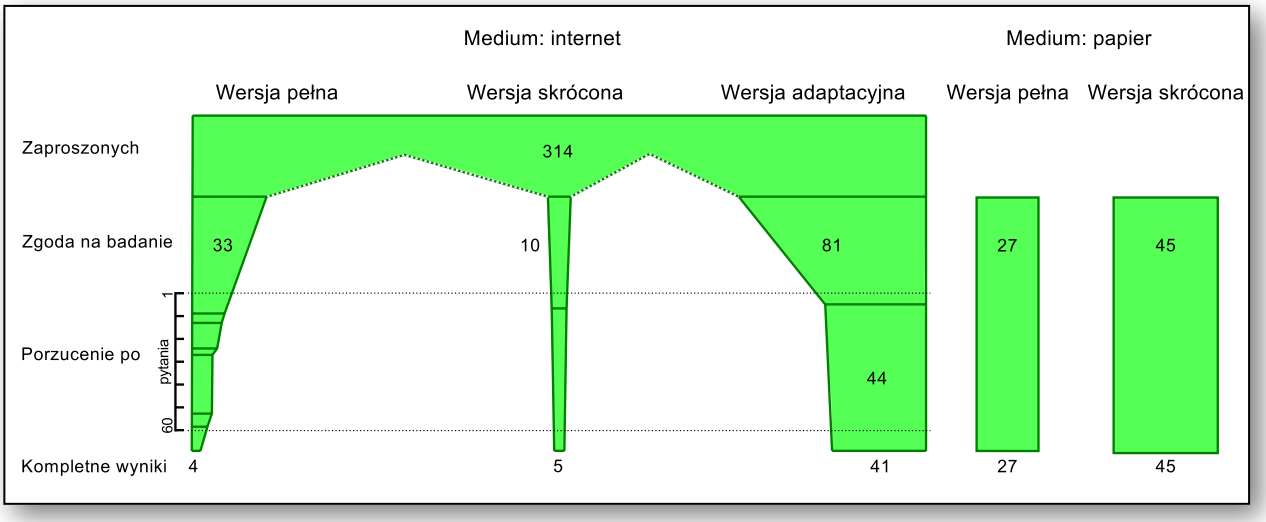
Źródło: badania własne.

Badania papierową wersją pełną testu inteligencji OMNIBUS przeprowadzone zostały w całości w Poznaniu, wśród studentów Uniwersytetu im. Adama Mickiewicza. Dobór do próby opierał się na zasadzie dostępności i dobrowolności – studentom 2 i 3 roku obecnym na wykładach zaproponowano udział w badaniu. W 61% badana próba składała się ze studentów kierunku psychologia, pozostałe 39% stanowili studenci kierunku geografia, skandynawistyka oraz pedagogika. W każdej grupie testy wypełniane były w salach wykładowych z ograniczeniem czasowym wynoszącym 60 minut. Na początku badania zapoznano osoby uczestników z instrukcją, a następnie rozdano testy i arkusze odpowiedzi. Wypełnione arkusze zbierano w sposób zapewniający anonimowość.

Badania za pomocą internetu przeprowadzone zostały w latach 2011 – 2012 za pomocą systemu do badań, który powstał w ramach grantu badawczego IP11/1 na Wydziale Nauk Społecznych (opis systemu: patrz załącznik 4). Poniżej przedstawiono dwie ryciny (4.6 i 4.7) ilustrujące proces zbierania danych za pomocą obu narzędzi badawczych.

W przypadku medium internetowego osoby zaproszone do badań były losowo przydzielane do jednego z trzech warunków badawczych. Zaproszenia wysyłane były

na adres mejlowy zgromadzony w bazie danych Instytutu Psychologii Uniwersytetu im. Adama Mickiewicza. Po kliknięciu na indywidualny link osoba zapoznawała się z opisem badania na stronie internetowej i mogła udzielić zgody na udział w badaniu. W przypadku testu Omnibus takich zgód uzyskano najwięcej od osób zaproszonych do wersji adaptacyjnej (81 osób) a najmniej od tych, którym zaproponowano wypełnienie wersji skróconej (10 osób). Po wyrażeniu zgody osoby uczestniczące w badaniu zapoznawały się z instrukcją narzędzia badawczego i przystępowały do rozwiązywania zadań. Zostały one poinformowane, że w każdej chwili mogą zrezygnować z badania, oraz że badanie jest anonimowe. Wśród tych, które zgodziły się na badanie i je ukończyły, najwyższy odsetek osób obserwowano dla badań adaptacyjnych (50,6%) i wersji skróconych (50%) a najniższy dla wersji pełnej (12,1%).



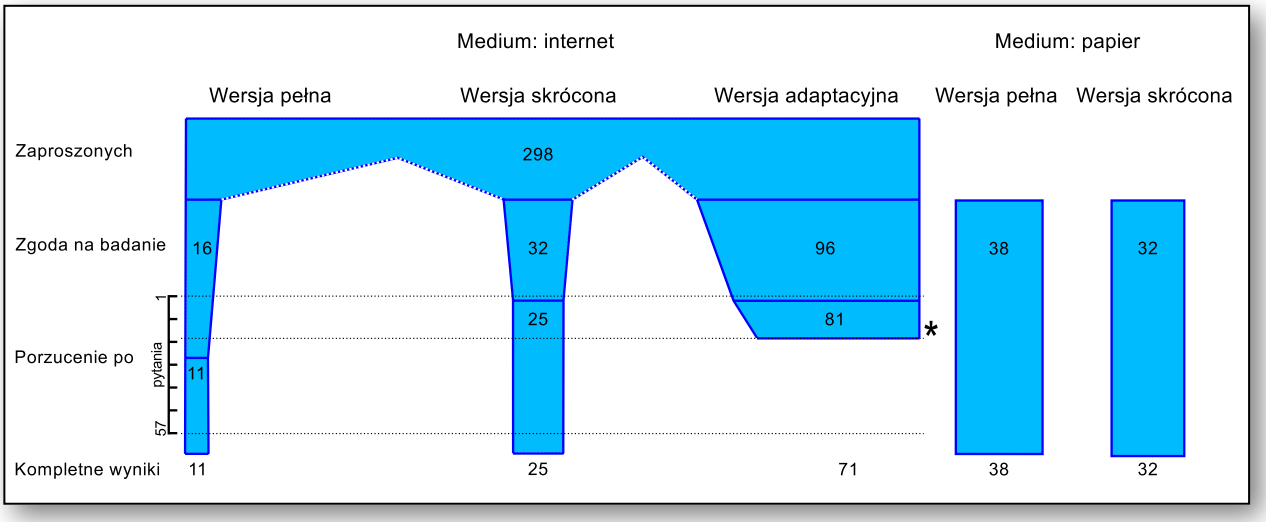
Ryc 4.6. Proces zbierania danych za pomocą testu Omnibus.

Liniami kropkowanymi przedstawiono schematycznie rozpiętość testu, aby oddać miejsce, w którym osoby porzucały badanie. Dla wersji adaptacyjnej tylko 3 osoby zrezygnowały w trakcie badania i stało się to na samym jego początku. Podobnie w wersji skróconej, gdzie zrezygnowało 5 osób. Bardziej zróżnicowany przebieg rezygnacji z badań wystąpił w przypadku wersji pełnej, gdzie najwięcej (19) osób zrezygnowało już na pierwszym pytaniu, a potem sukcesywnie co kilka pytań rezygnowały kolejne osoby. Pokazuje to jak ważny dla realizacji badań przez internet jest czas, który muszą poświęcić osoby badane – narzędzia długie zwiększają prawdopodobieństwo zebrania niekompletnych danych.

Czynnikiem, który prawdopodobnie różnicował realizację próby dla trzech wersji narzędzia badawczego była podana w instrukcji zróżnicowana informacja o maksymalnym czasie potrzebnym do wypełnienia testu. W wersji pełnej i adaptacyjnej

było to (zgodnie z wersją papierową) maksymalnie 60 minut, w wersji skróconej ze względu na stałą, mniejszą liczbę pozycji testowych podano 29 minut. Podanie dokładnej wartości mogło zwracać uwagę osób badanych i wpływać na zmniejszenie motywacji do udziału w badaniu.

Rycina 4.7 przedstawia analogiczny proces dla kwestionariusza PTS.



Ryc. 4.7. Proces zbierania danych za pomocą kwestionariusza PTS.

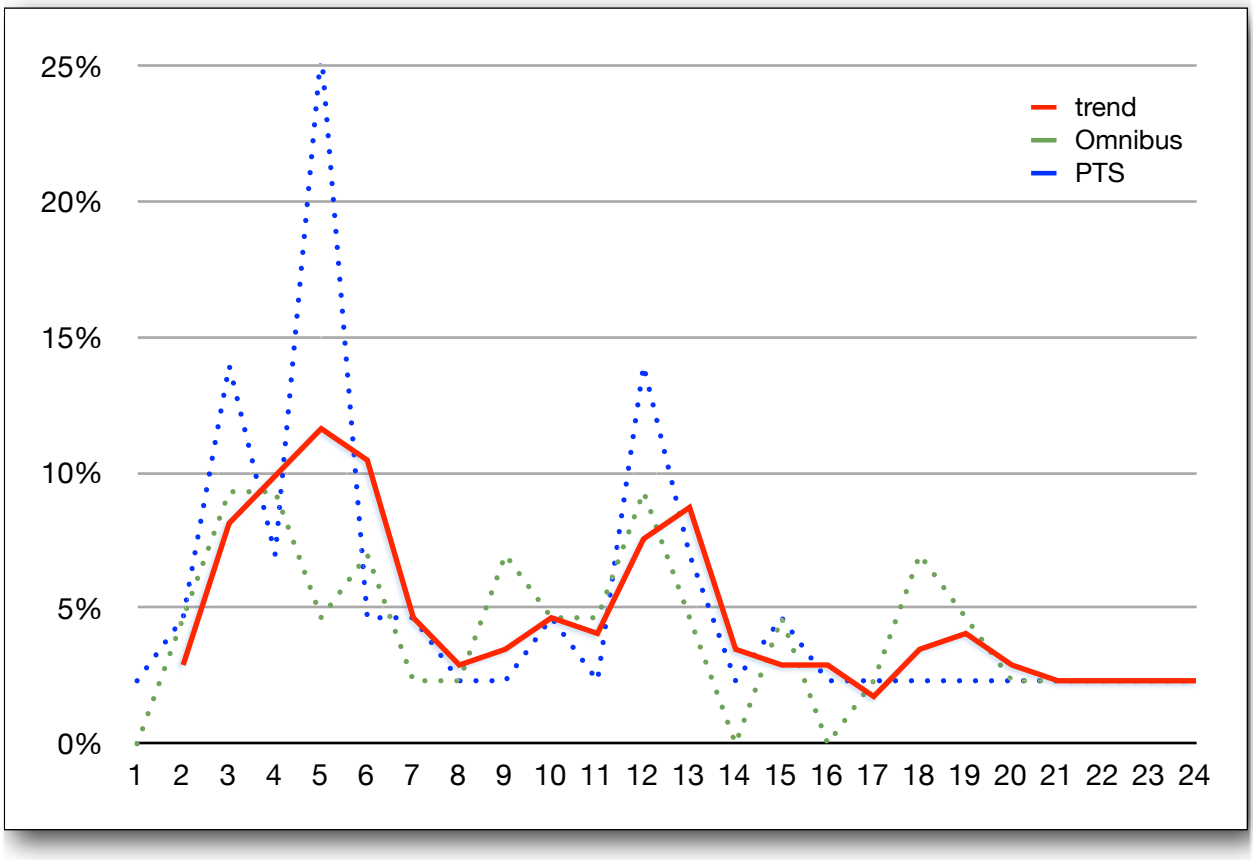
Na rycinie 4.7 zaakcentowano (*) różnicę związaną z przeprowadzeniem badania dla wersji adaptacyjnej, która obejmowała tylko jeden czynnik temperamentu reprezentowany przez pozycje kwestionariuszowe dotyczące ruchliwości procesów nerwowych. Czynnik ten wybrano ze względu na konieczność zachowania jednowymiarowości zmiennej latentnej w badaniu adaptacyjnym. Tym samym w tej wersji osoby badane odpowiadały na pytania z puli 19, a nie 57 pozycji jak w pozostałych przypadkach.

Ponownie bezwzględnie najwięcej osób ukończyło wersję adaptacyjną, przy czym w tej części badania dla żadnej z wersji nie podano czasu potrzebnego do jej ukończenia.

Zaobserwowana różnica między realizacją próby w przypadku Omnibusu a PTS może być tłumaczona stopniem trudności pytań w obu narzędziach. Ze względu na charakter i mierzone cechy test Omnibus składał się z subiektywnie trudniejszych pozycji. Wymagał dokonywania obliczeń, wnioskowania oraz sprawdzał wiedzę, natomiast kwestionariusz PTS odwoływał się tylko do wyrażenia zgody i potwierdzenia prawdziwości twierdzeń w stosunku do własnej osoby. Najprawdopodobniej z tego powodu zaobserwowano mniej porzuceń w przypadku właśnie drugiego narzędzia.

Dla porównania wielkości prób między wersjami przedstawiono ich wielkość dla pełnego i skróconego narzędzia w badaniach wersjami papierowymi.

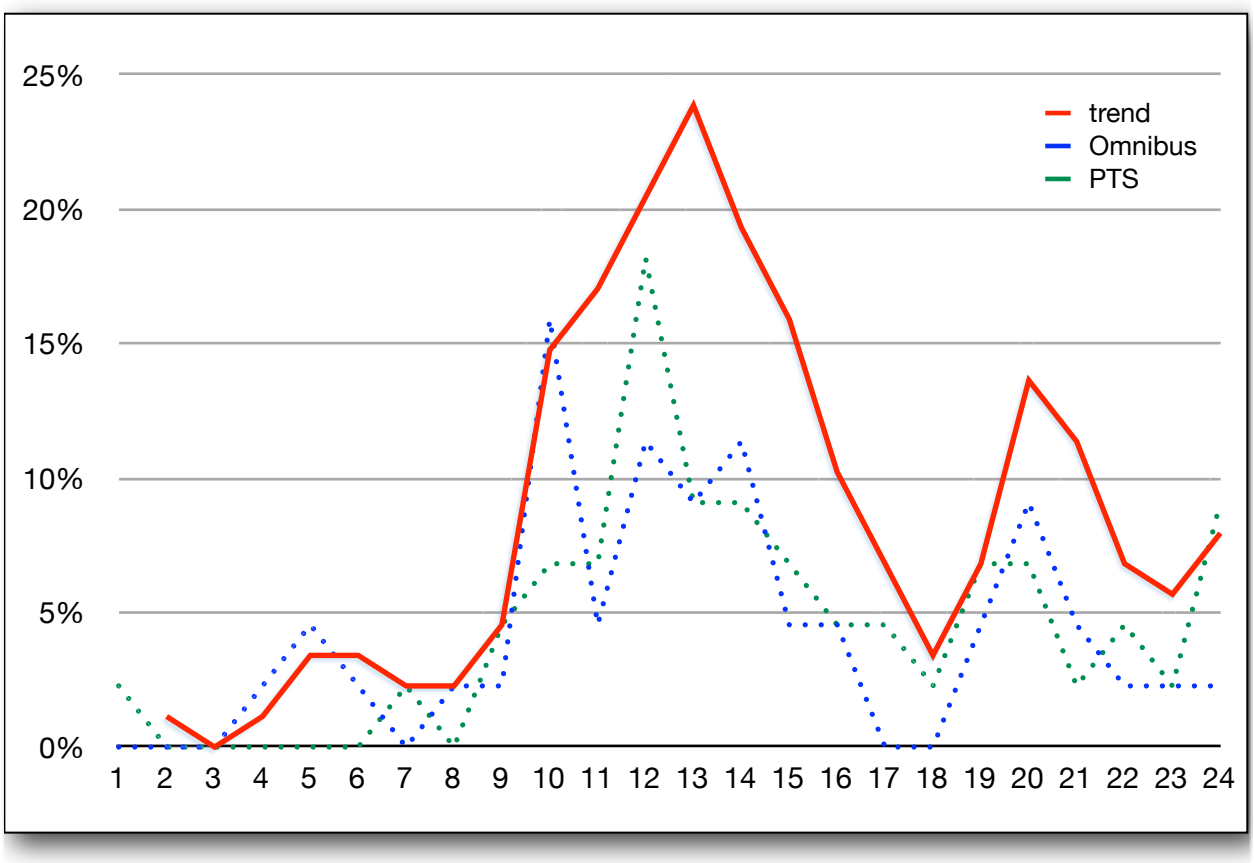
Podczas badania za pośrednictwem internetu dodatkowo zebrano informację o dacie oraz godzinie badania. Obie zmienne podsumowują ryciny 4.8 i 4.9.



Ryc. 4.8. Trend aktywności osób uczestniczących w badaniach w jego kolejnych dniach.

Największą aktywność w badaniach asynchronicznych obserwuje się na początku badania – nie inaczej było i w tym przypadku. Po wysłaniu zaproszeń najwięcej osób zdecydowało się na wzięcie udziału w badaniu podczas pierwszego tygodnia. Później zaproszenie przesłaniają bieżące sprawy i niewielki odsetek osób podejmuje wysiłek wzięcia udziału w badaniu. Wysyłanie przypomnień o możliwości wzięciu udziału w badaniach miało miejsce 12 dnia od rozpoczęcia badań i wtedy też przypada drugi szczyt liczby odwiedzin. Osoby, które nie odpowiedzą na zaproszenie w ciągu kilku (czterech do siedmiu) dni od wysłania zaproszeń – raczej nie podejmą już takiej decyzji w terminie późniejszym.

Pierwszy szczyt aktywności dobowej przypada natomiast na godziny 13-14, a następny na godziny 20-21 i odpowiadają one przeciętnemu wzorowi aktywności ludzi. Są to godziny czy to w pracy, czy w domu, które ludzie poświęcają na mniej ważne czynności pod koniec pracy lub przed snem.

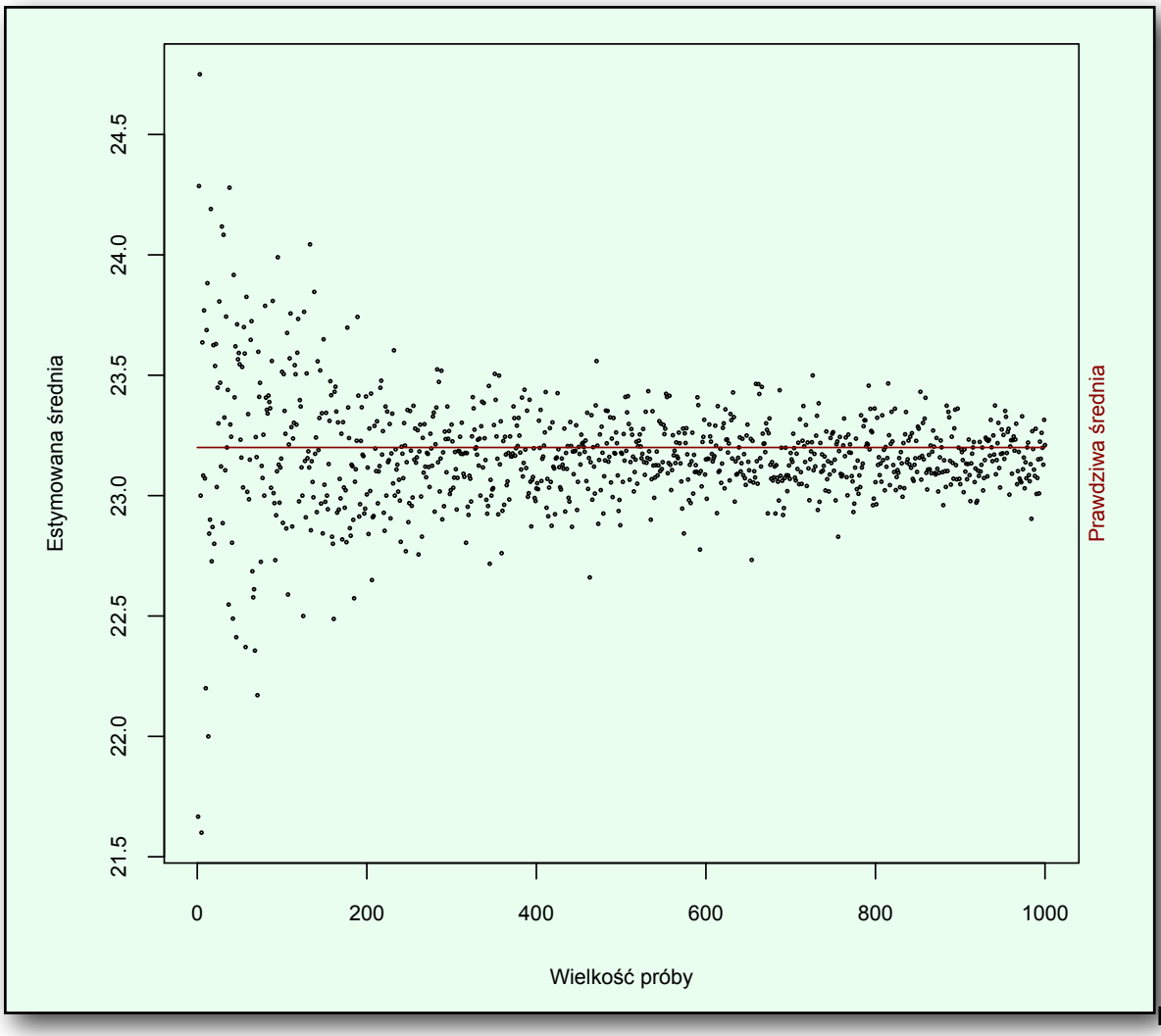


Ryc. 4.9. Trend dobowy aktywności osób uczestniczących w badaniach.

Zaobserwowany rozkład aktywności może sugerować sposób, w jaki ludzie traktują badania internetowe w warunkach braku kontroli. Jednocześnie jeśli jakieś badanie wymagałoby maksymalnego skupienia, należało by kontrolować porę badania, ponieważ ludzie z własnej woli odkładają uczestnictwo w nim na okres mniejszej wydajności.

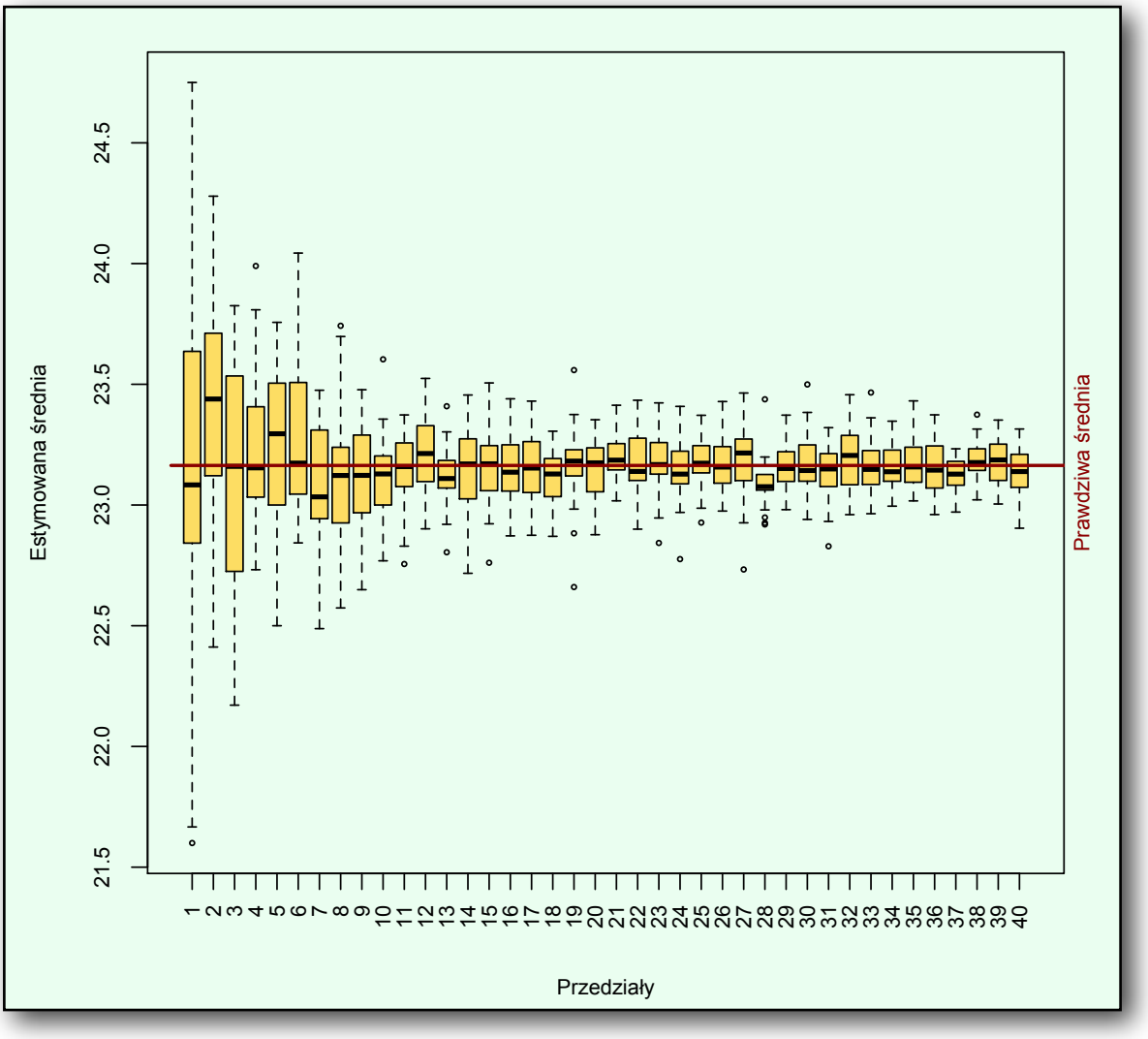
4.4. Szacowanie wielkości próby kalibracyjnej

Aby określić w sposób eksperymentalny wielkość próby potrzebną do wyznaczenia stabilnych parametrów dla pozycji testowych lub kwestionariuszowych, przeprowadzono symulację z użyciem metody MCMC. Założono wielkość populacji N = 100.000 oraz wyznaczono średni wynik (na rycinie 3.1 pozioma czerwona linia) w hipotetycznym teście składającym się z 60 pozycji testowych o znanych parametrach a oraz *b*. Taką długość testu przyjęto *ad hoc* opierając się na długości narzędzi używanych w dalszych analizach: Omnibus posiada 60 pozycji, PTS – 57 pozycji. Następnie dokonano szeregu estymacji średniej w oparciu o te same parametry co dla populacji, ale dla wielu różnych prób w wielkości od 5 do 1.000 osób. Rozrzut uzyskanych średnich przedstawia rycina 3.1.



Ryc. 4.1. Symulowany wynik prawdziwy (czerwona linia) i jego estymacje w oparciu o metodę MCMC dla prób o różnej wielkości. Źródło: opracowanie własne.

Analizując wykres można zauważyć, iż rozrzut wyników wokół wyniku prawdziwego maleje wraz ze wzrostem wielkości próby, lecz poprawa w dokładności estymacji zatrzymuje się dla próby liczącej około 300 osób. Dalsze zwiększanie liczebności próby nie przynosi już wyraźnej poprawy. Aby lepiej zobrazować zmiany i wyznaczyć wystarczającą wielkość próby sporządzono dla średnich przedziały o szerokości 25 osób w grupie wykreślone na rycinie 4.2.



Ryc 4.2. Rozrzut średnich w kolejnych przedziałach (1=1:25, 2=26:50, 3=51:75, itd.). Wykresy skrzynkowe w oparciu o mediany. Źródło: opracowanie własne.

Obciążoną niewielkim błędem estymację przy niezbyt dużym rozrzucie udało się uzyskać już dla 10 przedziału, czyli dla 250 osób. Adekwatność tego szacowania można odnieść do wyników meta-analizy, jakiej poddano próby w różnych badaniach psychologicznych opublikowanych w latach 1969 – 1998 w czasopismach: *Applied Psychology, Personnel Psychology* i *Academy of Management Journal* (Aguinis i inni, 2005). Meta-analiza ta wykazała, że przeciętna wielkość próby w uwzględnionych badaniach naukowych wynosiła 272 osoby. Opierając się na tych doniesieniach oraz na wynikach przeprowadzonych obliczeń w dalszych badaniach przyjęto wielkość próby na poziomie 275 osób jako wystarczającą do wyznaczenia stabilnych parametrów dla pozycji narzędzia badawczego.

4.5. Opis statystyczny uzyskanych wyników

Test Ominibus

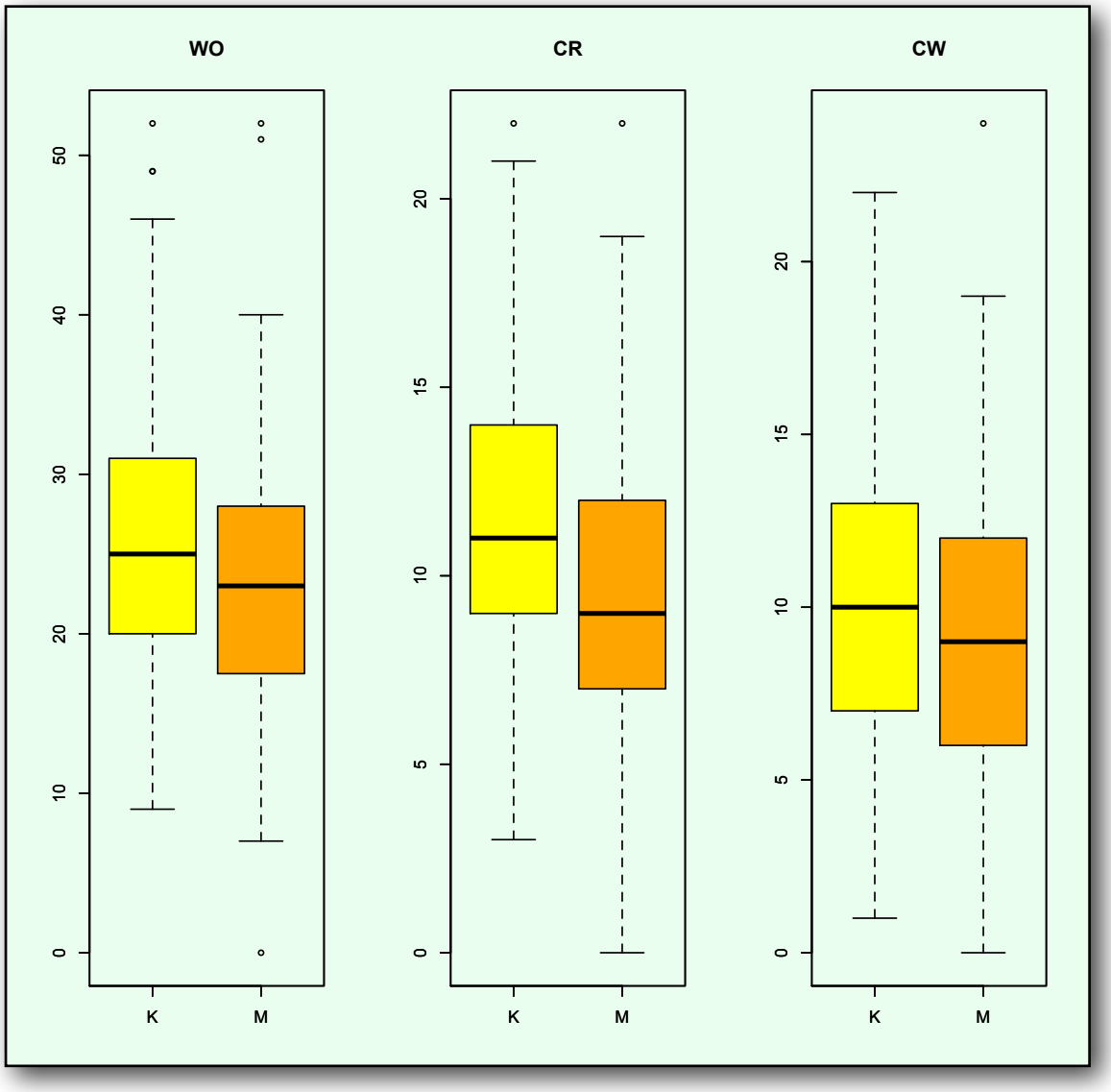
Rozkład wyników uzyskany przez osoby uczestniczące w badaniach miał charakter lekko prawoskośny. W odniesieniu do norm, kobiety uzyskały wyniki wyższe niż grupa normalizacyjna na wymiarach rozumowania oraz wiedzy, natomiast mężczyźni uzyskali istotnie niższe wyniki na wymiarze rozumowania (patrz tabela 4.1).

Tabela 4.1. Statystyki opisowe wyników w teście Omnibus i istotność różnic w stosunku do norm oraz między grupami wyróżnionymi ze względu na płeć

	średnia ± SD						
Skala	mężczyźni	kobiety	skośność	kurtoza	$1Q$	median a	$3Q$
wynik ogólny	23,7 ± 9,08	25,7 ± 5,66 [b]	0.52	0.53	19	24	31
czynnik wiedzy	9,6 ± 4,36	10,21 ± 4,34 [c]	0.41	-0.11	7	10	13
czynnik rozumowania*	9,9 ± 4,31[a]	11,4 ± 4,09 [d]	0.32	-0.02	8	11	14

Różnica w stosunku do norm: a – istotna dla $p < 0,05$ ($t_{(66)} = -2,24$); b – istotna dla $p < 0,001$ ($t_{(228)} = 9,99$); c – istotna dla $p < 0,001$ ($t_{(228)} = 8,69$); d – istotna dla $p < 0,001$. Różnica między grupami: * – istotna dla $p < 0,05$ ($t_{(294)} = 2,49$).

Jednocześnie można zaobserwować istotną na poziomie $p < 0,05$ różnicę między mężczyznami i kobietami na korzyść tych ostatnich, pod względem wyników w czynniku rozumowania (patrz ryc. 4.3). Dla czynnika wiedzy i wyniku ogólnego nie stwierdzono istotnych różnic między płciami.



Rycina 4.3. Wykresy skrzynkowe w oparciu o medianę rozkładu wyników w teście Omnibus z uwzględnieniem płci. WO - wynik ogólny, CW - czynnik wiedzy, CR - czynnik rozumienia. Źródło: badania własne.

Wyniki uzyskane w poszczególnych pytaniach testu posłużyły do obliczenia parametrów IRT według modelu 3PL dla każdej z pozycji testowych. Parametry te przedstawiono w tabeli 4.2.

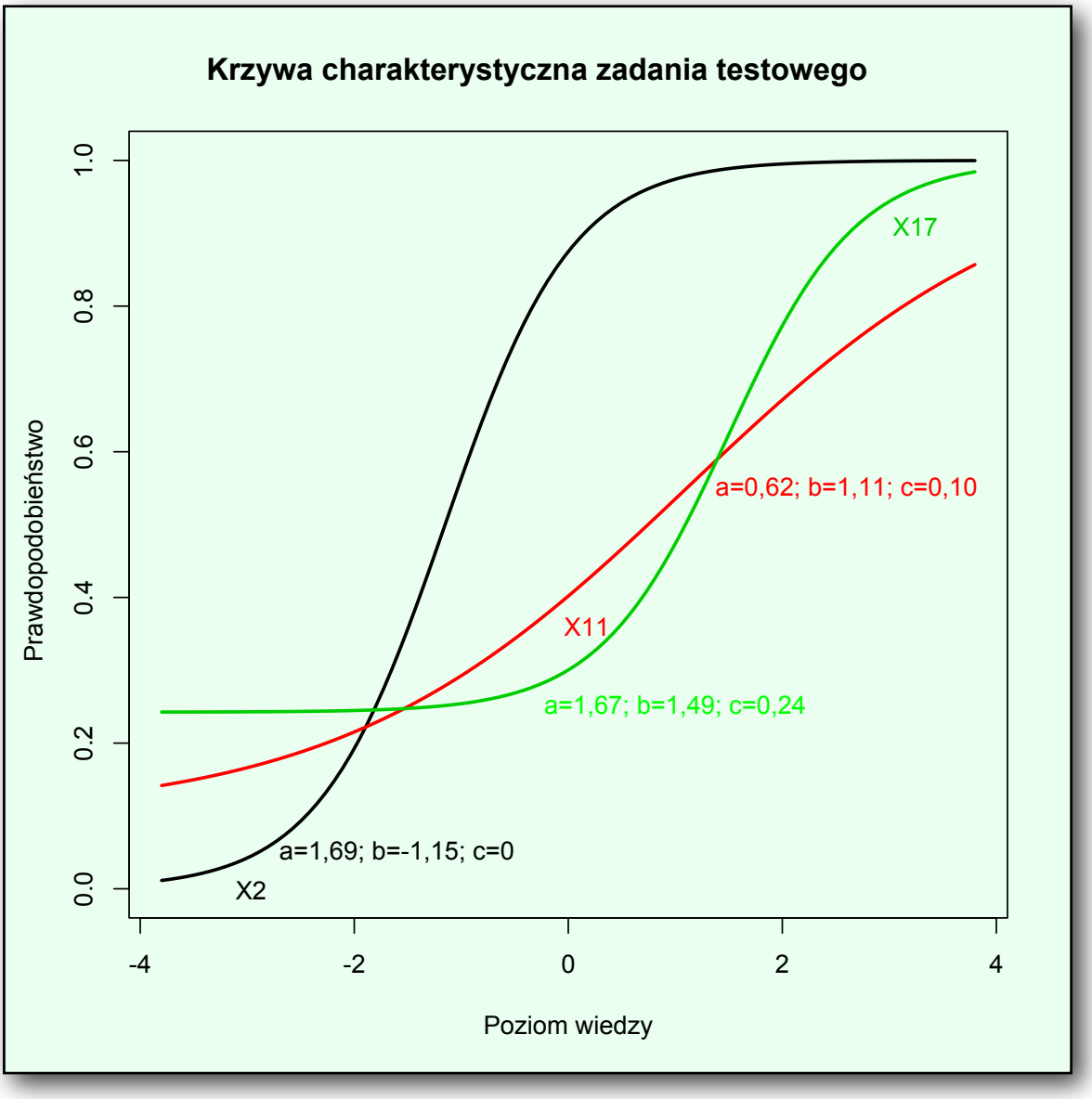
Tabela 4.2. Parametry IRT poszczególnych pozycji testu Omnibus

	c	*b*	*a*		*c*	*b*	*a*
OMN1	0.471	-0.091	1.779	OMN31	0.126	0.129	1.643
OMN2	0	-1.151	1.690	OMN32	0.219	1.535	2.545
OMN3	0.018	-0.403	1.014	OMN33	0.024	-0.135	1.378
OMN4	0	0.064	1.061	OMN34	0.192	2.011	12.282
OMN5	0.329	1.038	1.151	OMN35	0.122	1.920	4.257
OMN6	0.430	1.443	10.408	OMN36	0	0.337	0.973
OMN7	0	-0.700	1.090	OMN37	0.224	1.443	2.095
OMN8	0	-1.234	0.680	OMN38	0.150	1.762	2.682
OMN9	0.249	0.920	2.032	OMN39	0.053	1.061	1.107
OMN10	0.030	0.109	0.623	OMN40	0.286	2.001	1.776
OMN11	0.101	1.109	0.619	OMN41	0	1.210	0.804
OMN12	0.181	2.981	1.247	OMN42	0.115	3.412	0.798
OMN13	0.557	1.713	2.606	OMN43	0.001	0.369	0.678
OMN14	0	-0.977	0.755	OMN44	0.292	1.973	3.427
OMN15	0.463	1.446	2.667	OMN45	0.497	1.177	2.191
OMN16	0.131	1.873	0.721	OMN46	0.068	0.824	1.170
OMN17	0.243	1.492	1.671	OMN47	0.101	0.949	1.555
OMN18	0.253	0.610	1.780	OMN48	0.217	1.111	3.592
OMN19	0.005	1.329	0.353	OMN49	0.196	1.917	1.884
OMN20	0	0.762	0.803	OMN50	0.121	2.085	13.255
OMN21	0	0.255	0.875	OMN51	0.052	2.743	0.680
OMN22	0	-0.625	0.870	OMN52	0.141	1.347	1.002
OMN23	0.001	-0.046	0.768	OMN53	0.164	1.753	3.422
OMN24	0.194	1.861	1.631	OMN54	0.244	1.588	1.452
OMN25	0.134	0.404	0.635	OMN55	0.125	1.268	2.295
OMN26	0.263	0.716	1.561	OMN56	0.236	1.655	1.359
OMN27	0.001	-0.597	0.791	OMN57	0.166	2.550	1.626
OMN28	0.477	-0.121	1.367	OMN58	0.199	1.965	1.457
OMN29	0	0.142	0.790	OMN59	0	-1.059	0.901
OMN30	0.006	0.709	0.417	OMN60	0	1.099	0.749

Objaśnienia: c – parametr zgadywalność, b – parametr trudność, a – parametr moc różnicująca.
Źródło: badania własne.

13 pozycji testowych charakteryzowało się zerowym poziomem zgadywalności, dla dodatkowych 7 pozycji parametr ten nie przekraczał poziomu c = 0,05. Najtrudniejszym zadaniem okazało się pytanie 42 dotyczące wyrażenia frazeologicznego „Bodaj się tacy na kamieniu rodzili", najłatwiejszym: pytanie 8 dotyczące ciągu liczb „1 2 8 48 ?". Analizując poszczególne parametry zadań, których trudność – w zamyśle autorów – miała charakter narastający (Jaworowska, Matczak, 202, s. 15) można zauważyć, iż na podstawie wyników empirycznych, w oparciu o model IRT, teza ta nie znajduje potwierdzenia.

Na poniższej rycinie przedstawiono krzywe charakterystyczne dla 3 przykładowych pozycji testowych, różniących się poziomem trudności, mocy różnicującej oraz zgadywalności.



Rycina 4.4. Krzywe charakterystyczne dla pozycji 2, 11 i 17 testu Omnibus

Pozycja nr 2 („Koherencja znaczy coś przeciwnego niż… a-Nieobecność, b-Niespójność, c-Czołobitność, d-Podległość, e-Uznanie") była najłatwiejszą pozycją testową o zerowym poziomie zgadywalności, jednocześnie charakteryzowała się dobrą mocą różnicującą. Dla porównania pozycja 17 („Reglamentować znaczy coś przeciwnego niż… a-Sygnalizować, b-Szafować, c-Synchronizować, d-Racjonować, e-Rejestrować") była jedną z trudniejszych pozycji testowych o wysokim poziomie parametru zgadywania. Z kolei pozycja 11 („Miecz Damoklesa… a-Bardzo ostry miecz, b-Miecz wodza Wikingów, c-Zabytek muzealny z czasów rzymskich, d-Stałe

zagrożenie, e-Symbol sprawiedliwości") na rycinie 4.4. reprezentuje przeciętną pozycję testową, ze słabą mocą różnicującą.

Kwestionariusz PTS

Rozkład wyników surowych uzyskanych przez osoby uczestniczące w badaniu różnił się od wyników w grupie normalizacyjnej: mężczyźni wykazywali istotnie niższą siłę procesów hamowania, natomiast kobiety przejawiały istotnie wyższą siłę procesów pobudzania oraz ruchliwość procesów nerwowych (por. tabela 4.3)

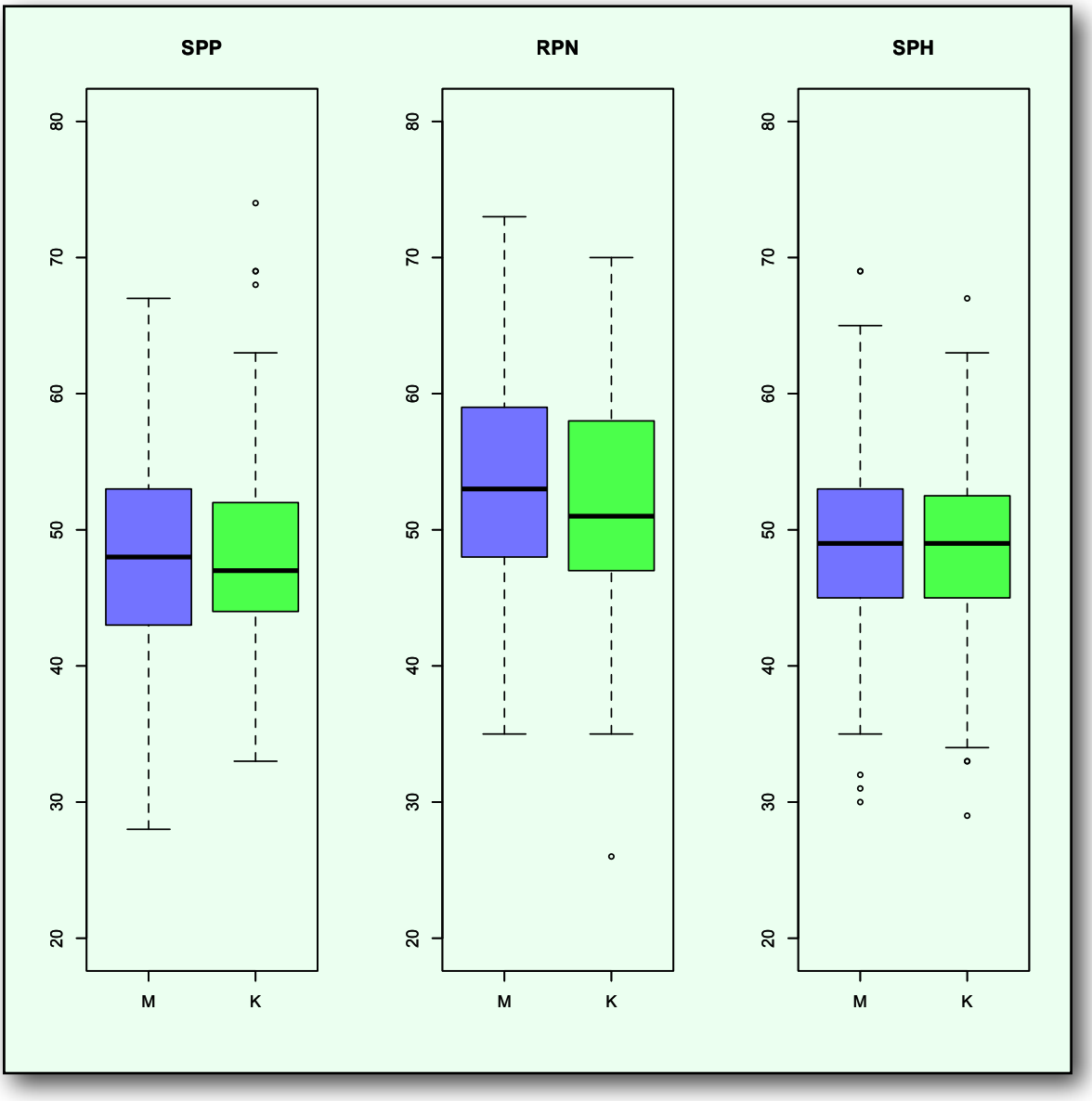
Tabela 4.3. Statystyki opisowe wyników w kwestionariuszu PTS i istotność różnic w stosunku do norm z uwzględnieniem płci

	średnia ± SD						
Skala	mężczyźni	kobiety	skośność	kurtoza	1Q	mediana	3Q
Siła procesów pobudzania	48,5 ± 7,24	48,1 ± 7,72 [b]	0,34	0,44	43	48	53
Siła procesów hamowania	49,0 ± 6,53 [a]	49,2 ± 6,87	-0,13	0,63	45	49	53
Ruchliwość procesów nerwowych	52,2 ± 8,12	53,6 ± 7,97 [c]	0,01	-0,53	47	52	59

Różnica w stosunku do norm: a – istotna dla $p < 0,05$ ($t_{(118)} = -2,26$); b – istotna dla $p < 0,001$ ($t_{(173)} = 5,13$); c – istotna dla $p < 0,01$ ($t_{(173)} = 2,97$); Brak istotnej różnicy między grupami. Źródło: badania własne.

Rozkład wyników przeliczonych z uwzględnieniem płci ilustruje rycina 4.3. Wyniki uzyskane przez 293 osoby uczestniczące w badaniu posłużyły do obliczenia parametrów IRT w oparciu o model GRM (por. tab. 4.4).

Parametr a świadczy o mocy różnicującej danej pozycji kwestionariuszowej, zaś parametry b określają próg poziomu mierzonej cechy między kategoriami odpowiedzi (*between category thresholds*). Parametry b należy interpretować jako wartości (poziomy) mierzonej cechy: siły procesów pobudzenia, hamowania lub ruchliwości procesów nerwowych, które odpowiednio zwiększają prawdopodobieństwo, iż osoby wypełniające kwestionariusz w danym stwierdzeniu wybiorą właśnie taką kategorię odpowiedzi. Na przykład dla pytania 34 w kwestionariuszu PTS (por. ryc. 4.4) osoba o sile procesów pobudzenia ocenianej na Θ = -1 wybierze najprawdopodobniej 2 kategorię odpowiedzi: „raczej się zgadzam". Zaś inna osoba o poziomie SPP wynoszącym np.: Θ = +1 najprawdopodobniej wybierze kategorię 4 – „zdecydowanie się nie zgadzam". Inaczej mówiąc z twierdzeniem: „Łatwo tracę głowę, jeśli znajdę się pod bardzo silną presją" osoby o sile procesów pobudzenia poniżej przeciętnej (łatwo pobudzający się układ nerwowy) raczej się zgodzą, a osoby o odporniejszym na pobudzenie układzie nerwowym – zdecydowanie nie zgodzą się.



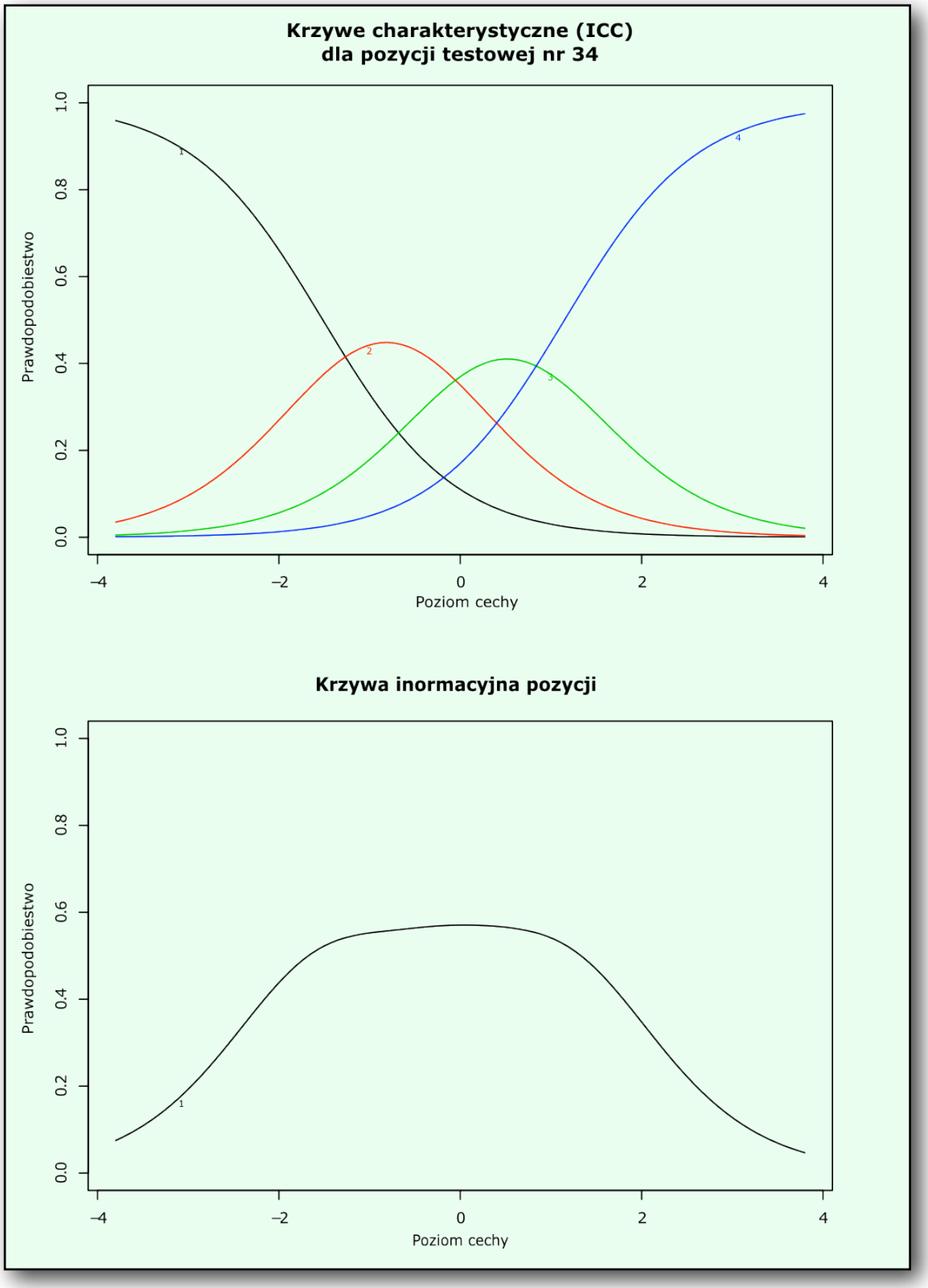
Rycina 4.3. Wykresy skrzynkowe w oparciu o medianę rozkładu wyników w kwestionariuszu PTS z uwzględnieniem płci. SPP - siła procesów pobudzania, RPN - ruchliwość procesów nerwowych, SPH - siła procesów hamowania. Źródło: badania własne.

Tabela 4.4. Parametry mocy różnicującej i trudności dla poszczególnych pozycji kwestionariusza PTS

	Informacja	b_1	b_2	b_3	a		*Informacja*	b_1	b_2	b_3	a
PTS 1	1.61	-3.70	-1.02	1.83	0.74	PTS 30	1.99	-2.02	-0.44	1.78	0.95
PTS 2	0	97.19	-15.94	-113.25	-0.01	PTS 31	1.23	-3.75	-1.34	1.06	0.64
PTS 3	1.06	-3.13	-0.36	2.26	0.56	PTS 32	1.38	-1.89	0.66	2.68	0.70
PTS 4	0.11	-13.96	-7.74	1.38	0.14	PTS 33	0.74	2.42	-0.82	-3.64	-0.43
PTS 5	1.45	-4.70	-2.59	0.18	0.71	PTS 34	2.90	-1.51	-0.12	1.15	1.38
PTS 6	0.40	-10.22	-4.15	3.77	0.28	PTS 35	1.54	-2.57	-0.87	1.65	0.77
PTS 7	0.67	-5.73	-1.25	3.02	0.38	PTS 36	0.70	-3.22	0.30	3.72	0.40
PTS 8	1.13	-2.33	0.11	2.61	0.60	PTS 37	2.04	-2.09	-0.63	1.42	0.99
PTS 9	0.37	-7.24	-1.61	4.50	0.26	PTS 38	1.97	-1.83	0.03	1.77	0.96
PTS 10	0.75	-2.43	0.93	4.58	0.42	PTS 39	0.67	-4.18	-1.75	1.74	0.41
PTS 11	0.60	-3.56	0.23	4.60	0.35	PTS 40	1.70	-3.03	-0.99	1.61	0.81
PTS 12	2.46	-1.71	-0.75	0.38	1.32	PTS 41	0.85	-4.63	-1.28	1.78	0.47
PTS 13	2.30	-1.81	-0.45	1.14	1.14	PTS 42	0	-121.91	-51.30	52.25	0.02
PTS 14	1.28	-3.10	-0.33	1.59	0.66	PTS 43	1.79	-2.73	-0.43	2.13	0.82
PTS 15	2.53	-1.77	-0.67	0.77	1.27	PTS 44	1.63	-2.36	-0.28	2.07	0.79
PTS 16	1.45	-2.36	0.34	2.37	0.72	PTS 45	0.01	-38.63	-10.53	31.43	0.04
PTS 17	2.34	-2.72	-1.23	0.88	1.09	PTS 46	1.43	-3.80	-1.10	1.79	0.68
PTS 18	2.49	-2.21	-0.38	1.35	1.14	PTS 47	2.76	-2.09	-0.78	0.67	1.32
PTS 19	0.18	-12.90	-2.12	6.12	0.17	PTS 48	0.55	-3.74	-0.74	3.37	0.35
PTS 20	2.23	-2.88	-1.00	0.94	1.03	PTS 49	0.75	-3.05	1.06	5.18	0.41
PTS 21	1.96	-3.00	-0.94	1.10	0.92	PTS 50	1.40	-2.03	-0.14	2.90	0.69
PTS 22	2.50	-1.90	-0.60	0.71	1.25	PTS 51	1	-3.24	0.07	3.81	0.50
PTS 23	0.76	-5.77	-2.28	1.57	0.42	PTS 52	0.25	-6.56	1.17	8.99	0.21
PTS 24	2.12	-2.65	-0.81	1.45	0.98	PTS 53	0.85	-3.33	0.84	5.71	0.43
PTS 25	1.13	-3.61	-0.94	2.46	0.57	PTS 54	0.90	-2.57	1.05	5.10	0.46
PTS 26	1.40	-2.37	0.44	2.92	0.68	PTS 55	0	63.14	27.67	-42.96	-0.03
PTS 27	0.57	-1.50	1.78	5.87	0.36	PTS 56	1.83	-2.31	-0.60	1.96	0.87
PTS 28	1.77	-2.29	0.11	2.46	0.82	PTS 57	0.72	-2.36	1.90	5.64	0.40
PTS 29	1.25	-3.68	-1.53	1.28	0.64						

Objaśnienia: a – moc różnicująca danej pozycji, b_1, b_2, b_3 – poziom trudności, dla $\Theta < b_1$ wartość 1 w kluczu odpowiedzi, dla $b_1 < \Theta < b_2$ – wartość 2 w kluczu odpowiedzi, dla $b_2 < \Theta < b_3$ – wartość 3 w kluczu odpowiedzi, dla $\Theta > b_3$ – wartość 4 w kluczu odpowiedzi. Źródło: badania własne.

Trzy pozycje kwestionariuszowe: 2, 42 oraz 55 okazały się pytaniami bardzo słabymi – ich moc różnicująca była zbliżona do zera, a funkcja informacyjna była prawie płaska. Pozycje te nie „działały" w sposób diagnostyczny w badanej grupie, ponieważ poziom wartości cechy, na który były wrażliwe wypadał dla dużych liczb ujemnych. Poniżej na rycinie 4.4. przedstawiono interpretację graficzną parametrów przykładowej pozycji, która miała przeciętne parametry.



Rycina 4.4. Wykresy krzywych charakterystycznych dla kategorii oraz funkcji informacyjnej dla pozycji nr 34 kwestionariusza PTS.

4.6. Testowanie hipotez

4.6.1. Wpływ wielkości próby na dokładność estymacji parametrów modeli IRT

Analizując pierwszą hipotezę założono, że wraz ze zwiększaniem wielkości próby w sposób wyraźny poprawiać się będzie dokładność estymowanych parametrów, czyli będzie malał błąd standardowy pomiaru. W analizie kontrolowano wpływ takich parametrów pozycji, jak: moc różnicująca pozycji testowej lub kwestionariuszowej (a) oraz jej trudność (b). Wyniki zostały wygenerowane z użyciem założonych parametrów populacji w oparciu o model 3PL. Zestawy zakładanych parametrów (por. tab. 3.1) zostały użyte do wygenerowania *quasi*-odpowiedzi dla przyjętego narzędzia badawczego poprzez określenie prawdopodobieństwa danej odpowiedzi na dana pozycję i porównanie go z losową liczbą z przedziału od 0 do 1.

Analiza oparta była o wyniki uzyskane w symulacji metodą łańcuchów Markowa Monte Carlo – zakładając parametry a oraz b generowano losowe 1000 prób n-osobowych wyników, a następnie wyznaczano parametry rozrzutu, skośności, kurtozy i błędu standardowego dla każdego zestawu parametrów (por. tabela 4.5).

Tabela 4.5. Przykładowe wyniki dla porównań próby 50 osobowej z populacją 10.000-czną

		różnica						
a	b	wielkość	istotność	SD	rozrzut	skośność	kurtoza	SEM
niski	niski	-0,00074	0,3898	0,06947	0,46	-0,09584	-0,01346	0,00220
niski	przeciętny	0,00251	0,0135	0,06957	0,48	0,11293	0,34637	0,00220
niski	wysoki	-0,00164	0,0901	0,06302	0,36	0,14154	-0,16718	0,00199
przeciętny	niski	-0,00210	0,8992	0,07166	0,44	-0,12704	-0,03161	0,00227
przeciętny	przeciętny	-0,00247	0,1243	0,06726	0,48	0,11931	0,14598	0,00213
przeciętny	wysoki	0,00076	0,9776	0,05735	0,34	0,03470	-0,19496	0,00181
wysoki	niski	-0,00370	0,8084	0,06720	0,40	-0,07134	-0,14133	0,00213
wysoki	przeciętny	0,00331	0,5763	0,06896	0,44	0,07308	-0,11372	0,00218
wysoki	wysoki	0,00104	0,1136	0,05732	0,34	0,10086	0,07534	0,00181

Podobne dane zostały sporządzone dla 20 wielkości prób – od 50 do 1000, co 50. Oznaczenia: a – poziom mocy różnicującej dla pozycji; b – poziom trudności pozycji, *różnica* – między średnim wynikiem w danej próbie a średnią w quasi-populacji, istotność – na podstawie wyniku testu Wilcoxona, SD – odchylenie standardowe, *rozrzut* – rozstęp wyników w danej próbie, podobnie *skośność* i *kurtoza*; SEM – błąd standardowy pomiaru. Dla każdej kombinacji a oraz b wygenerowano 1000 wyników.

Dysponując średnią różnicą między wynikiem uzyskanym w próbie, a wynikiem prawdziwym dla populacji (wygenerowanej) przeanalizowano wpływ wielkości próby na błąd standardowy pomiaru za pomocą analizy wariancji. Wyniki (por. tabela 4.6)

potwierdzają sprawdzaną hipotezę – wraz ze zwiększeniem próby maleje błąd standardowy pomiaru ($F_{(19, 158)} = 541{,}9$; $p < 0{,}001$).

Tabela 4.6. Parametry oszacowań ze względu na wielkość próby oraz wyniki analizy wariancji

	różnica						
n	wielkość	istotność	*SD*	rozrzut	skośność	kurtoza	*SEM*
50	-0,000337	0,444	0,066	0,416	0,032	-0,011	0,00208
100	0,001440	0,284	0,047	0,298	0,059	-0,077	0,00149
150	-0,000223	0,569	0,038	0,248	0,027	-0,076	0,00121
200	0,000158	0,531	0,033	0,216	0,011	-0,048	0,00105
250	0,000512	0,623	0,030	0,195	0,030	0,022	0,00094
300	0,000229	0,356	0,027	0,168	0,015	-0,059	0,00086
350	0,000143	0,451	0,025	0,163	0,062	-0,039	0,00078
400	-0,000101	0,466	0,023	0,148	0,020	0,007	0,00074
450	0,000002	0,266	0,022	0,148	0,000	-0,021	0,00070
500	-0,000173	0,413	0,021	0,136	0,028	-0,079	0,00066
550	0,000003	0,470	0,020	0,130	0,031	0,045	0,00063
600	-0,000282	0,546	0,019	0,125	0,051	0,014	0,00060
650	0,000088	0,627	0,018	0,117	0,041	-0,112	0,00058
700	-0,000263	0,522	0,018	0,115	-0,012	0,029	0,00056
750	-0,000004	0,474	0,017	0,111	0,011	0,012	0,00054
800	0,000033	0,547	0,016	0,107	0,040	0,046	0,00052
850	0,000133	0,580	0,016	0,100	-0,001	-0,070	0,00050
900	0,000111	0,436	0,015	0,097	0,007	-0,119	0,00049
950	0,000137	0,495	0,015	0,099	0,035	0,079	0,00048
1000	0,000165	0,373	0,015	0,093	-0,015	-0,049	0,00047
F	1,443	0,883	541,911	163,785	0,630	1,333	541,911
Istotność	0,114	0,605	< 0,001	< 0,001	0,879	0,170	< 0,001

Współzmienne a oraz b występujące w modelu zostały oszacowane jako następujące wartości: $a = 0{,}987$, $b = 0{,}5$. Każde F testuje efekt wielkości próby. Ten test jest oparty na liniowo niezależnych porównaniach parami pomiędzy oszacowanymi średnimi brzegowymi. $df_1 = 19$, $df_2 = 158$.

Sam fakt zmniejszania się błędu wraz ze wzrostem wielkości próby nie jest niczym zaskakującym – wszak maleje względne zróżnicowanie wyników (w stosunku do liczby obserwacji w próbie) a tym samym wariancja. Za istotny należy uznać fakt, że dla wielkości różnicy między wynikiem prawdziwym a estymowanym, nieistotny jest wpływ poziomu mocy różnicującej pozycji ($F_{(2, 170)} = 0{,}85$; $p = 0{,}430$), a istotna jest jej trudność ($F_{(2, 170)} = 5{,}31$; $p = 0{,}006$). Analiza kontrastów z poprawką Bonferroniego wskazuje, że najtrudniejsze pozycje ($b > 1{,}0$) różnią się od przeciętnych zadań wtedy, gdy poziom mocy różnicującej jest wysoki ($a > 1{,}35$; por. tabela 4.7).

Tabela 4.7. Istotność różnic ze względu na poziom trudności pozycji b z uwzględnieniem poziomu mocy różnicującej a. Zmienna zależna: różnica między wynikiem estymowanym a prawdziwym.

poziom mocy różnicującej		Suma kwadratów	df	Średni kwadrat	F	Istotność
niski	trudność	0,000001992	2	0,000000996	1,116	0,330
	błąd	0,000151721	170	0,000000892		
przeciętny	trudność	0,000005130	2	0,000002565	2,874	0,059
	błąd	0,000151721	170	0,000000892		
wysoki	trudność	0,000005897	2	0,000002948	3,304	0,039
	błąd	0,000151721	170	0,000000892		

Źródło: badania własne

Jakie jest praktyczne znaczenie uzyskanych wyników? Otóż podejmując decyzję o zbudowaniu skróconej wersji kwestionariusza lub testu badacz musi mieć świadomość, iż wybieranie pozycji o wysokim poziomie mocy różnicującej nie przynosi samych korzyści. Oczywiście test lub kwestionariusz składający się z takich pozycji będzie lepiej stratyfikował osoby uczestniczące w badaniu pod względem badanej cechy, lecz jednocześnie wzrośnie też prawdopodobieństwo popełnienia błędu w jej szacowaniu. Prawdopodobieństwo błędu jest tym większe, im trudniejsze pozycje zostaną włączone do nowo powstałej wersji. Kwestia wielkości próby okazuje się tutaj sprawą drugorzędną i nie jest niespodzianką konstatacja, iż im większa próba tym lepiej – jej wielkość oparta będzie raczej na zakładanym poziomie błędu pomiarowego (a w praktyce na przesłankach ekonomicznych), niż na jej wpływie na dokładność estymacji.

4.6.2. Wpływ długości skróconej wersji testu na wyniki

Wraz z decyzją o doborze pozycji do zbudowania skróconej wersji danego narzędzia badawczego, badacz musi podjąć decyzję o liczbie pozycji, które wykorzysta. Tego problemu jest pozbawiony w przypadku badania adaptacyjnego, gdzie o liczbie wykorzystanych pozycji decydują odpowiedzi udzielone przez osoby uczestniczące w badaniu, w relacji do reguły zatrzymania algorytmu testowania. Tym niemniej w procesie budowania „statycznej" skróconej wersji narzędzia badawczego interesujący jest wpływ liczby wykorzystanych pozycji na wynik i jego dokładność. Aby to sprawdzić sporządzono skrócone wersje testu Omnibus w oparciu o model 3PL. W celu uniknięcia tendencyjności w doborze pozycji testowych, wykorzystano podejście oparte na MCMC losując z 60 dostępnych pozycji zestawy liczące od 5 do 55 pozycji testowych. Uwzględniając odpowiedzi osób badanych (N = 306) uzyskano w ten sposób 15606

zestawów wyników (51 wersji x 306 osób w grupie) z różnych skróconych wersji testu Omnibus. Dla każdej i-tej długości uśredniono zmienne kontrolowane i wyjaśniane dla każdej z 306 osób uczestniczących w badaniach (liczba iteracji = 50). Zmiennymi niezależnymi kontrolowanymi były uśrednione parametry a, b oraz c pozycji testowych wchodzące w skład skróconej wersji testu. Zmiennymi zależnymi były: wynik surowy (suma poprawnych odpowiedzi), wynik przeliczony (Θ) oraz wynik uzyskany przez daną osobę w pełnej wersji narzędzia badawczego. Rozkłady wyników okazały się niesymetryczne (test Kołgomorowa-Smirnowa $p < 0{,}001$) i dlatego do zbadania prostych zależności liniowych użyto współczynnika korelacji rang Spearmana (porównaj tabela 4.8).

Tabela 4.8. Korelacje między długością skróconej wersji testu a wynikiem osób uczestniczących w badaniach oraz parametrami IRT.

		Uśrednione parametry				Wynik		
Zmienna	Statystyka	a	b	c	SEM	WS	Θ	WP
długość [i]	korelacja	0,012	0,008	-0,005	-0,936**	-0,001	0	0
	istotność	0,127	0,308	0,552	< 0,001	0,944	1,000	1,000
wynik surowy [WS]	korelacja	-0,011	-0,021**	0,008			,960**	,996**
	istotność	0,170	0,008	0,293			< 0,001	< 0,001
wynik przeliczony [Θ]	korelacja	-0,003	0,000	0,007				,964**
	istotność	0,676	0,979	0,412				< 0,001
wynik pełny [WP]	korelacja	-0,003	0,003	0,007	SEM	0,149**	0,043**	0,149**
	istotność	0,681	0,737	0,369		< 0,001	< 0,001	< 0,001

WS – wynik surowy, WP – wynik przeliczony, N = 306 osób x 51 wersji narzędzia = 15606. Źródło: badanie własne.

Jak wynika z analizy wartości współczynników korelacji, nie istnieje liniowy związek między wynikiem uzyskiwanym przez osoby uczestniczące w badaniach a długością skróconej wersji. Podobnie nie zaobserwowano zależności między liczbą pozycji a ich parametrami. Zgodnie z przewidywaniami, istnieje natomiast ujemna zależność między długością wersji skróconej a wielkością błędu pomiarowego – im dłuższe narzędzie, tym mniejszy błąd. Wykazano także, iż wynik przeliczony jest najmniej wrażliwy na wielkość błędu – współczynnik determinacji dla Θ od SEM wynosi tylko 0,185% zaś dla wyniku surowego i przeliczonego – 2,201%.

Dodatkowo przeprowadzono analizę regresji liniowej dla liczby pozycji testowych oraz ich parametrów, uzyskując nieistotne parametry dla wyników pełnych ($F_{(4, 15601)} = 0{,}59$; $p = 0{,}668$) i przeliczonych ($F_{(4, 15601)} = 0{,}39$; $p = 0{,}818$). Natomiast dla wyniku surowego, czyli sumy poprawnych odpowiedzi istotne ($p < 0{,}01$) okazały się

parametry b oraz c – trudność pozycji testowej i poziom zgadywalności danej pozycji są istotnymi predyktorami wyniku surowego (por. tabela 4.9).

Gdy dokonano przekształcenia i w miejsce zmiennej zależnej zamiast wyniku przyjęto wartość bezwzględną z różnicy między wynikiem otrzymanym a oczekiwanym, to wpływ długości narzędzia okazał się istotny ($F_{(1,\ 178)}$ = 83,9; p < 0,001). Wpływ ten nie był prostoliniowy. Najlepiej zależność tę wyjaśnia funkcja odwrotna (wzór 4.1; skorygowane R^2 = 31,6% w porównaniu do funkcji liniowej: skorygowane R^2 = 20,1%).

$$\Delta = 0,089 \cdot \frac{1}{i}, \qquad (4.1),$$

gdzie i to liczba pozycji testowych lub kwestionariuszowych.

Zależność o podobnym kształcie zaobserwowano podczas analizy wpływu wielkości próby na zróżnicowanie średnich wyników – przy zwiększaniu długości testu początkowo różnica szybko się zmniejsza, by po chwili „wyhamować" i zbliżać się asymptotycznie do zera.

Tabela 4.9. Współczynniki uzyskane w analizie regresji.

	Współczynniki niestandaryzowane	Współczynniki standaryzowane				Korelacje		
predyktory	β	Beta	SEM	t	Istotność	Rzędu zerowego	Cząstkowa	Częściowe (semicząstkowe)
(Stała)	28,930		3,321	8,712	<0,001			
średnie a	-1,698	-0,012	1,180	-1,439	0,150	-0,013	-0,012	-0,012
średnie b	-8,868	-0,027	2,870	-3,090	0,002	-0,022	-0,025	-0,025
średnie c	55,769	0,025	19,976	2,792	0,005	0,011	0,022	0,022
i	0,00004081	<0,001	0,005	0,009	0,993	-0,001	<0,001	<0,001

Zmienna zależna: wynik surowy, a,b,c – uśrednione parametry pozycji testowych, i – liczba pozycji testowych. Istotność modelu: $F_{(4,\ 15601)}$ = 3,96; p = 0,003). Źródło: badania własne.

Podsumowując wyniki można stwierdzić, że ponieważ zależność między długością testu a wielkością różnicy między wynikiem prawdziwym i obserwowanym jest krzywoliniowa, istnieje optymalna długość testu. Test zbyt krótki daje wyniki mocno obciążone błędem, z kolei test dłuższy wcale nie znaczy lepszy. Co prawda dokładność wzrasta wraz z długością testu, ale związek jest krzywoliniowy i dość szybko osiąga punkt przegięcia, poza którym dalsze wydłużanie narzędzia nie przynosi wymiernych korzyści (por. ryc 4.1).

Ciekawą zależnością jest związek wyniku surowego z poziomem błędu, który to związek zanika, jeśli wynik jest szacowany w oparciu o IRT i wyrażony na skali Θ. Pokazuje to wadę opierania wyniku końcowego na prostej sumie wyników pozycji

testowych bez uwzględniania ich wag. Wysoki poziom trudności pozycji testowych oddziaływuje negatywnie na wynik surowy, podobnie zgadywalność – im wyższa, tym wyższa liczba poprawnych odpowiedzi. Z kolei gdy pozycje testowe są łatwe, to wynik testu będzie wyższy, co ma miejsce właśnie wtedy, gdy jest on oparty o sumę punktów poszczególnych pozycji testowych.

4.6.3. Równoważność wyników skróconych wersji testu

Założono, iż stopień trudności testu w wersji skróconej nie będzie miał wpływu na wyniki uzyskiwane przez osoby uczestniczące w badaniu. Jednakże dla wyników surowych powinna zostać zaobserwowana różnica, ponieważ osoba o tym samym poziomie Θ test łatwy powinna rozwiązać lepiej, a trudny gorzej – co nie ma wpływu na szacowanie jej poziomu mierzonej cechy Θ.

Aby sprawdzić tak postawioną hipotezę sporządzono skrócone wersje testu, w oparciu o poziom trudności poszczególnych pozycji dobierając je tak, aby otrzymać wersję składającą się z pytań łatwych, przeciętnych oraz trudnych. Dodatkowo sporządzono także czwartą wersję skróconą, do której wybrano pozycje testowe w sposób losowy. Tabela 4.10 przedstawia statystyki opisowe dla czynników rozumienia i wiedzy oraz dla surowego wyniku ogólnego z uwzględnieniem wersji testu.

Tabela 4.10. Statystyki opisowe wyników surowych skróconych wersji testu

Wersja	Średnia ± SD	Mediana	Min	Max	Skośność	SEM
surowy wynik ogólny ($F_{(3, 1228)}$ = 177,0; p < 0,001; d = 0,25; 95%CI = (0,12; 0,38))						
1	11.75 ± 4.12 [a]	12	1	20	-0.37	0.23
2	8.10 ± 3.56 [b]	8	2	20	0.71	0.20
3	5.49 ± 2.81 [c]	5	0	17	1.10	0.16
4	7.59 ± 3.20 [b]	7	0	19	0.54	0.18
czynnik rozumienia – wynik surowy ($F_{(3, 1228)}$ = 135,7; p < 0,001; d = 0,27; 95%CI = (0,13; 0,40))						
1	6.91 ± 2.69 [a]	7	0	12	-0.40	0.15
2	5.06 ± 2.26 [b]	5	1	12	0.49	0.13
3	3.20 ± 1.96 [c]	3	0	11	1.02	0.11
4	5.68 ± 2.33 [d]	6	0	12	-0.08	0.13
czynnik wiedzy – wynik surowy ($F_{(3, 1228)}$ = 149,1; p < 0,001; d = 0,78; 95%CI = (0,65; 0,91))						
1	4.88 ± 2.01 [a]	5	0	8	-0.30	0.11
2	3.04 ± 1.96 [b]	3	0	8	0.58	0.11
3	2.29 ± 1.55 [c]	2	0	8	0.59	0.09
4	4.88 ± 2.00 [a]	5	0	8	-0.30	0.11

Wersja testu skróconego: 1 – łatwa, 2 – przeciętna, 3 – trudna, 4 – losowa. a, b, c, d – grupy jednorodne dla p < 0,05. N = 308. Istotność różnic obliczona za pomocą analizy wariancji i potwierdzona za pomocą testu median. Źródło: badania własne.

Zgodnie z oczekiwaniami, dla wyników surowych (suma odpowiedzi poprawnych) poszczególnych skal testu Omnibus (a więc i dla wyniku sumarycznego) zaobserwowano istotne różnice. Najwięcej poprawnych odpowiedzi było w łatwej wersji testu, a najmniej w wersji trudnej. Wyniki w wersji losowej dla czynnika rozumienie plasowały się pomiędzy wynikami wersji łatwej i przeciętnej (grupy jednorodne na podstawie HSD Tukey'a, $p < 0{,}05$). Dla czynnika wiedzy wersja losowa miała wynik taki sam jak wersja łatwa (grupy jednorodne na podstawie HSD Tukey'a, $p < 0{,}05$).

Do analizy różnic między wynikami przeliczonymi wykorzystano jednoczynnikową analizę wariancji dla powtarzanych pomiarów. Zgodnie z założeniami, przeliczony poziom inteligencji, czynnika wiedzy i czynnika rozumienia dla wszystkich testowanych wersji nie powinny się różnić, ponieważ dobór pozycji do skróconej wersji danego narzędzia nie powinien mieć wpływu na wyniki. Hipoteza ta bazuje na założeniu o lokalnej niezależności pozycji testowych w modelach IRT. Uzyskane wyniki przedstawia tabela 4.11.

Tabela 4.11. Różnice między przeliczonymi wynikami ogólnymi między różnymi wersjami narzędzia ($F_{(1, 306)} = 23637{,}9$; $p < 0{,}001$)

wersja	średni wynik przeliczony ($SEM = 0{,}005$)	przedział ufności 95%	
		dolna granica	górna granica
łatwa	0,537 [a]	0,526	0,547
przeciętna	0,535 [a]	0,525	0,545
trudna	0,626 [b]	0,617	0,636
losowa	0,576 [c]	0,567	0,584

$N = 308$; *a, b, c* – grupy jednorodne na podstawie porównań parami z poprawką Bonferroniego ($p < 0{,}05$).
Źródło: badania własne.

Zgodnie z oczekiwaniami teoretycznymi, wynik przeliczony wersji łatwej nie różni się od wyniku przeliczonego wersji przeciętnej, ale już dla wersji trudnej zaobserwowano istotny wzrost wyniku, co może być spowodowane tym, że poprawna odpowiedź na trudną pozycję testową przynosi więcej „korzyści". Test składający się tylko z pozycji, które cechuje wysoka trudność, a tym samym punkt przegięcia krzywej charakterystycznej leży wysoko na osi badanej cechy, uniemożliwia precyzyjne oszacowanie wyniku osoby badanej o przeciętnej wartości cechy, uwypuklając różnicę między osobami, które odpowiadają poprawnie i niepoprawnie na takie właśnie pozycje. Innymi słowy, test taki dyskryminuje wyniki niskie i przeciętne, działając zgodnie z zasadą „wszystko albo nic". Podobne wnioski można wysnuć analizując korelacje między wynikiem Θ reprezentującym wartość mierzonej cechy a wynikami

przeliczonymi dla danych wersji skróconej (patrz tabela 4.12). Wraz ze wzrostem poziomu cechy Θ, najszybciej rosną wyniki w łatwej wersji testu, wolniej zaś w wersji średniej i losowej. Dla wersji trudnej zaobserwowano natomiast najsłabszą korelację z wartością cechy oraz najsłabszą ujemną korelację z poziomem błędu.

Tabela 4.12. Korelacje między wynikami z różnych wersji a wynikiem przeliczonym (Θ) i błędem pomiaru (*SEM*).

			wyniki surowe		
wyniki w wersjach	Θ	SEM	wynik ogólny	czynnik rozumienia	czynnik wiedzy
losowej	,858** a	-,770**	,867**	,795**	,653**
łatwej	,922** b	-,868**	,855**	,722**	,736**
średniej	,836** a	-,720**	,891**	,717**	,761**
trudnej	,536** c	-,392**	,713**	,579**	,618**
wyniki surowe					
wynik ogólny	,953**	-,833**			
czynnik rozumienia	,780**	-,678**			
czynnik wiedzy	,838**	-,723**			

** – korelacja jest istotna na poziomie 0,01 (dwustronnie), a, b, c – różnice między współczynnikami korelacji określono za pomocą transformacji z Fishera.

Ten ostatni wynik wskazuje na to, że w wersjach trudnych narzędzi badawczych najwolniej będzie redukowany SE, a tym samym będą one stosunkowo dłuższe niż wersje łatwe i przeciętne. Wnioski praktyczne z uzyskanych wyników wskazują, że badacz konstruując skróconą wersję kwestionariusza lub testu nie może zupełnie swobodnie dobierać pozycji, ponieważ tworząc wersję składającą się głównie z trudnych pozycji, otrzyma narzędzie do oceny osób o wysokiej wartości danej cechy. Jednocześnie wyniki uzyskiwane za pomocą takiej trudnej wersji przez osoby uczestniczące w badaniach o przeciętnej i niskiej wartości cechy, będą obciążone większym błędem i zajmować będą relatywnie więcej czasu.

4.6.4. Równoważność skróconych wersji kwestionariuszy konstruowana za pomocą różnych technik

Rozpatrzę teraz użycie modelu IRT do skonstruowania skróconej wersji kwestionariusza w porównaniu do najczęściej używanych metod opartych na analizie czynnikowej (CFA) lub analizie regresji (MR). To, w jaki sposób będzie skracana forma pełna kwestionariusza często jest uzależnione od celu stawianego przed taką wersją. Tutaj dla celów porównawczych założono skrócenie każdej ze skal kwestionariusza temperamentu PTS z 19 do 8 pozycji, wybierając za każdym razem te pozycje, które w danym paradygmacie będą miały najwyższe parametry – ładunek czynnikowy dla

analizy czynnikowej, korelację semicząstkową dla analizy regresji i poziom informacji dla teorii odpowiadania na pozycje testowe.

W pierwszym kroku na wynikach zebranych wśród 293 osób przeprowadzono analizę czynnikową metodą głównych składowych z rotacją Varimax, ustalając liczbę czynników na 3. Miara adekwatności doboru próby KMO wyniosła 0,73, a wynik testu sferyczności Bartletta okazał się istotny na poziomie $p < 0{,}001$. Rozkład uzyskanych ładunków czynnikowych przedstawia tabela 4.13.

Tabela 4.13. Rozkład ładunków czynnikowych. Ładunki < 0,30 zostały ukryte.

	Czynniki					Czynniki			
nr pytania	1	2	3	czynnik	nr pytania	1	2	3	czynnik
1		0,484		SPP	30	0,385			SPH
2		0,328		SPH	31			0,336	RPN
3				SPP	32	**0,522**			SPP
4			**0,402**	SPH	33				SPH
5		0,335		RPN	34	**0,579**			SPP
6			**0,425**	SPH	35	0,396			RPN
7	0,408			RPN	36	0,516			SPH
8		0,337		SPP	37	0,451			RPN
9			**0,575**	SPH	38	**0,461**			SPP
10	**0,419**			SPP	39	0,364			SPH
11	0,425			SPH	40			0,431	SPP
12	0,381	**0,524**		RPN	41		**0,489**		RPN
13		**0,598**		RPN	42			**0,499**	SPH
14		0,557		SPP	43			0,355	SPP
15	0,363	**0,482**		RPN	44	**0,452**			SPP
16				SPP	45			**0,435**	SPH
17		**0,459**		RPN	46			0,559	RPN
18		0,537		SPP	47		**0,556**		RPN
19		**0,336**		RPN	48	0,376			SPH
20		0,384	0,366	RPN	49	**0,424**			SPP
21		0,420		SPP	50	**0,474**			SPP
22	0,407	**0,499**		RPN	51			0,481	SPP
23			**0,600**	SPH	52	0,409			SPH
24		**0,355**	0,424	RPN	53	0,581			SPH
25			**0,639**	SPH	54	**0,554**			SPP
26	0,336			RPN	55			**0,416**	SPH
27	0,405			SPH	56	0,489			RPN
28	0,483			RPN	57	0,501			SPH
29				SPP					

Źródło: badania własne.

Pogrubieniem zaznaczono pozycje włączone do danej skróconej wersji. Wybrano tylko te pozycje, które miały ładunek czynnikowy w danej składowej zgodny z układem teoretycznym. Warto zaznaczyć, że w badanej grupie struktura otrzymanych ładunków

nie pokrywa się w pełni ze strukturą teoretyczną, a procent wyjaśnianej wariancji przez trzy ustalone składowe wynosił 24%.

Kolejnym krokiem było przeprowadzenie analizy regresji liniowej dla każdej ze skal metodą wprowadzania, za zmienną zależną przyjmując wynik sumaryczny w danym czynniku, a za predyktory – wyniki odpowiedzi na poszczególne pytania. Wyniki analizy przedstawia tabela 4.14. Do wersji skróconej na podstawie poziomu korelacji semicząstkowych, wybrano te pozycje, które wykazywały najsilniejszy związek z wynikiem ogólnym (w tabeli zaznaczone są one pogrubieniem).

Tabela 4.14. Wyniki analizy regresji dla pozycji kwestionariusza względem wyniku danego czynnika

SPP	*Beta*	*r*	r_p	*RPN*	*Beta*	*r*	r_p	*SPH*	*Beta*	*r*	r_p
1	0,123	0,369	0,108	5	0,104	0,369	0,096	**2**	**0,147**	**0,366**	**0,139**
3	**0,141**	**0,295**	**0,135**	**7**	**0,119**	**0,249**	**0,113**	**4**	**0,153**	**0,372**	**0,139**
8	**0,145**	**0,424**	**0,130**	**12**	**0,134**	**0,61**	**0,103**	6	0,128	0,282	0,117
10	**0,138**	**0,332**	**0,127**	13	0,125	0,543	0,102	9	0,145	0,455	0,125
14	**0,139**	**0,307**	**0,126**	15	0,123	0,569	0,097	**11**	**0,151**	**0,268**	**0,142**
16	**0,188**	**0,338**	**0,173**	17	0,109	0,486	0,093	23	0,142	0,453	0,118
18	0,13	0,457	0,106	**19**	**0,116**	**0,314**	**0,103**	25	0,141	0,468	0,116
21	0,132	0,459	0,11	20	0,109	0,435	0,094	**27**	**0,155**	**0,298**	**0,144**
29	**0,131**	**0,271**	**0,122**	22	0,126	0,572	0,093	**30**	**0,145**	**0,355**	**0,136**
32	**0,137**	**0,408**	**0,123**	24	0,111	0,445	0,095	**33**	**0,157**	**0,044**	**0,149**
34	0,140	0,523	0,117	**26**	**0,117**	**0,366**	**0,106**	36	0,155	0,522	0,132
38	0,138	0,495	0,120	**28**	**0,114**	**0,356**	**0,104**	**39**	**0,158**	**0,413**	**0,148**
40	0,131	0,332	0,118	31	0,12	0,43	0,11	42	0,148	0,342	0,133
43	0,125	0,42	0,109	**35**	**0,121**	**0,432**	**0,106**	**45**	**0,158**	**0,345**	**0,144**
44	0,136	0,441	0,120	**37**	**0,119**	**0,45**	**0,103**	48	0,159	0,399	0,147
49	0,132	0,354	0,119	41	0,121	0,429	0,102	52	0,142	0,314	0,13
50	**0,137**	**0,414**	**0,123**	46	0,111	0,334	0,099	53	0,135	0,391	0,115
51	0,132	0,295	0,114	47	0,119	0,631	0,093	55	0,144	0,337	0,133
54	0,132	0,364	0,116	**56**	**0,117**	**0,422**	**0,103**	57	0,143	0,363	0,124

Objaśnienia: r – korelacje rzędu zerowego, r_p – korelacje częściowe (semicząstkowe).
Źródło: badania własne.

Trzecią wersję skróconą dla kwestionariusza temperamentu PTS sporządzono w oparciu o wyniki analizy probabilistycznej – na podstawie wielkości funkcji informacyjnej każdej pozycji wybrano po 8 pozycji dla każdego czynnika kwestionariusza temperamentu (por. tabela 4.4). W skład narzędzi skróconych weszły te pozycje, które miały najwyższy poziom funkcji informacyjnej.

Wszystkie trzy sposoby skracania kładą akcent na wymiar centralny faworyzując te pozycje, które są silniej z nim związane lub są obciążone mniejszym błędem pomiarowym. Tym niemniej skład skróconych wersji różni się i można zaobserwować, że dla każdej z nich wybierane są różne zestawy pozycji (tabela w załączniku nr 2).

Współczynnik zgodności α Krippendorffa (Krippendorff, 2004) dla trzech czynników temperamentu wynosił odpowiednio: -0,284; -0,070; -0,213, co wskazuje na niezgodność porównywanych zbiorów pozycji, nie odbiegającą od przypadkowości. Analiza parami wersji narzędzia (FA, MR oraz IRT) za pomocą współczynnika zgodności $\varkappa$ Cohena (Cohen, 1960) wykazuje brak podobieństwa zestawów pozycji kwestionariuszowych. Jednym wyjątkiem jest istotny, lecz ujemny współczynnik dla czynnika SPP i pozycji kwestionariusza wybranych za pomocą MR oraz IRT ($\varkappa$ = -0,73; T = 3,170; p = 0,002). Analiza tego wyniku ujawnia, że istotność spowodowana jest wskazaniem przez obie metody w 16 przypadkach przeciwstawnych pozycji – te wskazane przez analizę MR są wykluczone w oparciu o IRT i *vice versa*, a dodatkowo dla 3 pozycji kwestionariusza zgodność dotyczy wyłączania ich ze skróconej wersji czynnika SPP.

Wiedząc, że skrócone wersje różnią się doborem pozycji kwestionariuszowych porównano je pod względem rzetelności. Obliczono współczynnik α Cronbacha (1951) dla poszczególnych składowych temperamentu. Zaobserwowano dla IRT gorszą wewnętrzną spójność w stosunku do skali pełnej dla czynnika SPP (por. tabela 4.15) i jednocześnie wzrost tej spójności dla czynników SPH i RPN.

Tabela 4.15. Analiza rzetelności skróconych wersji kwestionariusza według 3 podejść: analizy czynnikowej, analizy regresji i teorii odpowiadania na zadania testowe oraz korelacja wyników z wynikami pełnej wersji

Metoda	Skale	Alfa		średnia korelacja	średnia	SD	korelacja z pełną wersją
		surowe	standaryzowane				
FA	SPP	0,638	0,638	0,180	2,382	0,517	,952***
	RPN	0,049	0,073	0,010	2,414	0,366	,910***
	SPH	0,652	0,654	0,191	2,159	0,526	,697***
MR	SPP	0,148	0,170	0,025	2,375	0,406	,905***
	RPN	0,516	0,513	0,116	2,591	0,465	,958***
	SPH	0,270	0,271	0,044	2,398	0,419	,876***
IRT	SPP	0,198	0,223	0,035	2,352	0,369	,863***
	RPN	0,185	0,186	0,028	2,475	0,381	,927***
	SPH	0,370	0,377	0,070	2,372	0,429	,885***

Rzetelność pełnych skal wynosiła odpowiednio: SPP: α = 0,37, RPN: α = 0,36, SPH: α = 0,51.
*** – p < 0,001. Źródło: badania własne.

Podsumowując trudno jednoznacznie wskazać, które z podejść jest lepsze lub gorsze. Na pewno ze względu na sposób obliczeń i obecność technik w pakietach statystycznych bardziej dostępne są metody skracania oparte na analizie czynnikowej

lub analizie regresji niż na teorii odpowiadania na zadania testowe. Mimo zróżnicowanego doboru pozycji kwestionariuszowych do poszczególnych wersji narzędzia mierzącego temperament, analizowane metody dostarczają równie rzetelnych narzędzi. Najsłabiej wypadła analiza czynnikowa, w oparciu o którą uzyskano zestaw pozycji kwestionariuszowych dla RPN o bardzo niskim współczynniku rzetelności α Cronbacha (0,073) i najsłabszą korelację dla czynnika SPH z wynikiem pełnej wersji (r = 0,697 istotnie mniejsze od współczynników dla MR oraz IRT, na podstawie transformacji Fishera $p < 0{,}0001$).

4.6.5. Wersje papier-i-ołówek oraz adaptacyjna a wyniki kwestionariuszy osobowości

Kolejnym aspektem używania skróconego narzędzia badawczego jest wielkość błędu standardowego związanego z medium, jakie zostało użyte do zebrania wyników. Wiadomo (por. rozdział 2.4.1), że pełne wersje papierowe i komputerowe dostarczają takich samych wyników, natomiast otwartym pozostaje pytanie o obciążenie błędem wyników zebranych za pomocą wersji skróconych. W tym celu obliczono wielkość błędu w oparciu o rzetelność dla wyników otrzymanych za pomocą trzech wersji (1. pełnej, 2. skróconej w oparciu o IRT oraz 3. adaptacyjnej), przy każdą z nich zastosowano w badaniach przeprowadzonych za pośrednictwem internetu i tradycyjnie – metodą „papier-i-ołówek" (por. tab. 4.16).

Tabela 4.16. Wielkość błędu standardowego w poszczególnych wersjach kwestionariusza PTS.

wersja kwestionariusza		SEM
internet	pełna	2,944
	skrócona	5,880
	GRM	5,331
papier	pełna	6,661
	skrócona	6,752
ogółem	internet	6,090
	papier	6,701

Źródło: badania własne.

Analiza otrzymanych wyników pokazuje, że najmniejszy błąd uzyskano dla narzędzia pełnego użytego w internecie. Jest to związane z automatyczną walidacją wpisywanych wyników i zredukowaniem w ten sposób pomyłek mechanicznych oraz braków danych. Na drugim miejscu są obie wersje skrócone – nieznacznie różniące się od siebie poziomem błędu standardowego. Zredukowanie długości narzędzia

spowodowało wzrost poziomu błędu, lecz jest on i tak niższy, niż poziom błędu w pełnej wersji papierowej. Podsumowując: zastosowanie komputerów w procesie zbierania wyników pozwoliło w każdym przypadku obniżyć błędy standardowe pomiaru w stosunku do takich samych wersji papierowych.

4.6.6. Wpływ wersji narzędzia na zróżnicowanie wyników w teście i kwestionariuszu

Aby sprawdzić wpływ formy zastosowanego narzędzia na wyniki uzyskiwane przez osoby badane, obliczono miarę delta opartą na bezwzględnej wartości różnicy między wynikiem danej osoby w danej grupie a wartością średnią w grupie. Następnie zastosowano jednoczynnikową analizę wariancji, aby przekonać się, czy zróżnicowanie wyników jest w jakiś sposób uzależnione od wersji narzędzia badawczego. W przypadku kwestionariusza PTS, ze względu na wymaganie jednowymiarowości zmiennej latentnej, do badania wybrano tylko pozycje kwestionariuszowe dotyczące ruchliwości procesów nerwowych. Wyniki średnie oraz istotność różnic między nimi przedstawia tabela 4.17.

Tabela 4.17. Uśrednione bezwzględne różnice między wynikami indywidualnymi i grupowymi

medium badania	forma narzędzia	N	Średnia bezwzględna różnica	Odchylenie standardowe	Błąd standardowy	95% przedział ufności dla średniej	
						Dolna granica	Górna granica
test Omnibus - wynik ogólny							
internet	pełna [a]	27	8,40	5,632	1,084	6,17	10,63
	skrócona [ab]	45	10,10	6,925	1,032	8,02	12,18
	3pl [ac]	40	5,42	4,853	0,767	3,86	6,97
kwestionariusz PTS - ruchliwość procesów nerwowych							
internet	pełna	10	2,77	2,51	1,651	-0,49	6,027
	skrócona	24	5,43	3,661	1,065	3,325	7,532
	GRM	70	7,14	5,687	0,624	5,91	8,373
papier	pełna	38	7,05	5,262	0,847	5,374	8,717
	skrócona	32	8,08	5,632	0,923	6,258	9,901

a, b, c – grupy jednorodne na podstawie testu post hoc T3 Dunnetta. Źródło: badania własne.

Najmniejsze zróżnicowanie zaobserwowano w grupie osób badanych za pomocą pełnej wersji kwestionariusza PTS za pośrednictwem internetu. Jednakże niska liczebność tej grupy (N=10) wpływa na mało dokładne oszacowanie tej zmienności, czego wyrazem jest wysoki SE. Dla czynnika RPH kwestionariusza PTS nie stwierdzono interakcji między formą narzędzia a medium badania (F(1, 169) = 0,487; p = 0,486). Istotne okazały się natomiast różnice między wynikami zebranymi za pomocą internetu i papieru (F(1, 169) = 8,841; p = 0,003) oraz między poszczególnymi

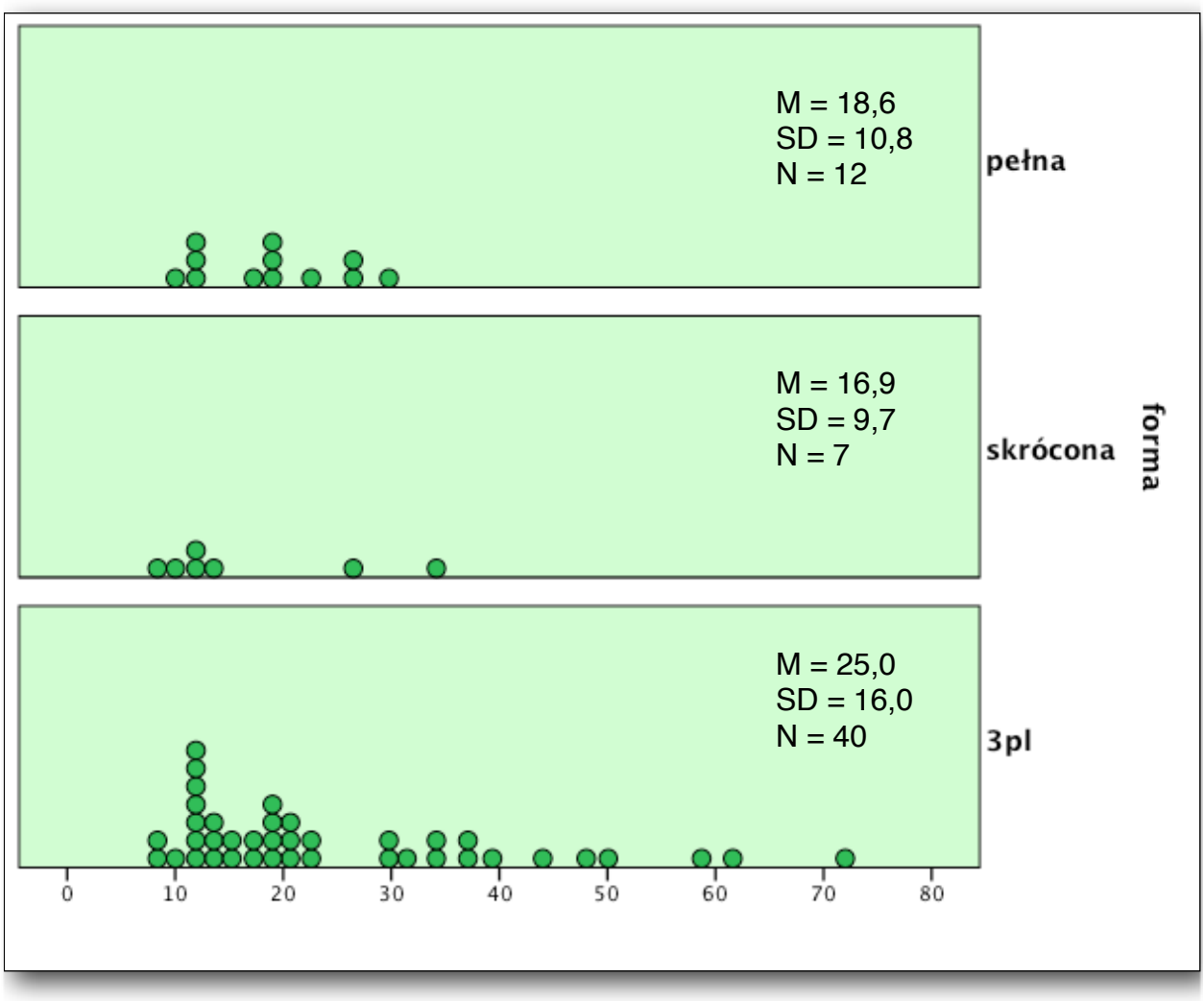
formami kwestionariusza (F(2, 169) = 3,786; p = 0,025). W tym drugim przypadku analiza wielkości efektu wykazała jednak, że są to różnice pozorne, związane z nieliczną, odstającą grupą wyników skali pełnej zebranych w internecie (d = -0,0983; $95\%CI$ = (-0,76;0,56)). Po jej wykluczeniu wariancja bezwzględnych odchyleń w poszczególnych grupach nie jest istotnie zróżnicowana (p = 0,179).

Dla testu Omnibus zróżnicowanie wyników okazało się istotnie różne (F(2, 109) = 6,65; p = 0,002; d = 0,784; $95\%CI$ = (0,34; 1,23)) – wyniki zebrane za pomocą testu adaptacyjnego miały niższą różnorodność i niższy błąd standardowy niż wyniki z wersji skróconej.

4.6.7. Zróżnicowanie czasu odpowiedzi w różnych typach testów i kwestionariuszy

Badanie za pośrednictwem komputerów pozwala mierzyć nie tylko same odpowiedzi, ale też na przykład czas, jaki jest potrzebny na ich udzielenie. Podczas przeprowadzonych badań poddano analizie także tą zmienną.

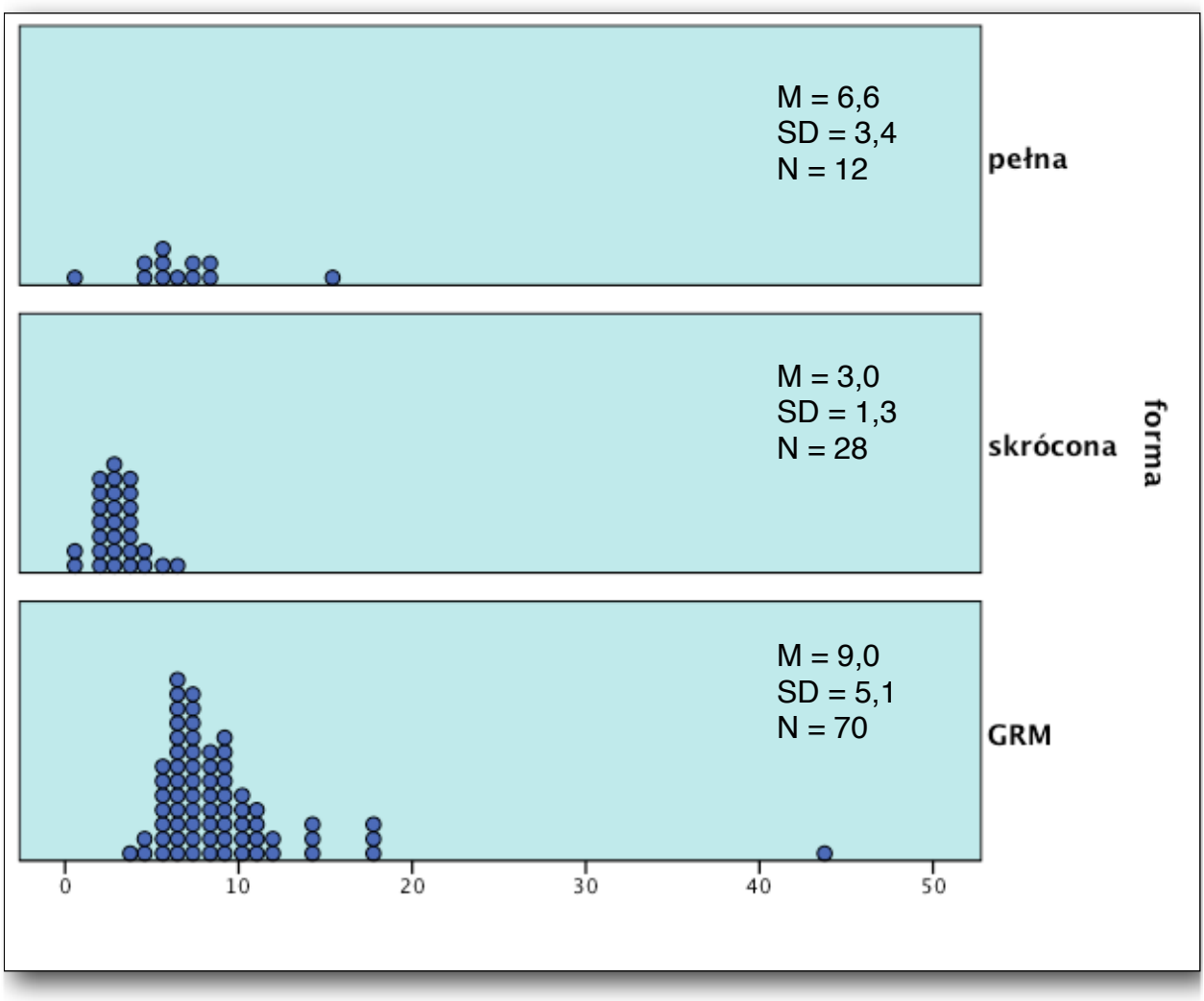
Średni czas potrzebny na rozwiązanie jednego zadania w teście Omnibus wynosił we wszystkich formach badania ok. 22 sekundy (± 14). Między zastosowanymi formami brak jest istotnych różnic ($X^2_{(2)}$ = 1,03; p = 0,598), lecz analiza rozkładu czasów pokazuje, że dla formy adaptacyjnej ma on kształt dużo bardziej prawoskośny, niż niż dla pozostałych (por. ryc. 4.5).



Ryc 4.5. Średni czas w sekundach potrzebny na udzielenie odpowiedzi w teście Omnibus.

Co ciekawe, osoby, które nie wypełniły testu do końca odpowiadały na pytania zdecydowanie szybciej ($p < 0,05$). Tutaj średni czas wynosi 14 sekund (± 4).

W przypadku kwestionariusza PTS czasy odpowiedzi ze względu na charakter pozycji były krótsze i wynosiły około 7,5 sekundy (± 5). Rozkłady czasów dla poszczególnych form kwestionariusza mają zbliżony kształt (por. ryc. 4.6), a średni czas formy skróconej jest krótszy niż pozostałych obu, które się między sobą nie różnią ($X^2_{(2)} = 60,5$; $p < 0,001$; na podstawie post hoc z poprawką Bonferroniego dla $p < 0,01$).



Ryc 4.6. Średni czas w sekundach odpowiadania na jedną pozycję kwestionariusza PTS.

Dla kwestionariusza PTS osoby, które go nie ukończyły, nie różniły się istotnie pod względem czasu odpowiadania na pozycje kwestionariuszowe od pozostałych badanych.

Zaobserwowane wyniki pozwalają sformułować wniosek, iż pomiar czasu może być dodatkowym wskaźnikiem jakości wyników i znając przeciętny czas potrzebny na wypełnienie testu można oznaczyć nierzetelne odpowiedzi osób uczestniczących w badaniach. W przypadku kwestionariuszy, gdzie udzielenie odpowiedzi zajmuje mniej czasu, różnica między rzetelną a prawidłową odpowiedzią jest trudniejsza do uchwycenia, choć przypuszczam, że w badaniach poświęconych temu obszarowi także udałoby się dowieść istotnych różnic. Wymagałoby to jednak podjęcia badań poświęconych wyłącznie temu zagadnieniu, bez wprowadzania dodatkowych czynników w postaci formy kwestionariusza oraz medium przeprowadzania badań.

Rozdział 5. Dyskusja

Symulacja łańcuchów Markowa Monte Carlo, którą wykorzystano w tej pracy została zaprojektowana w celu zbadania psychometrycznych skutków zastosowania modeli opartych na teorii odpowiadania na zadania testowe (IRT) do sporządzania skróconych wersji narzędzi badawczych w psychologii. Konieczność, a jednocześnie możliwość zastosowania podejścia symulacyjnego wynikała z założenia przeprowadzenia analiz na dużych, 1000-osobowych próbach. W literaturze przedmiotu można napotkać kilka prac, które poświęcone są zastosowaniu IRT w praktyce psychometrycznej, ale większość z nich dotyczy tylko modeli dychotomicznych, dwuparametrycznych (np. Kang i Waller, 2005). Analizy przeprowadzone w prezentowanej pracy rozszerzają wyniki na modele politomiczne, dla danych z kwestionariuszy z pozycjami typu Likerta, w celu zwiększenia możliwości uogólnień na narzędzia często stosowane w psychologii. Jednocześnie wyniki symulacji były wykorzystywane do zbadania właściwości teoretycznych na wynikach obserwowanych w trakcie badań empirycznych. Z tego powodu pierwszym celem tej pracy było rozszerzenie wiedzy na temat potencjalnie korzystnych zastosowań IRT poprzez wdrożenie modeli politomicznych do badań praktycznych.

Drugim celem pracy było zbadanie równoległości narzędzi badawczych wykorzystywanych w psychologii, gdy badania przeprowadzane są w ich „naturalnym" środowisku – za pomocą papieru i ołówka – oraz coraz częściej za pośrednictwem komputerów i internetu. Upowszechnianie się dostępu, rozwój i popularyzacja technologii oraz coraz większa obecność internetu w życiu codziennym, skłania do refleksji nad możliwościami i ograniczeniami wynikającymi ze stosowania elektronicznych wersji narzędzi pierwotnie opracowanych w tradycyjnych warunkach. Przytoczone wyniki badań (por. rozdział 2.4) oraz przeprowadzone analizy pozwalają sądzić, że w większości przypadków przeniesienie papierowej wersji psychologicznego narzędzia badawczego „do komputera" nie tylko nie zmienia jego właściwości psychometrycznych, lecz także wzbogaca jego możliwości pomiarowe, np. o różne parametry czasu związanego z udzielaniem odpowiedzi. Co więcej, wersje komputerowe dostarczają wyniki mniej obciążone błędem pomiarowym dzięki np. automatycznej walidacji, ograniczaniu zakresu odpowiedzi, czy też kontrolowaniu braków danych.

Trzecim celem pracy było uogólnienie wniosków wynikających z zastosowania IRT do promowania metod probabilistycznych jako użytecznego narzędzia

statystycznego. Mimo, że wielu badaczy zwraca uwagę na znaczenie teorii pomiaru dla jakości uzyskiwanych wyników (Phillips i Lord, 1986, Smith i Stanton, 1998, Scandura i Williams, 2000; Austin i in., 2002) i równie wielu podkreśla użyteczność IRT (Hulin i Ilgen, 1990; Zickar, 1998), to w oparciu o tę teorię jak dotąd powstało niewiele narzędzi praktycznych. W niniejszej pracy podjęto próbę dodania podejścia IRT do repertuaru technik statystycznych wykorzystywanych przez badaczy w psychologii.

5.1. Wnioski i znaczenie wyników

Używanie modeli opartych o teorię odpowiadania na pozycje testowe przy analizowaniu wyników dotyczących zmiennych latentnych ma dwie niewątpliwe zalety: 1) wyniki przypisywane osobom uczestniczącym w badaniach wyrażane są na mocnej skali interwałowej Θ (Embretson i DeBoeck, 1994; Harwell i Gatti, 2001) oraz 2) modele IRT pozwalają na dokładniejszą estymację błędu pomiarowego SEM (Fraley, Waller i Brennan, 2000; Mellenbergh, 1999; Reise i Haviland, 2005). Przeprowadzone badania wykazały, że poziom błędu jest mocniej związany z wynikami surowymi i przeliczonymi, niż z wynikami określonymi za pomocą poziomu Θ. Wynik ten wskazuje, że traktowanie wszystkich pozycji testowych bądź kwestionariuszowych równorzędnie – a tak się dzieje przy większości narzędzi badawczych w psychologii – ma wpływ na jakość wyników uzyskiwanych za pomocą skróconych wersji. Sumowanie odpowiedzi z poszczególnych pozycji bez uwzględniania ich wagi, chociażby przez odniesienie się do ich trudności i mocy różnicującej, powoduje niekontrolowany wzrost błędu pomiarowego. Wyniki uzyskane w tej pracy wspierają tezę, iż wyniki przeliczone są stałe, bez względu na liczbę wybranych pozycji, ale należy pamiętać, że im jest ich mniej, tym większy staje się błąd pomiarowy.

Badania dostarczyły także informacji na temat kształtu zależności między długością narzędzia badawczego a zróżnicowaniem wyników. Istnieje optymalna długość narzędzia wyznaczona przez funkcję krzywoliniową, gdzie dodawanie kolejnych pozycji zmniejsza rozrzut wyników. Nie jest to jednak zależność prosta i od punktu przegięcia funkcji dalsze wydłużanie narzędzia nie przynosi znaczącej poprawy jakości wyników.

Przeprowadzone w pracy analizy wykazały ponadto, że różne podejścia statystyczne prowadzą do uzyskania różnych skróconych zestawów pozycji testowych bądź kwestionariuszowych. Opierając się na korelacjach, ładunkach czynnikowych czy poziomie informacji zawartej w poszczególnych pozycjach otrzymano różne zestawy, które jednocześnie nie różniły się wynikami przeliczonymi (poza testem Omnibus, gdzie

wyniki skrócone były istotnie niższe). Mimo odrzucenia hipotezy o zgodności składu narzędzi, należy zwrócić uwagę, że porównanie rzetelności tych różnych wersji nie pozwala wybrać lepszej metody. Jednocześnie w oparciu o IRT wiadomo, że poziom trudności pozycji testowych wpływa na poziom błędu standardowego pomiaru. A z kolei poziom SEM wpływa pośrednio na długość narzędzia – te składające się z pozycji łatwych i o średnim poziomie trudności w mniejszej liczbie kroków dostarczają informacji o poziomie Θ osoby uczestniczącej w badaniu.

Znając parametry pozycji wchodzących w skład narzędzia, badacz przygotowujący jego skróconą wersję może podjąć decyzję ze świadomością konsekwencji swojego wyboru. W oparciu o techniki probabilistyczne może spodziewać się określonego poziomu błędu, tym większego im więcej pozycji trudnych zostanie przez niego wybranych. Potwierdzeniem tej tezy jest zaprezentowana wcześniej obserwacja, zgodnie z którą poziom szacowanej cechy dla różnych wersji narzędzia nie różni się, poza wersją składającą się z trudnych pozycji.

Rezultaty uzyskane w tej pracy pokrywają się z wynikami przedstawionymi przez innych badaczy. Zarówno Embretson (1996), jak i Kang oraz Waller (2005) stwierdzili niewielki wpływ długości testu na wielkość błędu, niezależnie od sposobu przeliczania wyniku. Co prawda w klasycznej teorii testu zwiększanie długości skali jest jedną z metod zwiększania rzetelności pomiaru, jednak - jak wspomniano wcześniej - badania przeprowadzone w oparciu o IRT sugerują, że takie podejście jest mało efektywne.

5.1.1. Implikacje psychometryczne

Badania w zakresie zastosowania modeli IRT w różnych obszarach psychometrii są jej ważnym nurtem. Na gruncie teorii testów niejednokrotnie podkreślano psychometryczne zalety modeli Rascha w odniesieniu do skal dychotomicznych (Embretson i Reise, 2000; Reise i Haviland, 2005), jednak znacznie mniej uwagi poświęcano modelom politomicznym. W tym kontekście niniejsza praca wypełnia pewną lukę w obszarze badawczym. W 1996 Susan E. Embretson potwierdziła teoretyczne przesłanki, że wyniki Θ uzyskane w dychotomicznym modelu IRT są odporne zarówno na błędy I, jak i II rodzaju. Jednocześnie, podczas czynnikowej analizy wariancji wykazała, że wyniki surowe takiej odporności nie wykazują. Kang i Waller (2005) rozszerzyli te wyniki dla modeli dwuparametrycznych. Symulacje przeprowadzone w ramach powyższego opracowania sugerują, że wyniki uzyskane w oparciu o modele GRM także są odporniejsze na błędy niż wyniki surowe. Badania te wpisują się w trend sprawdzający teoretyczne modele w warunkach i dla narzędzi coraz

bardziej odpowiadających realiom rzeczywistych badań psychologicznych. Z uwagi na fakt, że w Polsce jak dotąd modele politomiczne nie były na gruncie psychologii badane, przedstawiona na kartach tej pracy próba przyczyni się do ich szerszego stosowania.

5.1.2. Implikacje wyników dla praktyki psychologicznej

Mimo, że pomiar uznawany jest za kluczowy element badań w psychologii stosowanej, przez ostatnie 70 lat dominuje tylko jedno podejście psychometryczne oparte na klasycznej teorii testu. Związane jest ono z kilkoma ograniczeniami narzędzi badawczych: z tendencją do długich skal, gdzie interpretacja wyników zależy od próby normalizacyjnej; z przyjmowaniem założenia, że każda pozycja testowa lub kwestionariuszowa wnosi tyle samo informacji do wyniku końcowego; z trudnym tworzeniem równorzędnych wersji. Wszystkie te problemy rozwiązuje podejście oparte na IRT, które dostarcza wyniki na mocnej skali ilorazowej. Niestety, podejście probabilistyczne jest wciąż mało popularne, mimo że wielu badaczy opracowuje narzędzia na jego podstawie i używa ich w różnych dziedzinach psychologii: osobowości (Reise i Waller, 1990; Ferrando, 1994; Steinberg i Thissen, 1995; Gray-Little, Williams i Hancock, 1997; Rouse, Finger i Butcher, 1999), postaw (Fraley, Waller i Brennan, 2000), psychopatologii (Reise i Waller, 2003; Waller i Reise, 2009), psychologii klinicznej dzieci (Lanza, Foster, Taylor i Burns, 2005), psychologii kryminalnej (Osgood, McMorris i Potenza, 2002). Ten stan rzeczy można tłumaczyć brakem popularyzatorskich opracowań pokazujących zastosowanie IRT, deficytem programów statystycznych opartych na IRT i nieobecnością IRT w procesie edukacji psychologów i psychometrów. Być może poniższa praca przyczyni się do zmiany w tym obszarze, szczególnie poprzez udostępnienie prostego w obsłudze narzędzia do przeprowadzania badań w oparciu o IRT, dostępnego pod adresem badanet.amu.edu.pl (por. załącznik 4).

Użycie IRT poprawia jakość analiz parametrycznych i lepiej oddaje właściwości cech latentnych wpływających na wyniki obserwowane, niż klasyczna teoria testu uznająca wynik za składową części prawdziwej i błędu pomiarowego. Mimo skomplikowanego aparatu matematycznego, poprawne zastosowanie IRT sprowadza się do prostego przestrzegania trzech opisanych poniżej kroków.

Po pierwsze, należy zebrać odpowiednią liczbę obserwacji, aby określić parametry pozycji testowych lub kwestionariuszowych według wybranego modelu IRT. Analizy w tej pracy oraz najnowsze badania wskazują, że dla modeli politomicznych

wystarczająca jest liczebność próby w granicach 250-300 osób (Chuah, Drasgow i Luecht, 2006; Ostini i Nering, 2006).

Po drugie, należy określić wymiarowość cechy latentnej. Informacje te można często założyć w oparciu o przesłanki teoretyczne lub sprawdzić w analizie czynnikowej. Problem braku jednowymiarowości można rozwiązać przeprowadzając osobno analizy dla poszczególnych podskal (Hulin i Ilgen, 2000).

Po trzecie, należy dobrać odpowiedni model do danych. Obecnie istnieje wiele modeli zarówno dychotomicznych, jak i politomicznych jedno- i wielowymiarowych. Są też nawet odpowiednie modele dla nieparametrycznych zmiennych latentnych. Ze względu na charakter większości zmiennych w psychologii, model GRM Samejima wydaje się najlepszym wyborem.

Rezultaty tej pracy zdecydowanie wskazują, że wyniki szacowane w oparciu o model 3PL i GRM charakteryzują się lepszą ogólną dokładnością niż wyniki surowe. Biorąc to pod uwagę, nie waham się rekomendować badaczom w obszarze psychologii częstszego wykorzystywania modeli IRT, aby zwiększyć dokładność analiz parametrycznych. Mam nadzieję, że wyniki przedstawione w tej pracy stanowią wystarczającą zachętę dla bardziej powszechnego stosowania modeli probabilistycznych.

5.2. Ograniczenia i przyszłe obszary badań

Analizy zawarte w tej pracy dotyczyły wybranych dwóch narzędzi. Jeden test i jeden kwestionariusz to zbyt mało, aby stwierdzić, że odkryte właściwości dotyczyć będą większości narzędzi wykorzystywanych w badaniach psychologicznych. Tym niemniej, jest to obiecujący początek procesu włączania do listy narzędzi badawczych współczesnych psychologów także wersji komputerowych, a w szczególności adaptacyjnych. Podejmując próby stworzenia takich wersji należy jednak pamiętać, że w zaprezentowanych wyżej analizach uzyskano niejednoznaczne wyniki dla poszczególnych parametrów. W porównaniu do wersji pełnych, dla wersji skróconych uzyskano różne wyniki w obu narzędziach. Dla pomiaru temperamentu zróżnicowanie wyników skróconego kwestionariusza było mniejsze; z kolei dla testu inteligencji w wersji skróconej zróżnicowanie wyników było większe. Wciąż otwartym pozostaje pytanie, czy czynnikiem różnicującym jest obszar narzędzia, czy tylko różna forma pozycji testowych i kwestionariuszowych.

Metodologia budowania skróconych wersji narzędzi badawczych w oparciu o IRT jest bardziej skomplikowana, niż za pomocą analizy czynnikowej lub analizy regresji.

Ma jednak ogromną zaletę, ponieważ potrafi określić przydatność poszczególnych pozycji bez względu na ich poziom pomiarowy. Odpowiednie modele matematyczne opisują zarówno pozycje dychotomiczne jak i wielokategorialne, gdzie te pierwsze są problematyczne w analizach czynnikowych i regresyjnych. Tę zaletę można wykorzystać budując wersje skrócone wielu narzędzi badawczych w psychologii, tak jak ma to miejsce np. w psychiatrii (Streiner, 2010; Cooper i Petrides, 2010; Calamia i in., 2011; Khan, Lewis i Lindenmayer, 2011). Takie wersje narzędzi pozwalają ograniczyć do niezbędnego minimum czas potrzebny na przeprowadzanie badania, dostarczając jednocześnie parametrycznych informacji na temat poziomu badanych cech ukrytych.

Ponadto zastosowanie komputera w procesie zbierania wyników z testu lub kwestionariusza pozwala wprowadzić pomiar nowych parametrów, np. czasu odpowiadania na poszczególne pozycje. Dostarcza to informacji o przebiegu badania w czasie, które można wykorzystać do poprawy jakości danych, szczególnie w testach inteligencji, gdzie np. osoby nierzetelnie wypełniające test robią to istotnie szybciej.

Obszary badawcze wymagające głębszej eksploracji, dotyczą odpowiedzi na pytania o długość skróconych wersji narzędzi badawczych oraz wielkość prób kalibracyjnych.

Pierwszy przypadek dotyczy tworzenia wersji skróconych. W takim przypadku, gdy nie jest wykorzystywany algorytm adaptacyjny, a badacz chce tylko uzyskać krótką wersję danego narzędzia, aby skrócić czas potrzebny na badanie, otwarta pozostaje kwestia optymalnej długości narzędzia. Czy odkryta w zaprezentowanych wcześniej analizach funkcja odwrotna opisywać będzie także inne testy i/lub kwestionariusze, poza użytymi w tej pracy?

Drugie pytaniem wymagające odpowiedzi można sformułować następująco: Czy dla każdej mierzonej cechy optymalna wielkość próby kalibracyjnej będzie wynosiła 275 osób? W przypadku tej pracy, a także wspomnianych już opracowań innych badaczy (Chuah, Drasgow i Luecht, 2006; Ostini i Nering, 2006) szacunki opierają się na analizach symulacyjnych. Na ile stabilne są te wyniki? Czy zostaną potwierdzone w badaniach rzeczywistych? Na te pytania należałoby odpowiedzieć, zanim przystąpi się do stosowania IRT w budowaniu narzędzi w wersjach skróconych lub adaptacyjnych.

Mimo tych wątpliwości opracowane tutaj skrócone wersje stanowią obiecującą alternatywę dla wersji oryginalnych.

Słownik symboli

1pl – odmiana jednoparametryczna (pozycje różnią się tylko trudnością) modelu dwukategorialnego IRT

1Q – pierwszy kwartyl

2pl – odmiana dwuparametryczna (pozycje różnią się trudnością i mocą dyskryminacyjną) modelu dwukategorialnego IRT

3pl – odmiana trójparametryczna (różnią się trudnością, mocą dyskryminacyjną oraz poziomem zgadywalności) modelu dwukategorialnego IRT

3Q – trzeci kwartyl

4pl – odmiana czwórparametryczna (pozycje różnią się trudnością, mocą dyskryminacyjną, poziomem zgadywalności i niedbałości) modelu dwukategorialnego IRT

a – moc różnicująca danej pozycji

b – poziom trudności pozycji

B.D. – brak danych

c – parametr zgadywania

CAT– sposób badania polegający na tym, że osoby uczestniczące w badaniu otrzymują pytania ustalane w oparciu o dotychczasowe odpowiedzi i dobierane tak, aby zmaksymalizować ilość informacji i z jak najmniejszym błędem dokonać estymacji

wartości θ (*Computerized-Adaptive Test*)

CI – granice przedziału ufności (*Confidence Interval*)

d – parametr niedbałości

D – stały parametr maksymalizujący dopasowanie krzywej logistycznej do ogiwy

df – stopnie swobody (*degree of freedom*)

DIF – wskaźnik określający na ile osoby o tym samym poziomie θ uzyskają różne wyniki

w związku z pochodzeniem z różnych grup (*Differential Item Functioning*)

e – podstawa logarytmu naturalnego – stała matematyczna

F – wartość rozkładu F Snedecora

FA – analiza czynnikowa (*Factor Analysis*)

FIT – sposób przeprowadzania badania polegający na tym, że osoby uczestniczące w badaniu rozwiązują taki sam zestaw pytań (*Fixed-Item Test*)

GPCM – wielokategorialny uogólniony model punktów częściowych (*Generalised Partial Credit Model*)

GRM – model IRT klasy odpowiedzi (*Graded Response Model*)

I – informacja, w IRT odwrotność SEM

ICC – krzywe charakterystyczne dla pozycji (*Item Characteristic Curve*)

IRT – teorii odpowiadania na pozycje testu (*Item Response Theory*)

KTT – klasyczny model wyniku prawdziwego lub klasyczna teoria testów

kwestionariusz – narzędzia badawcze, w których odpowiedzi udzielane przez osoby uczestniczące w badaniu wskazują na poziom natężenia badanych cech

LR – stopień podobieństwa (*Likelihood Ratio*)

MCMC – klasa algorytmów próbkowania z rozkładu prawd

MR – wielokrotna analiza regresji (*Multivariate Linear Regression*)

OL – sposób prezentacji pozycji testowych z wykorzystaniem komputerów

Omnibus – nazwa testu inteligencji

p – poziom istotności

p(Θ) – prawdopodobieństwo posiadania umiejętności, wiedzy itp. na danym poziomie Θ

PCM – wielokategorialny jednoparametryczny model IRT

pozycje (testowe, kwestionariuszowe) – elementy w postaci zdań, pytań lub zadań, z których składają się testy lub kwestionariusze

PP – sposób prezentacji pozycji testowych z wykorzystaniem papieru i ołówka

PTS – nazwa kwestionariusza do mierzenia temperamentu (*Pavlovian Temperament Survey*)

q – odwrotność prawdopodobieństwa p – prawdopodobieństwo nie posiadania umiejętności itp.

RPN – nazwa czynnika w kwestionariuszu PTS - ruchliwość procesów nerwowych

RST – teoria losowego doboru próby (*Random Sampling Theory*)

SAT – sposób badania, w którym procedura dobierania pytań opiera się na wyborze przez osoby uczestniczącej w badaniu poziomu trudności następnej pozycji (*Self-Adapted Test*)

SD – odchylenie standardowe (*Standard Deviation*)

SEM – standardowy błąd pomiaru (*Standard Error Measurement*)

SMD – standaryzowana średnia różnica,

SPH – nazwa czynnika w kwestionariuszu PTS - siła procesów hamowania

SPP – nazwa czynnika w kwestionariuszu PTS - siła procesów pobudzenia

t – wartość rozkładu t-Studenta

test – narzędzie, za pomocą którego ocenia się poprawność lub jakość odpowiedzi w odniesieniu do pewnego standardu

T_S, T, S_{3ij}, Q_3, $\chi^2_{G/D}$ – statystyczne wskaźniki określające jednowymiarowość zmiennej

latentnej

WP – wynik przeliczony

WS – wynik surowy

Z – standaryzowana wartość statystyki testu różnic

α – wielkość błędu pierwszego rodzaju

β – wielkość błędu drugiego rodzaju

Θ – zmienna latentna – właściwość, cecha, która nie jest dostępna bezpośredniemu pomiarowi; poprzez wartość cechy latentnej rozumie się tu zarówno natężenie cechy badanej, jak i poziom umiejętności

Bibliografia

Aguinis, H., Beaty, J. C., Boik, R. J., Pierce, C. A. (2005). Effect size and power in assessing moderating effects of categorical variables using multiple regression: A 30-year review. *Journal of Applied Psychology, 90*, 94–107.

Akdemir, O., Oguz, A. (2008). Computer-based testing: An alternative for the assessment of Turkish undergraduate students. *Computers & Education 51*, 1198–1204.

Alexander, M. W., Truell, A. D., Bartlett, J. E., Ouwenga, K. (2001). Testing in a computer technology course: An investigation of equaivalency in performance between online and paper and pencil methods. *Journal of Career and Technical Education, 18*. Pobrano 20.07.2011 ze strony: scholar.lib.vt.edu/ejournals/JCTE/v18n1/alexander.html.

Alfonseca, E., Rodriguez, P., Perez, D. (2007). An approach for automatic generation of adaptive hypermedia in education with multilingual knowledge discovery techniques. *Computers & Education, 49,* 495–513.

Alkhadher, O., Clarke, D. C., Anderson, N. (1998). Equivalence and predictive validity of paper-and-pencil and computerized adaptive formats of the Differential Aptitude Tests. *Journal of Occupational and Organisational Psychology*, *71*, 205–217.

Aluja, A., Rossier, J., Zuckerman, M. (2007). Equivalence of paper and pencil vs Internet forms of the ZKPQ-50-CC in Spanish and French samples. *Personality and Individual Differences*, *43*, 2022–2032.

Anastasi, A., Urbina S. (1997). Testy psychologiczne. Warszawa: Pracownia Testów Polskiego Towarzystwa Psychologicznego.

Andrich, D. (1982). An index of person separation in latent trait theory, the traditional KR-20 index, and the Guttman scale response pattern. *Education Research and Perspectives, 9*, 95–104.

Austin, J. T., Scherbaum, C. A., & Mahlman, R. A. (2002). History of research methods in industrial and organizational psychology: Measurement, design, analysis. W: S. G. Rogelberg (red.), *Handbook of research methods in industrial and organizational psychology* (s. 3-33). Malden, MA, Blackwell Publishing.

Baker, F. B. (1987). Methodology review: Item parameter estimation under the one-, two- and three-parameter logistic models. *Applied Psychological Measurement, 11,* 111–142.

Baker, F. B. (2001). *The basic of item response theory.* Portsmouth: Hainemann.

Bartram, D., Brown, A. (2004). Online testing: Mode of administration and the stability of OPQ 32i scores. *International Journal of Selection and Assessment*, *12*, 278–284.

Batorski, D. (2005). Społeczne aspekty korzystania znowych technologii. W: J. Czapiński i T. Panek (red.), *Diagnoza społeczna 2005: Warunki i jakość życia Polaków* (s. 27–28, 214–231). Warszawa: Wyższa Szkoła Finansów i Zarządzania.

Batorski, D. (2006). Cyfrowy podział w Polsce: nowe technologie a szanse życiowe i wykluczenie społeczne. W: D. Batroski, M. Marody, A. Nowak (red.), *Społeczna przestrzeń internetu* (s. 317–336). Warszawa: Wydawnictwo Academica SWPS.

Batorski, D., Olcoń-Kubicka, M. (2006). Prowadzenie badań przez internet – podstawowe zagadnienia metodologiczne. *Studia Socjologiczne, 182*, 99–132.

Baumer, M., Roded, K., & Gafni, N. (2009). *Assessing the equivalence of Internet-based vs. paper-and-pencil psychometric tests*. Zaprezentowano na CAT Research and Applications Around the World Poster Session, June 2, 2009. Pobrano 22.06.2011 ze strony: www.psych.umn.edu/psylabs/catcentral/pdf%20files/cat09roded.pdf.

Bayles, M. (2000). *Just how 'blind' are we to advertising banners on the web?* Pobrano 12.09.2009 ze strony: www.surl.org/usabilitynews/22/pdf/Usability%20News%2022%20-%20Bayles.pdf.

Beckers, J. J., Schmidt, H. G. (2001). The structure of computer anxiety: a six-factor model. *Computers in Human Behavior, 17*, 35–49.

Benway, J. P., Lane, D. M. (1998). *Banner blindness: web searchers often miss „obvious" links.* Pobrano 12.06.2009 ze strony: www.internettg.org/newsletter/dec98/banner_blindness.html.

Bernt, F. M., Bugbee, A. C., Arceo, R. D. (1990). Factors influencing student resistance to computer administered testing. *Journal of Research on Computing in Education, 22* (3), 265–275.

Bickart, B., Schmittlein, D. (1999). The distribution of survey contact and participation in the United States: Constructing a survey-based estimate. *Journal of Marketing Research, 36*, 286–294.

Binet, A., Simon, Th. A. (1905). Méthode nouvelle pour le diagnostic du niveau intellectuel des anormaux. *L'Année Psychologique, 11*, 191–244.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. W: F. M. Lord, M. R. Novick (red.), *Statistical theories of mental test scores* (s. 397–472). Reading, MA: Addison-Wesley.

Bishop, F. L., Lewis, G., Harris, S., McKay, N., Prentice, Ph., Thiel, H. Lewith, G. T. (2010) A within-subjects trial to test the equivalence of online and paper outcome measures: The Roland Morris Disability Questionnaire. *BMC Musculoskeletal Disorders, 11*, 113–117.

Bock, R.D., Moustaki, I. (2007). Item response theory in a general framework. W: C.R. Rao, S. Sinhary (red.), *Handbook of statistics on psychometrics* (s. 472–490). Amsterdam: Elsevier.

Bolt, L. (2010). Comparision of a paper-and-pencil administered and an Internet administered health questionnaire among Dutch adults. Master Graduation Research Project. Pobrano 13.07.2011 ze strony: www.ggdkennisnet.nl/kennisnet/atoom.asp?atoom=56045&atoomsrt=2&actie=2.

Boone, D. E. (1991). Item-reduction vs subset-reduction short forms on the WAIS-R witj psychiatic inpatients. *Journal of Clinical Psychology, 47*(2), 271–276.

Booth-Kewley, S., Larson, G. E., Miyoshi, D. K. (2007). Social desirability effects on computerized and paper-and-pencil questionnaires. *Computers in Human Behavior 23*, 463–477.

Bosnjak, M., Tuten T. L. (2001). Classifying response behaviors in web-based surveys. *Journal of Computer-Mediated Communication, 6*. Pobrano 27.08.2008 ze strony: jcmc.indiana.edu/vol6/issue3/boznjak.html.

Broos, A. (2005). Gender and Information and Communication Technologies (ICT) Anxiety: Male Self-Assurance and Female Hesitation. *CyberPsychology & Behavior, 8*, 21–33.

Brosnan, M. J. (1999). Modeling tehnophobia: a case word procesing. *Computers in Human Behavior, 15*, 105–121.

Brzezińska, A. I., Brzeziński, J. M. (2004). Skale szacunkowe w badaniach diagnostycznych. W: J. M. Brzeziński (red.), *Metodologia badań psychologicznych. Wybór tekstów* (s. 233–306). Warszawa: Wydawnictwo Naukowe PWN.

Brzeziński, J. M. (2000). Metodologia badań naukowych i diagnostycznych. Podstawowe metody badawcze – teoria i praktyka testowania. W: J. Strelau (red.), *Psychologia. Podręcznik akademicki* (s. 389–434). Gdańsk: GWP.

Butcher, J. N., Perry, J,. Hahn, J. (2004). Computers in clinical assessment: historical developments, present status, and future challenges. *Journal of Clinical Psychology, 60*, 331–345.

Calamia, M., Markon, K., Denburg, N. L., Tranel, D. (2011). Developing a short form of Benton's Judgement of Line Orientation Test: An Item Response Theory approach. *The Clinical Neuropsychologist, 25*(4), 670–684.

Campbell, N. J., Dobson, J. E. (1987). An inventory of student computer anxiety. *Elementary School Guidance and Counseling, 22,* 149–156.

Carlbring, P., Brunt, S., Bohman, S., Austin, D., Richards, J., Ostd, L-G., Andersson, G. (2007). Internet vs. paper and pencil administration of questionnaires commonly used in panic/agoraphobia research. *Computers in Human Behavior 23*, 1421–1434.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*, 629–637.

Cattell, R. B. (1978). *The scientific use of factor analysis.* New York: Plenum.

Ceranoglu, T. A. (2010). Video games in psychoterapy. *Review of General Psychology, 14*, 141–146.

Choi, I-Ch., Kim, K. S., Boo, J. (2003). Comparability of a paper-based language test and a computer-based language test. *Language Testing*, 20, 295–320.

Chou, C. (2003). Incidences and correlates of Internet anxiety among high school teachers in Taiwan. *Computers in Human Behavior, 19,* 731–749.

Choynowski, M. (1971). Podstawy i zastosowania teorii rzetelności testów psychologicznych. W: J. Kozielecki (red.), *Problemy psychologii matematycznej* (s. 65–118). Warszawa: PWN.

Chua, S. L. (1997). *A review on studies of computer anxiety in the 1990s*. Pobrano 23.04.2009 ze strony: www.aare.edu.au/97pap/chuas535.htm.

Chua, S. L., Chen, D.-T., Wong, A. F. L. (1999). Computer anxiety and its correlates: a meta-analysis. *Computers in Human Behavior, 15*, 609–623.

Chuah, S. Ch., Drasgow, F., Luecht, R. (2006). How big is big enough? Sample size requiments for CAST item parametr estimation. *Applied Measurement in Education, 19*(3), 241–255.

Clara, I. P., Huynh, C.-L. (2003). Four short-form linear equation estimates of Wechsler Adult Intelligence Scale III IQs in an elderly sample. *Measurement and Evaluation in Counseling and Development, 35*, 251–262.

Claycomb, C., Porter, S. S., Martin, C. L. (2000). Riding the wave: Response rates and the effects of time intervals between successive mail survey follow-up efforts. *Journal of Business Research, 48*, 157–162.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46.

Cole, M. S., Bedeian, A. G., Feild, H. S. (2006). The measurement equivalence of web-based and paper-and-pencil measures of transformational leadership: A multinational test. *Organizational Research Methods, 9*, 339–368.

Comrey, A. L., Lee, H. B. (1992). *A first course in factor analysis.* Hillsdale, NJ: Erlbaum.

Cook, A. J., Roberts, D. A., Henderson, M. D., Van Winkle, L. C., Chastain, D. C., Hamill-Ruth, R. J. (2004). Electronic pain questionnaires: A randomized, crossover comparison with paper questionnaires for chronic pain assessment, *Pain, 110*, 310–317.

Cooper, A., Petrides, K. V. (2010). A psychometric analysis of the Trait Emotional Intelligence Questionnaire–Short Form (TEIQue–SF) using Item Response Theory. *Journal of Personality Assessment, 92*(5), 449–457.

Cooper, J. (2006). The digital divide: special case of gender. *Journal of Computer Assited Learning, 22*, 320–334.

Coste, J., Guillemin, F., Pouchot, J., Fermanian, J. (1997). Methodological approaches to shortening composite measurement scales. *Journal of Clinical Epidemiology, 50*, 247–252.

Couper, M., Conrad, F. G., Tourangeau, R. (2007). Visual context effects in web surveys. *Public Opinion Quarterly, 71*, 623–634.

Couper, M., Rowe, B. (1996). Evaluation of a computer-assisted self-interview component in a computer-assisted personal interview survey. *Public Opinion Quarterly, 60*, 89–105.

Crawford, J. R., Allan, K. M., Jack, A. M. (1992). Short-forms of the UK WAIS-R: Regression equations and their predictive validity in a general population sample. *British Journal of Clinical Psychology, 31*, 191–202.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334.

Czapiński, J., Panek, T., (red.) (2011). Diagnoza społeczna 2011. Warunki i jakość życia Polaków. *Contemporary Economics, 5*(3), 1–461.

De Ayala, R. J., Dodd, B. G., Koch, W. R. (1992). A comparision of the Partial Credit and Graded Response Models in computerized adaptive testing. *Applied Measurement in Education, 5*, 17–34.

DeMars, Ch. (2010). *Items Response Theory.* New York: Oxford University Press.

Denscombe, M. (2006). Web-based questionnaires and the mode effect: an evaluation based on completion rates and data contents of near-identical questionnaires delivered in different modes. *Social Science Computer Review, 24*, 246–254.

Denscombe, M. (2008). The length of responses to open-ended questions: a comparison of online and paper questionnaires in terms of a mode effect. *Social Science Computer Review, 26*, 359–368.

DeRouvray, C., Couper, M. P. (2002). Designing a strategy for reducing „no opinion" responses in web-based surveys. *Social Science Computer Review, 20*, 3–9.

Dillman, D. A. (2000). *Mail and Internet surveys: The total design method.* New York: John Wiley.

Dillman, D. A., Sinclair, M. D., Clark, J. R. (1993). Effects of questionnaire length, respondent-friendly design, and a difficult question on response rates for occupant-addressed census mail surveys. *Public Opinion Quarterly, 57*, 289–304.

DiMaggio, P., Hargittai, E. (2001). *From the 'digital divide' to 'digital inequality': Studying Internet use as penetration increases.* Pobrano 21.01.2009 ze strony: www.webuse.umd.edu/webshop/resources/Dimaggio_Digital_Divide.pdf.

Do-Hong, K., Huynh, H. (2010). Equivalence of paper-and-pencil and online administration modes of the statewide english test for students with and without disabilities. *Educational Assessment, 15*, 107–121.

Drasgow, F., Levine, M. V., McLaughlin, M. E. (1991). Appropriateness measurement for some multidimensional test batteries. *Applied Psychological Measurement, 15*, 171–191.

Egberink, I. J. L., Veldkamp, B. P. (2007). The development of a computerized adaptive test for integrity. W: D. J. Weiss (red.), *Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing*. Pobrano 29.06.2012 ze strony: www.psych.umn.edu/psylabs/CATCentral/.

Embretson, S. E. (1996). Item response theory models and spurious interaction effects in factorial ANOVA designs. *Applied Psychological Measurement, 20*(3), 201–212.

Embretson, S. E., DeBoeck, P. (1994). Latent trait theory. W: R. J. Sternberg (red.), *Encyclopedia of Intelligence* (s. 644–647). New York: MacMillan.

Embretson, S. E., Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ, London: Lawrence Earlbaum Associates.

Emons, W. H. M., Sijtsma, K., Meijer, R. R. (2005). Global, local and graphical person-fit analysis using person response functions. *Psychological Methods, 10*, 101–119.

Epstein, J., Klinkenberg, W. D., Wiley, D., McKinley, L. (2001). Insuring sample equivalence across internet and paper-and-pencil assessments. *Computers in Human Behavior 17*, 339–346.

Ferguson, G. A., Takane, Y. (1997). *Analiza statystyczna w psychologii i pedagogice*. Warszawa: Wydawnictwo Naukowe PWN.

Ferrando, P. J. (1994). Fitting item response models to the EPI-A impulsivity subscale. *Educational and Psychological Measurement, 54*, 118–127.

Finger, M. S., Ones, D. S. (1999). Psychometric equivalence of the computer and booklet forms of the MMPI: a meta-analysis. *Psychological Assessment, 11*, 58–66.

Ford, B. D., Vitelli, R., Stuckless, N. (1996). The effects of computer versus paper-and-pencil administration on measures of anger and revenge with an inmate population. *Computers in Human Behavior, 12*, 159–166.

Fraley, R. C., Waller, N. G., Brennan, K. A. (2000). An item response theory analysis of self-report measures of adult attachment. *Journal of Personality and Social Psychology, 78*, 350 – 365.

Fritts, B. E., Marszalek, J. M. (2010). Computerized adaptive testing, anxiety levels, and gender differences. *Social Psychology of Education, 13*, 441–458. DOI 10.1007/s11218-010-9113-3.

Gao, F., Chen, L. (2005). Bayesian or non-Bayesian: A comparison study of item parameter estimation in the three-parameter logistic model. *Applied Measurement in Education, 18*, 351–380.

Gati, I., Saka, N. (2001). Internet-based versus paper-and-pencil assessment: Measuring carreer decision-making difficulties. *Journal of Career Assessment, 9*, 397–416.

Gerardi, M., Cukor, J., Difede, J., Rizzo, A., Rothbaum, B. O. (2010). Virtual reality exposure therapy for post-traumatic stress disorder and other anxiety disorders. *Current Psychiatry Reports, 12*, 298–305.

Gessaroli, M. E., De Champlain, A. F. (1996). Using an approximate chi-square statistic to test the number of dimensions underlying the responses to a set of items. *Journal of Educational Measurement, 33*, 157–179.

Glas, C. A. W., Falcon, J. C. S. S. (2003). A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement, 27*, 87–106.

Goldberg, L. R. (1972). Parameters of personality inventory construction and utilization: A comparison of prediction strategies and tactics. *Multivariate Behavioral Research Monographs, 72*, 1–59.

Gray-Little, B., Williams, V. S. L., Hancock, T. D. (1997). An item response theory analysis of the Rosenberg Self-Esteem Scale. *Personality and Social Psychology Bulletin, 23*, 443–451.

Grobler, A. (2006). *Metodologia nauk*. Kraków: Wydawnictwo Aureus; Wydawnictwo Znak.

Groves, R. M., Cialdini, R. B., Couper, M. P. (1992). Understanding the decision to participate in a survey. *Public Opinion Quarterly, 56*, 475–495.

Gruijter de, D. N. M., Kamp van der, L. J. Th. (2008). *Theory for the Behavioral Science*. Boca Raton, London, New York: Chapman & Hall/CRC.

Guilford, J . P. (1954). *Psychometric methods*. New York: McGraw-Hill.

Gulliksen, H. (1950). *Theory of mental* test*s*. New York: Wiley.

Hambleton, R. K., Swaminathan, H., Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.

Hambleton, R. K., Swaminathan, H., Rogers, H. J. (2007). Assessing the Fit of Item Response Theory Models. W: C. R. Rao, S. Sinhary, (red.). *Handbook of statistics on psychometrics* (s 683–718). Amsterdam: Elsevier.

Handwerk, P. G., Carson, C., Blackwell, K. M. (2000). *On-line vs. paper-and-pencil Surveying of Students: A Case Study. AIR 2000 Annual Forum Paper.* Pobrano 20.06.2011 ze strony: www.eric.ed.gov/PDFS/ED446512.pdf.

Hargittai, E. (2003). *The digital divide and what to do about it.* Pobrano 18.01.2009 ze strony: www.eszter.com/papers/c04-digitaldivide.html.

Harris, J. B., Neal, G. (1996). Correlates among teachers' anxieties, demographics, and telecomputing activity. *Journal of Research on Computing in Education, 28*, 300–318.

Harwell, M., Gatti, G. G. (2001). Rescaling ordinal data to interval data in educational research. *Review of Educational Research, 71*, 105–131.

Harwell, M., Stone, C. A., Hsu, T., Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement, 20*, 101–125.

Hays, W. L. (1973). *Statistics for the social sciences*. Holt, Rinehart and Winston, Inc.: New York.

Hedman, E., Ljótsson, B., Rück, Ch., Furmark, T., Carlbring, P., Lindefors, N., Andersson, G. (2010). Internet administration of self-report measures commonly used in research on social anxiety disorder: A psychometric evaluation. *Computers in Human Behavior, 26*, 736–740.

Heerwegh, D., Vanhove, T., Matthijs K., Loosveldt, G. (2005). The effect of personalization on response rates and data quality in web surveys. *International Journal Social Research Methodology, 8*, 85–99.

Heinssen, R. K., Glass, C. R., Knight, L. A. (1987). Assessing computer anxiety: Development and validation of the computer anxiety rating scale. *Computers in Human Behavior, 3*, 49–59.

Herrero, J., Meneses, J. (2006). Short Web-based versions of the perceived stress (PSS) and Center for Epidemiological Studies-Depression (CESD) Scales: A comparison to pencil and paper responses among Internet users. *Computers in Human Behavior 22*, 830–846.

Holländare, F., Andersson, G., Engström, I. (2010). A comparison of psychometric properties between internet and paper versions of two depression instruments (BDI-II and MADRS-S) administered to clinic patients. *Journal of Medical Internet Research,12,* DOI:10.2196/jmir.1392.

Hornowska, E. (2001). *Testy psychologiczne. Teoria i praktyka.* Warszawa: Wydawnictwo Naukowe Scholar.

Hornowska, E. (2007). Stare wino w nowych bukłakach – czyli od Bineta do testowania adaptacyjnego. W: J. M. Brzeziński (red.), *Psychologia. Między teorią, metodą i praktyką* (s. 257–269). Poznań: Wydawnictwo Naukowe UAM.

Huang, H-M. (2006). Do print and Web surveys provide the same results? *Computers in Human Behavior, 22*, 334–350

Hulin, C. L., Ilgen D. R. (2000). Introduction to computational modeling in organizations: The good that modeling does. W: C. L. Hulin, D. R. Ilgen (red.), *Computational Modeling of Behavior in Organizations* (s. 3 – 18). Washington, D.C.: American Psychological Association.

Hulin, C. L., Lissak, R. I., Drasgow, F. (1982). Recovery of two and three parameter logistic item characteristic curves: A Monte Carlo study. *Applied Psychological Measurement, 6*, 249–260.

Hulin, Ch. L., Drasgow, F., Parsons, Ch. K., (1983). *Item response theory. Application to psychological measurement.* Homewood: Dow Jones-Irwin.

Hull, C. L. (1943). *Principles of behavior*. New York: Appleton.

Jaworowska, A., Matczak, A., (2002). *Omnibus. Test inteligencji. Podręcznik.* Warszawa: Pracownia Testów Psychologicznych Polskiego Towarzystwa Psychologicznego.

Joubert, T., Kriek, H. J. (2009). Psychometric comparison of paper and-pencil and online personality assessments in a selection setting. *SA Journal of Industrial Psychology/ SA Tydskrif vir Bedryfsielkunde, 35, 1.* DOI: 10.4102/sajip.v35i1.727

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20*, 141–151.

Kang, S., Waller, G. (2005). Moderated multiple regression, spurious interaction effects, and IRT. *Applied Psychological Measurement, 29*, 87–105.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person fit statistics. *Applied Measurement in Education, 16*, 277–298.

Katz, J. E., Rice, R. E., Aspden, P. (2001). The Internet, 1995–2000: Access, civic involvement, and social interaction. *American Behavioral Scientist, 45*, 404–419.

Kaufman, A. S. (1972). A short form of the Wechsler Preschool and Primary Scale of Intelligence. *Journal of Consulting and Clinical Psychology, 39*, 361–369.

Kerlinger, E. N., Pedhazur, E. J. (1973). *Multiple regression in behavioral research*. New York: Holt, Rinehart & Winston.

Khan, A., Lewis, Ch., Lindenmayer, J.-P. (2011). Use of non-parametric Item Response Theory to develop a shortened version of the Positive and Negative Syndrome Scale (PANSS). *BMC Psychiatry, 178*(11). Pobrano 20.05.2012 ze strony: biomedcental.com/1471-244X/11/178.

Kim, D-H., Huynh, H. (2008). Computer-based and paper-and-pencil administration mode effects on a statewide end-of-course English test. *Educational and Psychological Measurement, 68*, 554–570.

King Jr., W. C., Miles, E. W. (1995). A quasi-experimental assessment of the effect of computerizing noncognitive paper-and-pencil measurements: A test of measurement equivalence. *Journal of Applied Psychology, 80*, 643–651.

Kline, P. (1979). *Psychometrics and psychology.* London: Acaderric Press.

Kobrin, J. L., Young, J. W. (2003). The cognitive equivalence of reading comprehension test items via computerized and paper-and-pencil administration. *Applied Measurement in Education, 16*, 115–140.

Konarski, R. (2004). Model cechy latentnej w analizie psychometrycznej testów i pozycji testowych. W: B. Niemierko, H. Szaleniec (red.), *Diagnostyka edukacyjna. Standardy wymagań i normy testowe w diagnostyce edukacyjnej*. Kraków: PTDE. Pobrano 26.07.12 ze strony: pbs.pl/e4u.php/1,ModFiles/Download/files/Artykuly/Roman_Konarski_Model_cechy_latentnej_Krakow_2004.pdf.

Krippendorff, K. (2004). Reliability in Content Analysis: Some common Misconceptions and Recommendations. *Human Communication Research 30*,3, 411–433.

Krosnick, J. A., Holbrook, A. L., Berent, M. K., Carson, R. T., Hanemann, W. M., Kopp, R. J., Mitchel, R. C., Presser, S., Ruud, P. A., Smith, V. K., Moody, W. R., Green, M. C., Conaway, M. C. (2002). The impact of „no opinion" response options on data quality. Non-attitude reduction or an invitation to satisfice? *Public Opinion Quarterly, 66*, 371–403.

Kveton, P., Jelinek, M., Voboril, D., Klimusova, H. (2007). Computer-based tests: the impact of test design and problem of equivalency. *Computers in Human Behavior, 23*, 32–51.

Lanza, S. T., Foster, M., Taylor, T. K., Burns, L. (2005). *Assessing the impact of measurement specificity in a behavior problems checklist: An IRT analysis. Technical Report 05-75*. Pobrano 20.05.2012 ze strony: methodology.psu.edu/media/bibliography/techreports/197934924805-75.pdf.

Lawley, D. N. (1943). On problems connected with item selection and test construction. *Proceedings of the Royal Society of Edinburgh, 61*, 273–287.

Lee, H. K. (2004). A comparative study of ESL writers' performance in a paper-based and a computer-delivered writing test. *Assessing Writing, 9*, 4–26.

Legg, S. M., Buhr, D. C. (1992). Computerized adaptive testing with different groups. *Educational Measurement: Issues and Practice, 11* (2), 23–7.

Leung, D. Y. P., Kember, D. (2005). Comparability of data gathered from evaluation questionnaires on paper through the internet. *Research in Higher Education, 46*, 5, DOI: 10.1007/s11162-005-3365-3.

Lewin, K. (1936). *Principles of typological psychology*. New York: McGraw-Hill.

Lewis, I. M., Watson, B. C., White, K. M. (2009). Internet versus paper-and-pencil survey methods in psychological experiments: Equivalence testing of participant responses to health-related messages. *Australian Journal of Psychology, 61*, 107–116.

Light, P., Littleton, K., Bale, S., Joiner, R., Messer, D. (2000). Gender and social comparison effects in computer-based problem solving. *Learning and Instruction, 10*, 483–496.

Linacre J. M., Wright B. D. (1994). Chi-Square Fit Statistics. *Rasch Measurement Transactions, 8*, 350.

Lonsdale, Ch., Hodge, K., Rose, E. A. (2006). Pixels vs. paper: Comparing online and traditional survey methods in sport psychology. *Journal of Sport and Exercise Psychology, 28*, 100–108.

Lord, F. M. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement, 13*, 517–548.

Lord, F. M., Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Lord, F. M., Novick, M. R. (1980). *Applications of Item Response Theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Loyd, B. H., Gressard, C. (1984). Reliability and factorial validity of computer attitude scales. *Educational and Psychological Measurement, 44*, 501–505.

Mair, P., Reise, S. P., Bentler, P. M. (2008). *IRT Goodness-of-Fit approaches from logistic regression*. UC Los Angeles: Department of Statistics, UCLA. Pobrano ze strony: www.escholarschip.org/uc/item/1m46j62q.

Maloney, M. P., Ward, M. P. (1976). Psychlogical tests as a method of data collection. W: M. P. Maloney, M. P. Ward, *Psychological assessment: A conceptual approach* (s. 51–75). New York: Oxford University Press.

Mangunkusumo, R. T., Moorman, P. W., van den Berg-de Ruiter, A. E., Van Der Lei, J., De Koning, H. J., Raat, H. (2005). Internet-administered adolescent health questionnaires compared with a paper version in a randomized study. *Journal of Adolescent Health, 36*, 701–706.

Manovich, L. (2006). *Język nowych mediów.* Warszawa: Wydawnictwo Akademickie i Profesjonalne.

Marcoulides, G. A. (1989). Measuring computer anxiety: The computer anxiety scale. *Educational and Psychological Measurement, 49*, 733–739.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149–174.

Marcoulides, G. A., Stocker, Y.-O. Marcoulides, L. D. (2004). Examining the psychological impact of computers technology: an updated cross-cultural study. *Educational and Psychological Measurement, 64*, 312.

Mead, A. D., Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114*, 449–458.

Meade, A. W., Lawrence, M. C., Lautenschlanger, G. J. (2007). Are internet and paper-and-pencil personality tests truly comparable? An experimental design measurement invariance study. *Organizational Research Methods 10,* 322. DOI: 10.1177/1094428106289393.

Mellenbergh, G. J. (1999). A note on simple gain score precision. *Applied Psychological Measurement, 23*, 87–89.

Meyerson, P., Tryon, W. W. (2003). Validating Internet research: A test of the psychometric equivalence of Internet and in-person samples. *Behavior Research Methods, Instruments, & Computers, 35*, 4, 614–620.

Miller, E. T., Neal, D. J., Roberts, L. J., Baer, J. S., Cressler, S. O., Metrik, J., Marlatt, G. A. (2002). Test–retest reliability of alcohol measures: Is there a difference between Internet-based assessment and traditional methods? *Psychology of Addictive Behaviors, 16*, 56–63.

Mogel, S., Satz, P. (1962). Abbreviation of the WAIS for clinical use: An attempt at validation. *Journal of Clinical Psychology, 18*, 77–79.

Monahan, J. S., Harke, M. A., Shelley, J. R. (2008). Computerizing the Mental Rotations Test: Are gender differences maintained? *Behavior Research Methods, 40*, 422–427.

National Telecommunication and Information Administration (1995). *Falling through the Net: a survey of the „Have Not" in rural and urban America*. Pobrano 12.01.2009 ze strony: www.ntia.doc.gov/ntiahome/fallingthru.html.

National Telecommunication and Information Administration (1998). *Falling through the Net II: new data of the digital divide.* Pobrano 12.01.2009 ze strony: www.ntia.doc.gov/ntiahome/net2/falling.html.

National Telecommunication and Information Administration (1999). *Falling through the Net: defining the digital divide.* Pobrany 12.01.2009 ze strony: www.ntia.doc.gov/ntiahome/digitaldivide.html.

Naus, M. J., Philipp, M. L., Samsi, M. (2009). From paper to pixels: A comparison of paper and computer formats in psychological assessment. *Computers in Human Behavior, 25*, 1–7.

Nicholson J., Gelpi A., Young S., Sulzby E. (1998). Influences of gender and open-ended software on first graders' collaborative composing activities on computers. *Journal of Computing in Childhood Education 9*, 3–42.

Norris, J. T., Pauli, R. A., Bray, D. E. (2007). Mood change and computer anxiety: A comparison between computerised and paper measures of negative affect. *Computers in Human Behavior, 23*, 2875–2887.

Norris, P. (2001). *Digital divide: civic engagement, information poverty and the internet in democratic societies.* New York: Cambridge University Press.

Nowakowska, M. (1975). *Psychologia ilościowa z elementami naukometrii*. Warszawa: Wydawnictwo PWN.

Nunnally, J. C., Bernstein, I. H. (1994). *Psychometric theory.* New York: McGraw-Hill.

Orlando, M., Marshall, G. N. (2002). Differential item functioning in a Spanish translation of the PTSD checklist: Detection and evaluation of impact. *Psychological Assessment, 14*, 50–59.

Osgood, D. W., McMorris, B. J., Potenza, M. T. (2002). Analyzing multiple-item measures of crime and deviance I: Item response theory scaling. *Journal of Quantitative Criminology, 18*, 267–296.

Ostini, R., Nering, M. L. (2006). *Polytomous item response theory models*. Thousand Oaks, CA: Sage Publications.

Ouimet, J. A., Hanson, G. R. (1997). *Research methodology in the information age: a comparision of two survey techniques.* Pobrano 20.06.2011 ze strony:www.eric.ed.gov/PDFS/ED408341.pdf.

Pagendarm, M., Schaumburg, H. (2006). Why are users banner-blind? The impact of navigation style on the perception of web banners. *Journal of Digital Information, 2*. Pobrane 12.06.2009, ze strony http://journals.tdl.org/jodi/article/view/36/38.

Paluchowski, W. J. (1991). *Diagnozowanie osobowości. Testowanie – interpretacja – interwencja*. Poznań: Wydawnictwo Naukowe UAM.

Paluchowski, W. J. (2007). Nowe pomysły, stare problemy – wykorzystanie komputerów w psychologii. W: J. M. Brzeziński (red.), *Psychologia. Między teorią, metodą i praktyką* (s. 269–282). Poznań: Wydawnictwo Naukowe UAM.

Perkins, B. (1993). *Differences between computer administered and paper administered Computer Anxiety and Performance measures.* Raport z badań pobrany 20.06.2011 ze strony http://www.eric.ed.gov/PDFS/ED355905.pdf.

Peteron, L., Johannsson, V., Carlsson, S. G. (1996). Computerized testing in a hospital setting: Psychometric and psychological effects. *Computer in Human Behavior, 12,* 339–350.

Peytchev, A., Couper, M. P., McCabe, S. E., Crawford S. D. (2006). Web survey design. Paging versus scroling. *Public Opinion Quarterty, 70*, 596–607.

Peytchev, A., Crawford, S. (2005). A typology of real-time validations in web-based surveys. *Social Science Computer Review, 23*, 235–249.

Phillips, J. S., Lord, R. G. (1986). Notes on the practical and theoretical consequences of implicit leadership theories for the future of leadership measurement. *Journal of Management, 12*, 21–42.

Pietrowicz, K. (2002). Nowa stratyfikacja społeczna? Digital divide a Polska. W: L. Haber (red.), *Formowanie się społeczeństwa informacyjnego. Krakowski eksperyment internetowy AGH* (s. 255–260). Kraków: Wydawnictwo Naukowe AGH.

Pomplun, M. (2007). A bifactor analysis for a mode-of-administration effect. *Applied Measurement in Education*, *20*, 137–152.

Porter, S. R., Whitcomb, M. E. (2007). Mixed-mode contacts in web surveys. Paper is not necessarily better. *Public Opinion Quarterly*, *71*, 635–648.

Pötschke, P. (2004). *Paper and pencil or online? Methodological experience from an employee survey.* Pobrano 20.06.2011 ze strony: www.uni-kassel.de/~poetschk/paper2.htm.

Pouwer, F., Snoek, F. J., van der Ploeg, H. M., Heine, R. J. Brand, A. N. (1998). A comparison of the standard and the computerized versions of the Well-being Questionnaire (WBQ) and the Diabetes Treatment Satisfaction Questionnaire (DTSQ). *Quality of Life Research*, *7*, 33–38.

Powella, M. B., Wilsonb, C., Hastya, M. K. (2002). Evaluation of the usefulness of 'Marvin'; a computerized assessment tool for investigative interviewers of children. *Computers in Human Behavior*, *18*, 577–592.

R Development Core Team (2011). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria.

Reardon, R., Loughead, T. (1988). A comparision of paper-and-pencil and computer version of the self-directed search. *Journal of Counseling and Development 67*, 249–252.

Reckase, M. D. (2009). *Multidimensional Item Response Theory*. New York: Springer

Reips, U.-D. (2001). Standards for Internet-based experimenting. *Experimental Psychology, 49,* 243–256.

Reise, S. P., Haviland, M. G. (2005). Item response theory and the measurement of clinical change. *Journal of Personality Assessment*, *84*, 228–238.

Reise, S. P., Waller, N. G. (1990). Fitting the two-parameter model to personality data. *Applied Psychological Measurement*, *14*, 45–58.

Reise, S. P., Waller, N. G. (2003). How many IRT parameters does it take to model psychopathology items? *Psychological Methods*, *8*, 164–184.

Reise, S. P., Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement*, *27*, 133–144.

Rice, R. E., Katz, J. E. (2003). Comparing internet and mobile phone usage: Digital divides of usage, adoption, and dropouts. *Telecommunications Policy*, *27*, 597–623.

Robinson, J. P., Barth, K., Kohut, A. (1997). Social impact research – personal computers, mass media, and use of time. *Social Science Computer Review*, *15*, 65–82.

Rosen, L. D., Sears, D. C., Weil, M. M. (1987). Computerphobia. *Behavior Research Methods, Instruments, & Computers*, *19*, 167–179.

Rouse, S. V., Finger, M. S., Butcher, J. N. (1999). Advances in clinical personality measurement: An item response theory analysis of the MMPI-2 PSY-5 scales. *Journal of Personality Assessment*, *72*, 282–307.

Russel, M., Haney, W. (1996). *Testing writing on computers: results of a pilot syudy to compare student Writing Test performance via computer or via paper-and-pencil.* Praca prezentowana na Mid-Atlantic Alliance for Computers & Writing Conference.

Russella, C. G., Flighta, I., Lepparda, P., van Lawick van Pabstb, J. A., Syrettea, J. A., Coxa, D. N. (2004). A comparison of paper-and-pencil and computerised methods of ‘‘hard’’ laddering. *Food Quality and Preference*, *15*, 279–291.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, *34*, 110–114.

Sanchez, M. E. (1992). Effects of questionnaire design on the quality of survey data. *Public Opinion Quarterly*, *56*, 206–217.

Sargent, S. L. (2007). Image effects on selective exposure to computer-mediated news stories. *Computers in Human Behavior*, *23*, 705–726.

Scandura, T. A., Williams, E. A. (2000). Research methodology in management: Current practices, trends, and implications for future research. *Academy of Management Journal*, *43*(6), 1248–1264.

Schulenberg, S. E., Yutrzenka, B. A. (2001). Equivalence of Computerized and Conventional Versions of the Beck Depression Inventory II (BDI-II). *Current Psychology*, *20*, 216–230.

Schwarz, N., Knauper, B., Hippler, H-J., Noelie-Neumann, E., Clark, F. (1991). Rating scales: Numeric values may change the meaning of scale labels. *Public Opinion Quarterly*, *55*, 618–630.

SdTSwPiP, (2007). Standardy dla testów stosowanych w psychologii i pedagogice. Gdańsk: Gdańskie Wydawnictwo Psychologiczne.

Sherman, J. (2003). *History of the Internet.* Berlin: Demco Media.

Shih, T-H., Fan, X. (2008). Comparing response rates from web and mail surveys: A meta-analysis. *Field Methods*, *20*, 249–271.

Shrestha, S., Owens, J. W. (2009). Eye movements analysis of text-based web page layouts. *Usability News*, *11*. Pobrano 12.06.2009 ze strony: www.surl.org/usabilitynews/111/eyetracking.asp.

Shrout, P. E., Yager, T. J. (1989). Reliability and validity of screening scales: effect of reducing scale length. *Journal of Clinical Epidemiology*, *42*, 69–78.

Silverstein, A. B. (1982). Validity of Satz-Mogel-Yudin type short forms. *Journal of Consulting and Clinical Psychology*, *50*, 20–21.

Silverstein, A. B. (1990). Short forms of individual intelligence tests. *Psychological Assessment: A Journal of Consulting and Clinical Psychology*, *2*, 3–11.

Simonson, M. R., Maurer, M., Montag-Torardi, M., Whitaker, M. (1987). Development of a standardized test of computer literacy and computer anxiety scale. *Journal of Educational Computing Research*, *3*, 231–247.

Sing, L.,Cheung, P. Ch. (2010). Creativity assessment: Comparability of the electronic and paper-and-pencil versions of the Wallach–Kogan Creativity Tests. *Thinking Skills and Creativity*, *5*, 101–107.

Slaughter, L., Harper, B., Norman, K. (1994). *Assessing the Equivalence of the Paper and On-line Formats of the QUIS 5.5.* 2nd Annual Mid-Atlantic Human Factors Conference. Pobrano 20.07.2011 ze strony: lap.umd.edu/quis/publications/slaughter1994.

Smith, P. C., Stanton, J. M. (1998). Perspectives on the measurement of job attitudes: The long view. *Human Resource Management Review*, *8*(4), 367–386.

Smyth, J. D. Dillman, D. A., Christian, L. M., Stern, M. J. (2006). Comparing check-all and forced-choice question format in web surveys. *Public Opinion Quarterly*, *70*, 66–77.

Spearman, C. E. (1904). „General intelligence" objectivly determined and measured. *American Journal of Psychology*, *15*, 201–293.

Steinberg, L., Thissen, D. (1995). Item response theory in personality research. W: P. E. Shrout, S. T. Fiske (red.), *Personality research, methods, and theory: A festschrift honoring Donald W. Fiske* (s. 161–181). Hillsdale, NJ: Erlbaum.

Streiner, D. L. (2010). Measure for measure: new developments in measurement and Items Respose Theory. *Canadian Journal of Psychiatry*, *55*(3), 180–186.

Strelau, J., Angleitner, A., Newberry, B. H. (1999). *The Pavlovian Temperament Survey (PTS). An international handbook.* Seattle, Toronto, Bern, Goettingen: Hogrefe & Huber Publishers.

Strelau, J., Zawadzki, B., Angleitner, A. (1995). Kwestionariusz Temperamentu PTS: próba psychologicznej interpretacji podstawowych cech układu nerwowego według Pawłowa. *Studia Psychologiczne*, *33*(1-2), 9–48.

Swaminathan, H., Hambleton, R. K., Rogers, H. J. (2007). Assessing the fit of Item Response Theory models. *Handbook of statistics on psychometrics*. B.V: Elsevier. DOI: 10.1016/S0169-7161(06)26021-8.

Terelak J., Kobos Z. , Tarnowski A., Truszczyński O. (1994). Porównawcza analiza psychometrycznych właściwości klasycznej i komputerowej wersji wybranych testów psychologicznych. *Przegląd Psychologiczny , 37* (3), 379–386.

Thurstone, L. L. (1928). Attitudes can be measured. *American Joural of Sociology*, *33*, 529–583.

Tourangeau, R. M., Couper, M. P., Conrad, F. G. (2004). Spacing, position, and order: interpretive heuristics for visual features of survey questions. *Public Opinion Quarterly*, *68*, 368–393.

Tourangeau, R., Couper, M., Conrad, F. (2007). Color, labels and interpretive heuristics for response scales. *Public Opinion Quarterly*, *71*, 91–112.

Trouteaud, A. R. (2004). How you ask counts: A test of Internet-related components of response rates to a web-based survey. *Social Science Computer Review*, *22*, 385–392.

Tseng, H-M., Tiplady, B., Macleod, H. A., Wright, P. (1998). Computer anxiety: A comparision of pen-based personal digidal assistants, conventional computer and paper assessment of mood and performance. *British Journal of Psychology*, *89*, 599–610.

Tsutakawa, R. K., Johnson, J. C. (1990). The effect of uncertainty of item parameter estimation on ability estimates. *Psychometrika*, *55*, 371–390.

Vallejo, M. A., Mananes, G., Comeche, M. I., Diaz, M. I. (2008). Comparison between administration via Internet and paper-and-pencil administration of two clinical instruments: SCL-90-R and GHQ-28. *Journal of Behavior Therapy and Experimental Psychiatry*, *39*, 201–208.

Van de Looij-Jansen, P., M., Jan de Wilde, E. (2008). Comparison of web-based versus paper-and-pencil self-administered questionnaire: Effects on health indicators in Dutch adolescents. *Health Service Research,* 43, 1708–1721.

Van de Vijver, F. J., Harsveld, M. (1994). The incomplete equivalence of the paper-and-pencil and computerized versions of the general Aptitude Test Battery. *Journal of Applied Psychology*, *79*, 852–859.

Vehovar, V., Manfreda, K. L., Batagelj, Z. (2000). Design issues in web surveys. *Proceedings of the American Statistical Association, Survey Research Methods Section,* 983–988.

Venette, S., Sellnow, D., McIntyre, K. (2010). Charting new territory: assessing the online frontier of student ratting of instruction. *Assessment and Evalution in Higher Education*, *35*, 101–115.

Vispoel, W. P., Rocklin, T. R., Wang, T. (1994). Individual differences and test administration procedures: A comparison of Fixed-Item, Computerized-Adaptive, and Self-Adapted Testing. *Applied Measurement in Education*, *7*, 53–79.

Waller, N. G., Reise, S. P. (2009). Measuring psychopathology with non-standard IRT models: Fitting the four parameter model to the MMPI. W: S. Embretson, J. S. Roberts (red.), *New directions in psychological measurement with model-based approaches* (s. 147–173). Washington, DC: American Psychological Association.

Walter, O. B., Becker, J., Bjorner, J. B., Fliege, H., Klapp, B. F., Rose M. (2007). Development and evaluation of computer adaptive test for Anxiety (Anxiety-CAT). *Quality of Life Research*, *16*, 143–155.

Walter, O. B., Holling, H. (2008) Transitioning from Fixed-Length Questionnaires to Computer-Adaptive Versions. *Journal of Psychology*, *216*, 22–28.

Wang, S., Jiao, H., Young, M. J., Brooks, T., Olson, J. (2007). A Meta-Analysis of testing mode effects in Grade K-12 Mathematics Tests. *Educational and Psychological Measurement*, *67*, 219–238.

Warrington, E. K., James, M., Maciejewski, C. (1986). The WAIS as a lateralizing and localizing diagnostic instrument: A study of 656 patients with unilateral cerebral lesions. *Neuropsychologia*, *24*, 223-239.

Whitaker, B. G. (2007). Internet-based attitude assessment: Does gender affect measurement equivalence? *Computers in Human Behavior*, *23*, 1183–1194.

Wiechmann, D., Ryan, A. M. (2003). Reactions to computerized testing in selection contexts. *International Journal of Selection and Assessment*, *11*, 215–229.

Williams, J. E., McCord, D. M. (2006). Equivalence of standard and computerized versions of the Raven Progressive Matrices Test. *Computers in Human Behavior*, *22*, 791–800.

Wilson E. J. 2000. *Closing the Digital Divide: An Initial Review. Briefing the President.* Washington: The Internet Policy Institute. Pobrane 15.05.2009 ze strony: www.internetpolicy.org/ briefing/ErnestWilson0700.html.

Woehr, D. J., Miller, M. J., Lane, J. A. S. (1998). The development and evaluation of a computer- administered measure of cognitive complexity. *Personality and Individual Differences*, *25*, 1037–1049.

Woo-Sam, J., Zimmerman, I. L. (1973). Note on the applicability of the Kaufman formula for abbreviating the WPPSI. *Perceptual and Motor Skills*, *36*, 1121–1122.

Wood, E., Nosko, A., Desmarais, S., Ross, C., Irvine, C. (2006). Online and trditional paper-and-pencil survey administration: Examining experimenter presence, sensitive material and long surveys. *The Canadian Journal of Human Sexuality*. Pobrano 22.06.2011 ze strony: findarticles.com/p/articles/mi_go1966/is_3-4_15/ai_n29356536.

Zakrzewska, M. (1994). *Analiza czynnikowa w budowaniu i sprawdzaniu modeli psychologicznych*. Poznań: Wydawnictwo Naukowe UAM.

Zawadzki, B. (2006). *Kwestionariusze osobowości. Strategie i procedura testowania.* Warszawa: Wydawnictwo Naukowe Scholar.

Zhang, Y. (2005). Age, gender, and Internet attitudes among employees in the business world. *Computers in Human Behavior*, *21,* 1–10.

Zickar, M. J. (1998). Modeling item-level data with item response theory. *Current Directions in Psychological Science*, *7*(4), 104–109.

Zimowski, M. F., Muraki, E., Mislevy, R. J., Bock, R. D. (1996). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items*. Chicago: Scientific Software International.

Załącznik 1: kod programu R wykorzystany w badaniu

```
### Kod symulacji wyników dla modelu dychotomicznego

oblicz.2pl <- function(a=0.5, b=0.5, numOsob=5, wykres=TRUE){
      #generowanie wyników w oparciu o model 2PL
      # a - moc dyskryminacyjna
      # b - trudność zadania

      # losuje rozkład normalny m=0, sd=1,
      theta <- rnorm(numOsob)*2
      # prawdopodobieństwa wyników osób wylosowane
      ptheta <- runif(numOsob)
      # prawdopodobieństwa wyznaczone z krzywej logitu
      pthetaObliczone <- 1/(1+exp(-1.702*a*(theta-b)))
      # przypisanie 0-1
      odp <- ifelse(ptheta < pthetaObliczone, 1, 0)


      if(wykres==FALSE){
            #zwróć 0-1 dla żądanej liczby osób
            return(odp)
            } else {
            #rysuj wykres
            plot(theta, pthetaObliczone, ylim=c(0,1), xlim=c(-3,3),
xlab="theta", ylab="p(theta)", pch=16, cex=.75, cex.axis=.8, cex.lab=.
8)
            points(theta, odp, col="red", pch=1, cex=.9)
            legend(1.5,.3,c(paste("a = ",a,"  b =
",b),"logit","wynik"),col=c("white","black","red"),pch=c(1,16,1),cex=.
8)
            mtext("levels", side=4, cex=.8, col="red", padj=1)
            mtext("1    ", side=4, cex=.8, col="red", adj=1)
            mtext("    0", side=4, cex=.8, col="red", adj=0)
      }
}



generuj.2pl <- function(n, nZadan, zestaw) {
      #dla zadanego zestawu generuje macierz odpowiedzi
      # wczytuje zestaw, czyli: a min, a max, b min, b max
      a.min <- dr[1,zestaw]
      a.max <- dr[2,zestaw]
      b.min <- dr[3,zestaw]

      #---definicja "miejsca" dla danych
      odp <- matrix(,n,nZadan)
      a <-  runif(nZadan,a.min,a.max)
      b <- rnorm(nZadan,b.min,1)

      #w petli wyznaczam prawdopodobne odpowiedzi przy danych a i b
oraz dla zadanej liczby osób
      for(i in 1:nZadan){
            odp[,i] <- oblicz.2pl(a[i], b[i], n, FALSE)
      }
      return(odp)
}
```

```
populacja.2pl <- function(listaA, listaB, nZadan=60, nOsob=1){
      #n = 10.000 po prostu populacja ;-)
      odp <- matrix(,nOsob,nZadan)
      for(i in 1:nZadan){
            odp[,i] <- oblicz.2pl(listaA[i], listaB[i], nOsob, FALSE)
      }
      return(odp)
}



omnibus.wyniki <- function(dane){
      #na wejściu matryca z danymi: 60 itemów x ilość osób
      numOsob = dim(dane)[1]

      #alokacja pamięci - osoby w zbiorze z danymi x (2 podskal + 1
wynik ogólny)
      omnibus.sumy <- matrix(,numOsob, 3)

      #zestawy pozycji testowych
      #wynik ogólny
      wo <- 1:60
      #czynnik wiedzy
      cw <-
c(1,2,3,10,11,12,16,17,18,25,26,27,31,32,33,40,41,42,46,47,48,55,56,57
)
      #czynnik rozumowania
      cr <-
c(7,8,9,13,14,15,22,23,24,28,29,30,37,38,39,43,44,45,52,53,54,58,59,60
)

      for(i in 1:numOsob){
            omnibus.sumy[i,1] <- sum(dane[i,wo])
            omnibus.sumy[i,2] <- sum(dane[i,cw])
            omnibus.sumy[i,3] <- sum(dane[i,cr])
      }
      srednia <- mean(omnibus.sumy[,1])
      odchsd  <- sd(omnibus.sumy[,1])
      return(c(srednia,odchsd))
}



oblicz.srednie <- function (granicaProby,zmienna=1){
      #liczy średnie dla grup o losowym składzie
      #dla liczebności od 5 do podanej granicy
      srednie <- NULL
      for(i in 5:granicaProby){
            srednie[i-5] <- omnibus.wyniki(h1[sample(1:100000,i),])
[zmienna]
      }
      return(srednie)
}
```

```
### - kod symulacji wyników w modelu GRM

oblicz.grm <- function(numOsob=5, a.low=0, a.high=2, b.mean=.5,
b.sd=2, wykres=FALSE){
      #generuje wyniki dla jednego zadania dla zadanej liczby osób

      a <- runif(1, a.low, a.high)
      b1 <- rnorm(1, b.mean, b.sd)
      #wartości reprezentują narastającą trudność itemów
      b2 <- b1 + .7 + runif(1,-.2,.2)
      b3 <- b2 + .7 + runif(1,-.2,.2)

      # losuje wyniki osób z rozkładu normalny m=0, sd=1,
      theta <- rnorm(numOsob)
      # prawdopodobieństwa wyników osób wylosowane
      ptheta <- runif(numOsob)
      # prawdopodobieństwa wyznaczone z krzywej logitu
      pthetaObliczone1 <- 1/(1+exp(-1.702*a*(theta-b1)))
      pthetaObliczone2 <- 1/(1+exp(-1.702*a*(theta-b2)))
      pthetaObliczone3 <- 1/(1+exp(-1.702*a*(theta-b3)))

      #przedziały dla prawdopodobieństwa
      ptheta1 <- 1 - pthetaObliczone1
      ptheta2 <- pthetaObliczone1 - pthetaObliczone2
      ptheta3 <- pthetaObliczone2 - pthetaObliczone3
      ptheta4 <- pthetaObliczone3

      # przypisanie 1-2-3-4
      odp <- ifelse(ptheta < ptheta1, 1, ifelse(ptheta >= ptheta1 &
ptheta < ptheta2, 2, ifelse(ptheta >= ptheta2 & ptheta < ptheta3, 3,
4)))

      if(wykres==FALSE){
            #zwraca wyniki liczbowe
            #zwróć 1-2-3-4 dla żądanej liczby osób
            return(odp)
            } else {
            #lub rysuje wykres
            plot(theta, ptheta1, ylim=c(0,1), xlab="theta",
ylab="p(theta)", pch=16, cex=.2, cex.axis=.8, cex.lab=.8, col="black",
main=paste("Parametry: a=",round(a,2),” b1=",round(b1,2),”
b2=",round(b2,2),” b3=",round(b3,2)), cex.main=.9)
            points(theta, ptheta2, col="black", pch=16, cex=.2)
            points(theta, ptheta3, col="black", pch=16, cex=.2)
            points(theta, ptheta4, col="black", pch=16, cex=.2)
            points(theta, (odp-1)/3, col="red", pch=1, cex=.7)
            legend(2,.4,c(paste("a = „,a,”
b = „,b),"logit","wynik"),col=c("white","black","red"),pch=c(1,16,1),c
ex=.8)
      }
}
```

```
### Kod do analizy problemu wpływu wielkości próby na wyniki
#ustalenie parametrów
ab <- matrix(,4,9, dimnames = list(c("a min","a max","b min","b max"),
1:9))
#zestaw parametrów do poszczególnych symulacji
#wielkości N od 50 do 1000 co 50
for(i in 1:20){n[i]<-(50*i)}
#a
ab[1,] <- c(rep(.25,3),rep(.65,3),rep(1.35,3))
ab[2,] <- c(rep(.64,3),rep(1.34,3),rep(1.69,3))
#b
ab[3,] <- rep(c(-1.5,0,1),3)
ab[4,] <- rep(c(0,1,2.5),3)
#k prób o wielkości k*50
k <- 20
proby <- matrix(,1000,k+1)
wyniki <- matrix(,k,10)
colnames(wyniki) <- c("a","b","n","mean","sd","range","p-
value","skew","kurosis","se")

#powtórka dla każdego warunku
for(j in 1:9){
#wczytanie parametrów dla pozycji testowych
a <- (ab[1,j]+ab[2,j])/2
b <- (ab[3,j]+ab[4,j])/2
#przy zadanych parametrach a i b z zestawu generuje 10.000 populację
#i obliczam proporcję odpowiedzi prawidłowych i nieprawidłowych
#powtarzam to 1000 razy
for(i in 1:1000){proby[i,k+1] <- (oblicz.2pl(a,b,10000,wykres=FALSE))/
10000}

#dla kolejnych wielkości próby od 50 do 1000
for(n in 1:k){
     #obliczam analogiczną proporcję 1000 razy, dla próby n
elementowej
     for(i in 1:1000){proby[i,n] <- (oblicz.
2pl(a,b,n*50,wykres=FALSE))/(n*50)}
          #różnica między rozkładami wyników
          wyniki[n,1] <- a
          wyniki[n,2] <- b
          wyniki[n,3] <- n*50
          wyniki[n,4] <- wilcox.test(proby[,k+1],proby[,n])$p.value
          tmp <- describe(proby[,n])
          wyniki[n,5] <- tmp$mean
          wyniki[n,6] <- tmp$sd
          wyniki[n,7] <- tmp$range
          wyniki[n,8] <- tmp$skew
          wyniki[n,9] <- tmp$kurtosis
          wyniki[n,10] <- tmp$se
     }
#zapis do pliku, za pierwszym razem nazwy kolumn
     if(j==1){
          write.table(wyniki, file="wyniki.csv", sep=";", dec=",",
row.names=FALSE, col.names=TRUE)
     }else{
          write.table(wyniki, file="wyniki.csv", sep=";", dec=",",
append=TRUE, row.names=FALSE, col.names=FALSE)
     }
}
```

```
### analiza wyników z testu Omnibus
#przygotować dane
#wczytanie z pliku
omni <- read.csv("~/Desktop/omni.csv", sep=";", header=T)

#opis danych
opis.omni <- descript(omni)

#informacja o brakach danych
opis.omni$missin

#same dane, usuwam informację o płci, wieku i lp
dane.omni <- omni[3:63]

#opis ogólny wyników w skali, rzetelności
omni.wyniki <- omnibus.wyniki(dane.omni, "przeliczone")
summary(omni.wyniki)
descriptive.table(vars = d(wo,cw,cr),data= omni.wyniki, func.names
=c("Mean","St. Deviation","Valid N","Skew","Kurtosis","Median"))

#wykresy
boxplot(ogólny ~ płeć, data=dw, subset=płeć=="M", col="orange",
xlim=c(1.5,3.5), main="WO")
boxplot(ogólny ~ płeć, data=dw, subset=płeć=="K", col="yellow", add=T)
boxplot(czynnik.rozumienie ~ płeć, data=dw, subset=płeć=="M",
col="orange", xlim=c(1.5,3.5), main="CR")
boxplot(czynnik.rozumienie ~ płeć, data=dw, subset=płeć=="K",
col="yellow", add=T)
boxplot(czynnik.wiedzy ~ płeć, data=dw, subset=płeć=="M",
col="orange", xlim=c(1.5,3.5), main="CW")
boxplot(czynnik.wiedzy ~ płeć, data=dw, subset=płeć=="K",
col="yellow", add=T)

###analiza wg modelu 3PL
library(ltm)

#model ze stałym a
fit3pl.r <- tpm(omni, type = "rasch")
fit3pl.r

#model ze zmiennym a
fit3pl <- tpm(omni.wyniki)
fit3pl

#wartości dla każdego badanego
#UWAGA - tylko unikalne wektory odpowiedzi, POSORTOWANE!
theta.3pl <- factor.scores(fit3pl)

theta <- theta.3pl$score.dat$z1
se.theta <- theta.3pl$score.dat$se.z1
parametry <- theta.3pl$coef

#wykres 3 przykładowych ICC
plot(omnifit, type="ICC", items=c(2,11,17), xlab="Poziom wiedzy",
ylab="Prawdopodobieństwo")
text(2.6,.55,"a=0,62; b=1,11; c=0,10", col="red")
text(1,.25,"a=1,67; b=1,49; c=0,24", col="green")
text(-1.6,.05,"a=1,69; b=-1,15; c=0")
#porównanie modeli
anova(fit3pl.r, fit3pl.dowolne.omni)
```

```
#analiza lokalnej niezależności: margins()
margins(fit.dowolne.omni)

#test jednowymiarowości
out <- unidimTest(rasch(LSAT))
plot(out, type = "b", pch = 1:2)
legend("topright", c("Real Data", "Average Simulated Data"), lty = 1,
    pch = 1:2, col = 1:2, bty = "n")

#wykresy
#po dwa wykresy dla najładniejszych i najbrzydszych IIC + OCC
par(mfrow = c(3,1))
pozycja <- 1
plot.grm(fit.dowolne.omni, type="ICC", items=pozycja)
plot.grm(fit.dowolne.omni, type="IIC", items=pozycja)
plot.grm(fit.dowolne.omni, type="OCCu", items=pozycja)

#zapisać do tabeli wartości a, b1 i c dla itemów, które zostały
summary(fit.dowolne.omni)
coef.grm(fit.dowolne.omni)
```

```
#hipoteza 2 - sprawdzanie różnic między wersjami: łatwa, średnia,
trudna i losowa



wersja.cw <-
matrix(c(c(25,31,10,1,33,3,27,2),c(55,41,48,11,47,46,26,18),c(42,12,57
,40,16,56,32,17),sample(cw,8)),8,4)
wersja.cr <-
matrix(c(c(43,36,21,29,4,23,28,22,7,14,59,8),c(15,6,37,52,19,45,60,39,
5,9,20,30),c(51,50,34,44,58,35,49,24,38,53,13,54),sample(cr,12)),12,4)

#theta dla 3PL
p.theta <- function(theta, a,b,c){
      tmp <- exp(a*(theta-b))
      p <- c + (1-c) * (tmp/(1+tmp))
      return(p)
}

#wynik sumaryczny pi (P dla i)
# pi <- 1/n*suma P(theta)
pi <- function(nrPytan, nrOsoby){
      #omnibus.tpm.parametry$  z factor.scores(tpm)
      theta <- omnibus.tpm.parametry$score.dat$z1[nrOsoby]
      #theta <- 1
      #wektor odpowiedzi
      odpowiedzi <- as.logical(omnibus.tpm.parametry$score.dat[nrOsoby,
nrPytan])
      pi <- p <- 0
      n <- length(nrPytan)
      for(i in 1:n){
            #parametry = parametry c,b,a
            p <- p.theta(theta, a=omnibus.tpm.parametry$coef[nrPytan[i],
3], b=omnibus.tpm.parametry$coef[nrPytan[i],2],
c=omnibus.tpm.parametry$coef[nrPytan[i],1])
            if(odpowiedzi[i]){
                  pi <- pi + p
            } else {
                  pi <- pi + (1-p)
            }
      }
      return(pi/n)
}



#generowanie wynikowego csv z danymi
export[,1] <- omnibus.tpm.parametry$score.dat$z1
export[,2] <- omnibus.tpm.parametry$score.dat$se.z1
for(i in 1:307){export[i,3] <- pi(omnibus.losowe, i)}
for(i in 1:307){export[i,4] <- pi(wersje[,3][wersje$trudność==1 &
wersje$czynnik==1], i)}
for(i in 1:307){export[i,5] <- pi(wersje[,3][wersje$trudność==2 &
wersje$czynnik==1], i)}
for(i in 1:307){export[i,6] <- pi(wersje[,3][wersje$trudność==3 &
wersje$czynnik==1], i)}
for(i in 1:307){export[i,7] <- pi(wersje[,3][wersje$trudność==1 &
wersje$czynnik==2], i)}
for(i in 1:307){export[i,8] <- pi(wersje[,3][wersje$trudność==2 &
wersje$czynnik==2], i)}
```

```
for(i in 1:307){export[i,9] <- pi(wersje[,3][wersje$trudność==3 &
wersje$czynnik==2], i)}
for(i in 1:307){export[i,10] <- pi(wersje[,3][wersje$trudność==1], i)}
for(i in 1:307){export[i,11] <- pi(wersje[,3][wersje$trudność==2], i)}
for(i in 1:307){export[i,12] <- pi(wersje[,3][wersje$trudność==3], i)}
for(i in 1:307){export[i,13] <- sum(omnibus.wyniki.posortowane[i,
1:60])}
for(i in 1:307){export[i,14] <-
sum(omnibus.wyniki.posortowane[i,c(7,8,9,13,14,15,22,23,24,28,29,30,37
,38,39,43,44,45,52,53,54,58,59,60)])}
for(i in 1:307){export[i,15] <-
sum(omnibus.wyniki.posortowane[i,c(1,2,3,10,11,12,16,17,18,25,26,27,31
,32,33,40,41,42,46,47,48,55,56,57) ])}
write.table(round(export,5), file="omnibus.wyniki-1.csv", sep=";",
dec=",")

#estymacja theta
for(i in 1:307){
     export[i,1] <- obliczTheta(omnibus.tpm.parametry$score.dat[i,
1:60],omnibus.tpm.parametry$coef,c(1:60))}
}

#wyniki skrócone jako sumy
export <- matrix(,307,15)
colnames(export) <- c("theta", "se", "losowy", "łatwy, rozumienie",
"średni, rozumienie", "trudny, rozumienie", "łatwy, wiedza", "średni,
wiedza", "trudny, wiedza", "łatwy", "średni", "trudny", "ogólny",
"czynnik rozumienie", "czynnik wiedzy")
export[,1] <- round(omnibus.tpm.parametry$score.dat$z1,5)
export[,2] <- round(omnibus.tpm.parametry$score.dat$se.z1,5)
for(i in 1:307){export[i,3] <- sum(omnibus.tpm.parametry
$score.dat[i,omnibus.losowe])}
for(i in 1:307){export[i,4] <- sum(omnibus.tpm.parametry
$score.dat[i,wersje[,3][wersje$trudność==1 & wersje$czynnik==1]])}
for(i in 1:307){export[i,5] <- sum(omnibus.tpm.parametry
$score.dat[i,wersje[,3][wersje$trudność==2 & wersje$czynnik==1]])}
for(i in 1:307){export[i,6] <- sum(omnibus.tpm.parametry
$score.dat[i,wersje[,3][wersje$trudność==3 & wersje$czynnik==1]])}
for(i in 1:307){export[i,7] <- sum(omnibus.tpm.parametry
$score.dat[i,wersje[,3][wersje$trudność==1 & wersje$czynnik==2]])}
for(i in 1:307){export[i,8] <- sum(omnibus.tpm.parametry
$score.dat[i,wersje[,3][wersje$trudność==2 & wersje$czynnik==2]])}
for(i in 1:307){export[i,9] <- sum(omnibus.tpm.parametry
$score.dat[i,wersje[,3][wersje$trudność==3 & wersje$czynnik==2]])}
for(i in 1:307){export[i,10] <- sum(omnibus.tpm.parametry
$score.dat[i,wersje[,3][wersje$trudność==1]])}
for(i in 1:307){export[i,11] <- sum(omnibus.tpm.parametry
$score.dat[i,wersje[,3][wersje$trudność==2]])}
for(i in 1:307){export[i,12] <- sum(omnibus.tpm.parametry
$score.dat[i,wersje[,3][wersje$trudność==3]])}
#wo / cw / cr
for(i in 1:307){export[i,13] <- sum(omnibus.tpm.parametry
$score.dat[i,c(1:60)])}
for(i in 1:307){export[i,14] <-
sum(omnibus.wyniki.posortowane[i,c(1,2,3,10,11,12,16,17,18,25,26,27,31
,32,33,40,41,42,46,47,48,55,56,57)])}
for(i in 1:307){export[i,15] <-
sum(omnibus.wyniki.posortowane[i,c(7,8,9,13,14,15,22,23,24,28,29,30,37
,38,39,43,44,45,52,53,54,58,59,60)])}
write.table(export, file="omnibus.wyniki-sumy.csv", sep=";", dec=",")
```

```
}


#dane do analizy regresji
parametry <- #wskazać tabelę z parametrami IRT dla Omnibusa
#wyniki <-
n.pytań <- nrow(parametry) #60
n.próby <- nrow(wyniki) #306

#definicje zmiennych
tabela.wyników <- matrix(,(n.pytań-9)*n.próby,8)      # tabela wyników
średnich
colnames(tabela.wyników) <- c("średnie a", "średnie b", "średnie c",
"i", "wynik skrócony", "se","theta", "wynik pełny")
}



#oblicz 1 osobę
o1o <- function(osoba=1, pytania=c(1,2,3,4,5)){
      sr.a <- mean(parametry[pytania,3])
      sr.b <- mean(parametry[pytania,2])
      sr.c <- mean(parametry[pytania,1])
      tmp <- wyniki[osoba,pytania]
      sr.suma.i <- sum(tmp)/length(pytania)
      war <- matrix(,1,length(pytania))
      for(k in 1:length(pytania)){war[k]<-((tmp[[k]]-sr.suma.i)^2)}
      stde <- sqrt(sum(war))/length(pytania)
      wo <- sum(wyniki[osoba,1:n.pytań])
      return(c(sr.a,sr.b,sr.c,length(pytania),
(sr.suma.i*n.pytań),stde,theta[osoba],wo))
}
```

```
#MCMC dla każdej osoby
mc.o1o <- function(osoba,liczba.iteracji=3, ile.pytań=5){
     tabela.mc <- matrix(,liczba.iteracji,8)               # tabela
surowych
     for(mc in 1:liczba.iteracji){    #iteracje dla MCMC
          próbka <- sample(1:n.pytań,ile.pytań)
          tabela.mc[mc,] <- o1o(osoba,próbka)
          }
     return(apply(tabela.mc,2,mean))
}


#powtórz dla wszystkich
doit <- function(liczba.iteracji=3, zapis=FALSE){ #ile razy różne
zestawy o tej samej liczbie pytań
     for(i in 5:55){                        #zestawy pytań od 5 do 55
z 60 możliwych w Omnibusie
          for(o in 1:n.próby){                  #dla każdej osoby
          wiersz <- o+(n.próby*(i-5))
               tabela.wyników[wiersz,] <- mc.o1o(o,liczba.iteracji,i)
          }
     }
     if(zapis==FALSE){
          return(tabela.wyników)
     } else {
          write.table(tabela.wyników, file="~/Desktop/
do_regresji.csv", sep=";", dec=",")
     }
}
```

```
### analiza wyników kwestionariusza PTS

#przygotować dane
#wczytanie z pliku
     dane.pts <- read.csv("~/Desktop/pts.csv", sep=";", header=T)
#rekodowanie
     dane.pts.r <- dane.pts
     #odwrócone itemy
     ldr <-
c(1,14,16,18,21,29,40,43,51,5,13,17,19,20,24,31,41,46,47,4,6,9,23,25,3
3,42,45,55) #28
     for(i in 1:length(ldr)){
          dane.pts.r[ldr[i]] <- 5-dane.pts.r[ldr[i]]
          }
#opis danych
     opis.pts <- descript(dane.pts)
     opis.pts$perc

#informacja o brakach danych
     opis.pts$missin

#wykluczenie obserwacji z brakami danych
     completne.pts <- subset(dane.pts, complete.cases(dane.pts))
#ALTERNATYWNIE: uzupełnione średnią
     uzupelnione.pts <- dane.pts
     for(i in 1:57){
          uzupelnione.pts[i][is.na(dane.pts[i])] <- (mean(dane.pts[i],
na.rm=TRUE)
     }
#na surowych
skale.c <- oblicz.pts(completne.pts)
skale.u <- oblicz.pts(uzupelnione.pts)
#porównanie wyników completnych i uzupełnionych
t.test(skale.c, skale.u, paired=T)

#do dalszych analiz przechodzą uzupełnione
     d.pts<-dane.pts.r
     for(i in 1:57){
          d.pts[i][is.na(d.pts[i])] <- round(mean(d.pts[i],
na.rm=TRUE),0)
          }

#opis ogólny wyników w skali, rzetelności
#Siła procesu pobudzenia
pts[1,] <- c(1,3,8,10,14,16,18,21,29,32,34,38,40,43,44,49,50,51,54)

#Ruchliwość procesów nerwowych
pts[2,] <- c(5,7,12,13,15,17,19,20,22,24,26,28,31,35,37,41,46,47,56)

#Siła procesu hamowania
pts[3,] <- c(2,4,6,9,11,23,25,27,30,33,36,39,42,45,48,52,53,55,57)


#rzetelność skal
library(psych)
     print(alpha(round(d.pts[spp],0))$total)
     print(alpha(round(d.pts[rpn],0))$total)
     print(alpha(round(d.pts[sph],0))$total)
```

```
#analiza wg modelu GRM
#model ze stałym a constrained=T
      fit.stale.pts <- grm(d.pts, constrained=T, Hessian=T)

#model ze zmiennym a: constrained=F
      fit.dowolne.pts <- grm(d.pts, constrained=F, Hessian=T)

#porównanie modeli
      anova(fit.stale.pts, fit.dowolne.pts)

#sprawdzanie założnien IRT
      #analiza lokalnej niezależności: margins()
      margins.grm(fit.dowolne.pts)
      #analiza reszt
      residuals(fit.dowolne.pts)

#wykresy
      #dla kategorii
      par(mfrow = c(2,2))
      plot.grm(fit.dowolne.pts, category=1)
      plot.grm(fit.dowolne.pts, category=2)
      plot.grm(fit.dowolne.pts, category=3)
      plot.grm(fit.dowolne.pts, category=4)


#po dwa wykresy dla najładniejszych (34, 47) i najbrzydszych (4, 19)
IIC + ICC + OCCu
      par(mfrow = c(3,1))
      pozycja <- 1
      plot.grm(fit.dowolne.pts, type="ICC", items=pozycja)
      plot.grm(fit.dowolne.pts, type="IIC", items=pozycja, ylim=c(0,1))
      plot.grm(fit.dowolne.pts, type="OCCu", items=pozycja)


#na wyjściu
#lista itemów do wykluczenia wg wartości funkcji informacyjnej
      informacja <- NULL
      for(i in 1:57){informacja[i] <- information(fit.dowolne.pts,
c(-4,4), items=i)$InfoTotal}
      round(informacja,2)

#zapisać do tabeli wartości a i b1, b2, b3 dla itemów, które zostały
      summary(fit.dowolne.pts)
      coef.grm(fit.dowolne.pts)

      par.pts <- coef.grm(fit.dowolne.pts)
      paar<-matrix(,57,5)
      rownames(paar) <- names(d.pts)
      for(i in 1:57){paar[i,] <-
c(informacja[i],if(length(par.pts[[i]])<5){c(0,par.pts[[i]])}
else{par.pts[[i]]})}
      xtable(paar)

#sporządzanie wyników dla testu PTS wg klucza
#UWAGA! pozycje są rekodowane!
oblicz.pts <- function(dane){
      numObs <- dim(dane)[1]

      #alokacja pamięci - osoby w zbiorze z danymi x (3 podskale)
      pts.sumy <- matrix(,numObs,3)
```

```
      colnames(pts.sumy) <- c("SPP","RPN","SPH")
      #alokacja na pytania w skalach
      pts <- matrix(,3,19)
      ldr <- matrix(,3,10) # (ldr = lista do rekodowania ;-)

      #skale - numery pytań wszystkie
      pts[1,] <-
c(1,3,8,10,14,16,18,21,29,32,34,38,40,43,44,49,50,51,54)         #Siła
procesu pobudzenia
      pts[2,] <-
c(5,7,12,13,15,17,19,20,22,24,26,28,31,35,37,41,46,47,56)  #Ruchliwość
procesów nerwowych
      pts[3,] <-
c(2,4,6,9,11,23,25,27,30,33,36,39,42,45,48,52,53,55,57)          #Siła
procesu hamowania

      #odwrócone itemy
      ldr[1,] <- c(1,14,16,18,21,29,40,43,51,0) #0 bo musi być 10
pozycji
      ldr[2,] <- c(5,13,17,19,20,24,31,41,46,47)
      ldr[3,] <- c(4,6,9,23,25,33,42,45,55,0) #0

      #suma wszystkich + k+1*n odwróconych - 2*suma odwróconych
      #trochę to pokręcone, ale mam wszystkie itemy z danej skali
w jednym miejscu (pts)
      for(s in 1:3){
            for(i in 1:numObs){
                  pts.sumy[i,s] <- sum(dane[i,pts[s,]], na.rm=T) +
5*length(subset(ldr[s,], ldr[s,]>0)) - 2*sum(dane[i,ldr[s,]], na.rm=T)
            }
      }

      return(pts.sumy)
}
```

Załącznik 2: skrócone wersje kwestionariusza PTS

Tabela przedstawiająca, które pozycje (oznaczone przez 1) wchodziły w skład skróconych wersji podskal

Skala SPP				Skala RPN				Skala SPH			
pełna	skrócona			pełna	skrócona			pełna	skrócona		
	FA	MR	IRT		FA	MR	IRT		FA	MR	IRT
1	0	0	1	5	0	0	0	2	0	1	0
3	0	1	0	7	0	1	0	4	1	1	0
8	0	1	0	12	1	1	1	6	1	0	0
10	1	1	0	13	1	0	1	9	1	0	0
14	0	1	0	15	1	0	1	11	0	1	1
16	0	1	0	17	1	0	1	23	1	0	1
18	0	0	1	19	1	1	0	25	1	0	1
21	0	0	1	20	0	0	1	27	0	1	0
29	0	1	0	22	1	0	1	30	0	1	0
32	1	1	0	24	0	0	1	33	0	1	1
34	1	0	1	26	0	1	0	36	0	0	1
38	1	0	1	28	0	1	0	39	0	1	1
40	0	0	1	31	0	0	0	42	1	0	0
43	0	0	1	35	0	1	0	45	1	1	0
44	1	0	1	37	0	1	1	48	0	0	0
49	1	0	0	41	1	0	0	52	0	0	0
50	1	1	0	46	0	0	0	53	0	0	1
51	0	0	0	47	1	0	0	55	1	0	0
54	1	0	0	56	0	1	0	57	0	0	1
	Kappa	*T*	*p*		*Kappa*	*T*	*p*		*Kappa*	*T*	*p*
IRT~FA	-0,08	-0,35	0,73		0,35	1,54	0,13		-0,3	-1,29	0,20
IRT~MR	-0,73	-3,17	0,00		-0,3	-1,29	0,20		-0,08	-0,35	0,73
FA~MR	-0,08	-0,35	0,73		-0,3	-1,29	0,20		-0,3	-1,29	0,20

Moetoda skracania: FA – analiza czynnikowa, MR – regresja wielokrotna, IRT – podejście probabilistyczne.
Źródło: badania własne

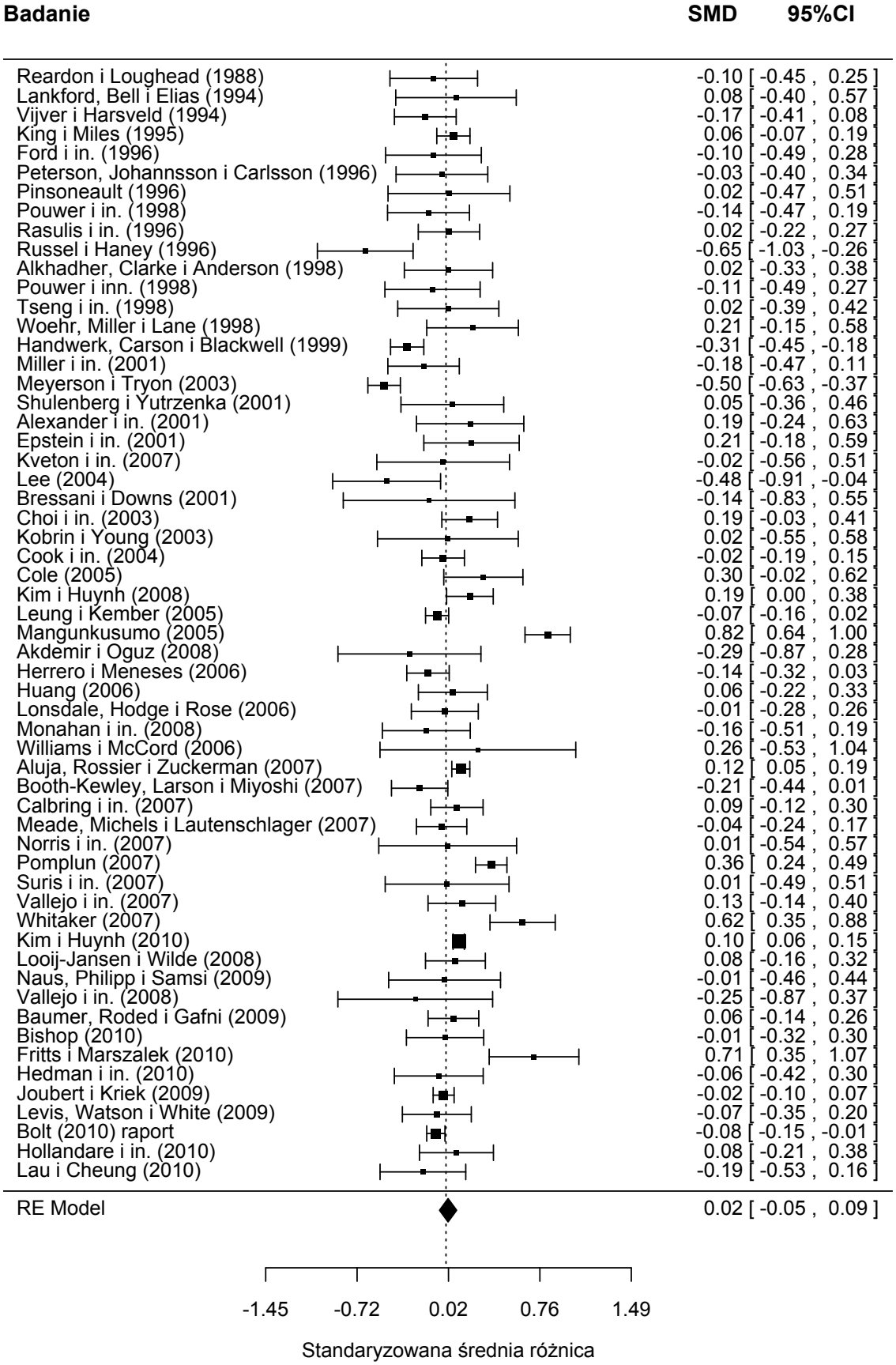
Załącznik 3.1: artykuły wykorzystane w meta-analizie

Badanie	liczba efektów	próba	plan badawczy	temat badania
Akdemir i Oguz (2008)	2	studenci	kontrola kolejności	wiedza / umiejętności
Alexander i in. (2001)	1	studenci		jakość życia
Alkhadher, Clarke i Anderson (1998)	3		równoważenie kolejności	wiedza / umiejętności
Aluja, Rossier i Zuckerman (2007)	10	studenci i pracownicy uczelni <30 lat		osobowość
Baumer, Roded i Gafni (2009)	2	studenci		wiedza / umiejętności
Bishop (2010)	2	pacjenci	kontrola kolejności	depresja
Bolt (2010) raport	2	dorośli	losowe grupy	jakość życia
Booth-Kewley, Larson i Miyoshi (2007)	4	studenci obligatoryjnie	losowe grupy	osobowość
Bressani i Downs (2001)	36	studenci i menadżerowie	kontrola kolejności	jakość życia
Calbring i in. (2007)	16	ochotnicy	kontrola kolejności	lęk / strach
Choi i in. (2003)	8	studenci	kontrola kolejności	językowy
Cole (2005)	5	dorośli	losowe grupy	jakość życia
Cook i in. (2004)	6	pacjenci		jakość życia
Epstein i in. (2001)	8	studenci	losowe grupy	osobowość
Ford i in. (1996)	8	więźniowie	równoważenie kolejności	lęk / strach
Fritts i Marszalek (2010)	3	uczniowie		lęk / strach
Handwerk, Carson i Blackwell (1999) raport	1	studenci	losowe grupy	
Hedman i in. (2010)	6	studenci	losowe grupy	lęk / strach
Herrero i Meneses (2006)	2	studenci	losowe grupy	jakość życia
Hollandare i in. (2010)	2	pacjenci	równoważenie kolejności	depresja
Huang (2006)	4		losowe grupy	wiedza / umiejętności
Joubert i Kriek (2009)	66	menadżerowie		osobowość
Kim i Huynh (2008)	2	studenci	kontrola kolejności	językowy
Kim i Huynh (2010)	2	studenci z niepełnospra wnością		językowy
King i Miles (1995)	5	studenci	losowe grupy	jakość życia
Kobrin i Young (2003)	4	studenci	równoważenie kolejności	wiedza / umiejętności
Kveton i in. (2007)	14	studenci	losowe grupy	osobowość
Lankford, Bell i Elias (1994)	6	studenci	losowe grupy	lęk / strach
Lau i Cheung (2010)	8	studenci	kontrola kolejności	osobowość
Lee (2004)	4	ochotnicy		wiedza / umiejętności
Leung i Kember (2005)	18	studenci		osobowość
Levis, Watson i White (2009)	3	studenci		wiedza / umiejętności
Lonsdale, Hodge i Rose (2006)	3	sportowcy	losowe grupy	depresja
Looij-Jansen i Wilde (2008)	6			jakość życia
Mangunkusumo (2005)	2	studenci	losowe grupy	jakość życia

Badanie	liczba efektów	próba	plan badawczy	temat badania
Meade, Lawrence i Lautenschlager (2007)	11	studenci	losowe grupy	osobowość
Meyerson i Tryon (2003)	13			jakość życia
Miller i in. (2001)	13	studenci	losowe grupy	jakość życia
Monahan i in. (2008)	2	studenci		wiedza / umiejętności
Naus, Philipp i Samsi (2009)	32	studenci	kontrola kolejności	jakość życia
Norris i in. (2007)	9	studenci	kontrola kolejności	lęk / strach
Peterson, Johannsson i Carlsson (1996)	6	pacjenci		osobowość
Pinsoneault (1996)	14		losowe grupy	osobowość
Pomplun (2007)	4	studenci	równoważenie kolejności	wiedza / umiejętności
Pouwer i in. (1998)	2	pacjenci	kontrola kolejności	jakość życia
Pouwer i inn. (1998)	8	studenci	losowe grupy	jakość życia
Rasulis i in. (1996)	3		losowe grupy	inteligencja
Reardon i Loughead (1988)	6	studenci		
Russel i Haney (1996)	5	studenci	losowe grupy	wiedza / umiejętności
Russel i in. (2004)	4	matki	losowe grupy	osobowość
Shulenberg i Yutrzenka (2001)	2	studenci	kontrola kolejności	depresja
Suris i in. (2007)	18	weterani	kontrola kolejności	osobowość
Tseng i in. (1998)	2	ochotnicy		wiedza / umiejętności
Vallejo i in. (2007)	2	studenci	kontrola kolejności	jakość życia
Vallejo i in. (2008)	34	studenci	kontrola kolejności	jakość życia
Vijver i Harsveld (1994)	7	studenci	losowe grupy	inteligencja
Whitaker (2007)	6	pracownicy	równoważenie kolejności	jakość życia
Williams i McCord (2006)	2	studenci	kontrola kolejności	inteligencja
Woehr, Miller i Lane (1998)	1	studenci	równoważenie kolejności	osobowość

Źródło: badania własne.

Załącznik 3.2: podsumowanie wielkości efektów wyników badań uwzględnionych w meta-analizie



Rycina 2.1. Podsumowanie wielkości efektów analizowanych wyników badań. Kolejność chronologiczna. Źródło: badania własne.

Załącznik 4: instrukcja do systemu badań internetowych

Spis funkcji

Wstęp 165
Logowanie 165
Panel administracyjny 166
Wprowadzenie nowego badania 167
Lista twoich badań 170
Wprowadzanie adresów e-mail Badanych 171
Pobieranie wyników 172
Zmiana hasła 173

Wstęp

System Badań Internetowych umożliwia przeprowadzanie następujących typów badań:

- zwykły,
- sekwencyjny,
- adaptacyjny 3PL,
- adaptacyjny GRM

oraz wysyłanie zaproszeń do udziału w badaniu.

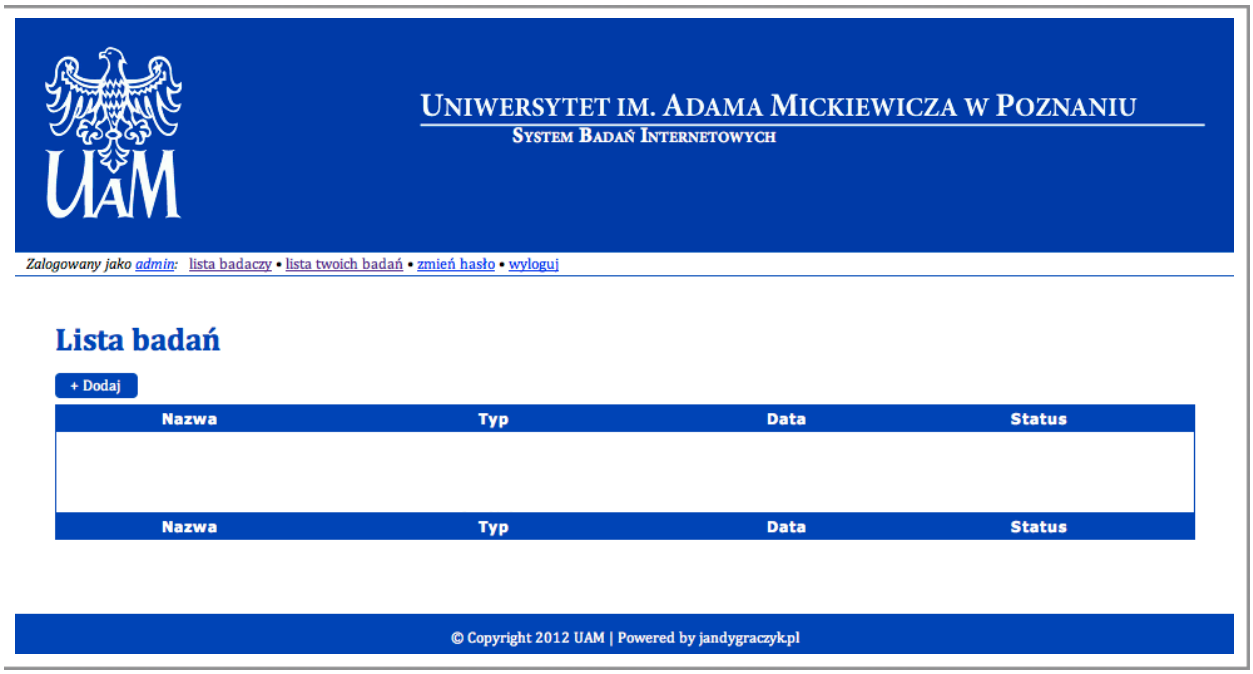
Osoba przeprowadzająca badanie zwana Badaczem winna posiadać do przeprowadzania badań przygotowane pytania w formacie CSV wg załączonego wzoru oraz bazę adresów e-mail osób, które wezmą udział w badaniu zwanych Badanymi.

Logowanie

Przejdź do strony System Badań Internetowych pod adresem badanet.amu.edu.pl. Wprowadź login i hasło uzyskane od Administratora.

Panel administracyjny

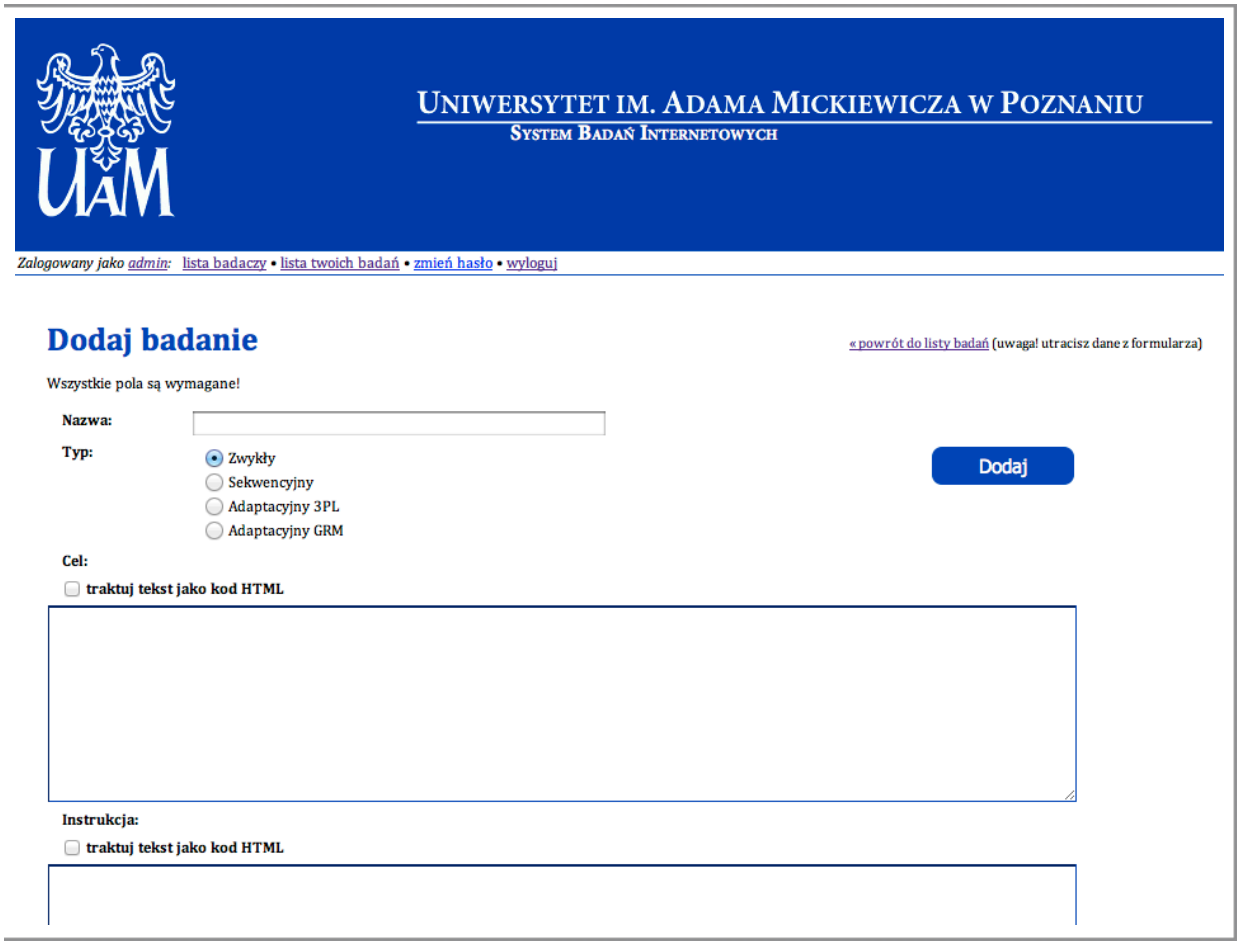
Po zalogowaniu uzyskujesz dostęp do Panelu administracyjnego. Zawiera on takie opcje jak: lista twoich badań, zmień hasło, wyloguj oraz listę Twoich badań.



Wprowadzenie nowego badania

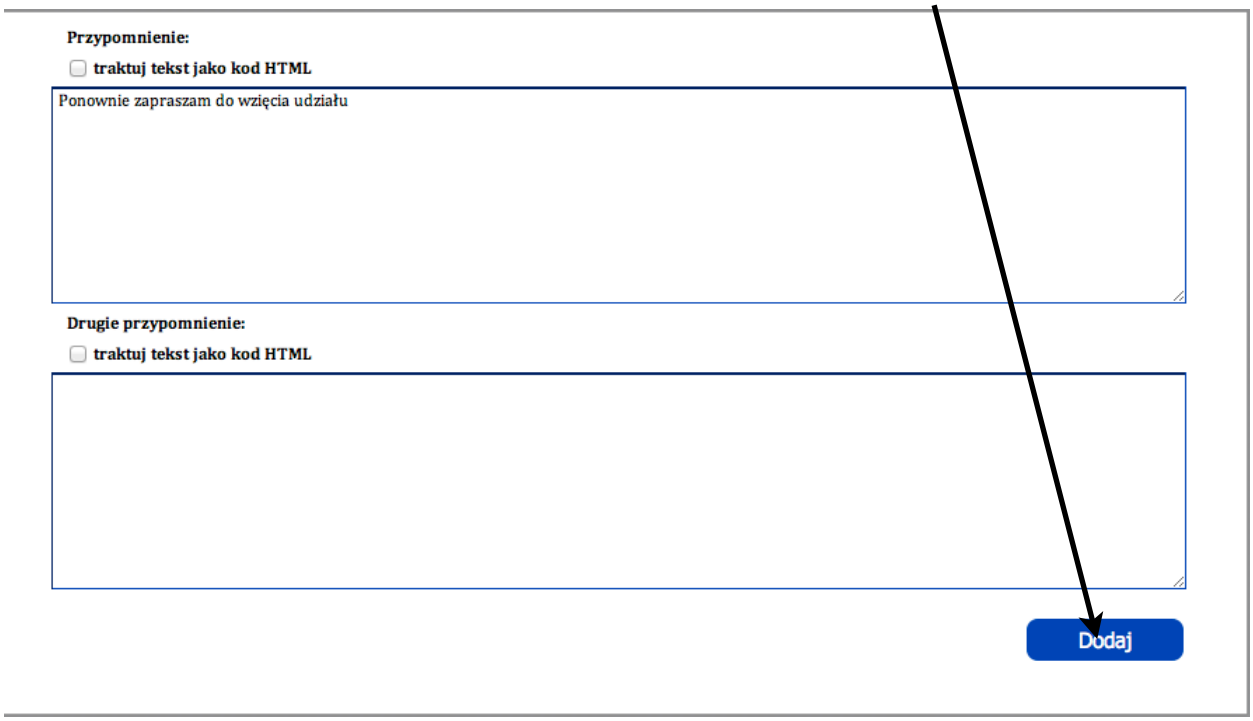
Aby wprowadzić nowe badanie, klikamy Dodaj.

Wprowadzamy informacje niezbędne do przeprowadzenia badania do pól Nazwa, Typ, Cel, Instrukcja, Zaproszenie, Przypomnienie (wysyłane do Badanych, którzy nie wzięli udziału w badaniu po wysłaniu Zaproszenia), Drugie przypomnienie (wysyłane do Badanych, którzy nie wzięli udziału w badaniu po wysłaniu Pierwszego przypomnienia).

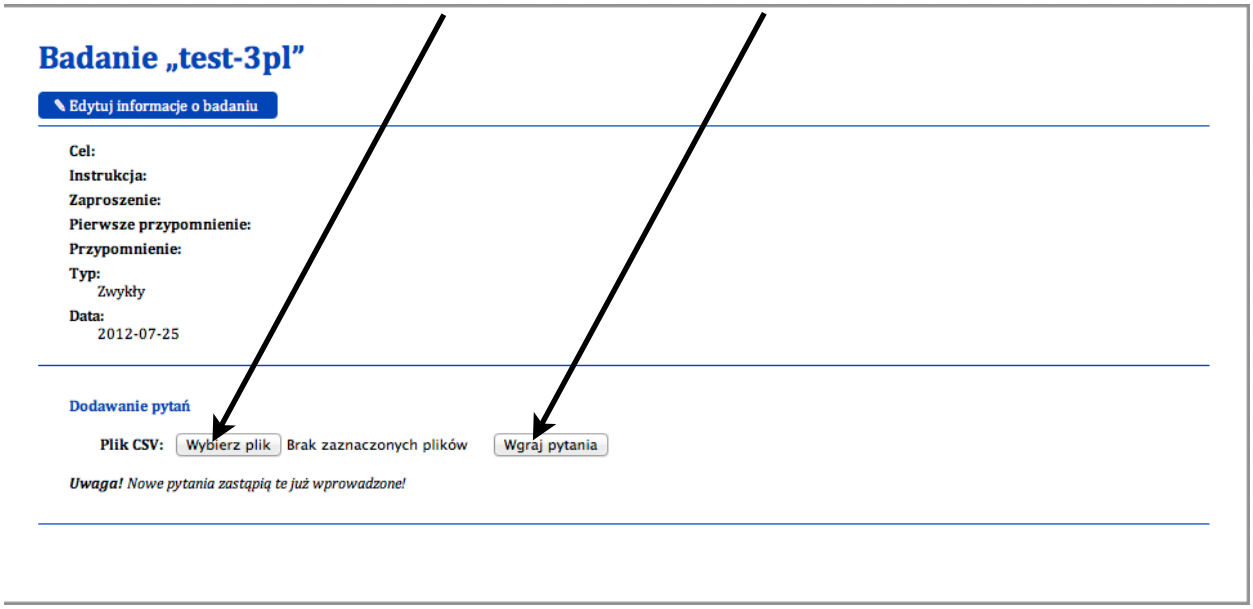


Pola począwszy od Cel umożliwiają wprowadzanie znaczników HTML. Poprzez zaznaczenie traktuj tekst jako HTML. Należy to stosować ostrożnie, ponieważ nieprawidłowe wprowadzenie znaczników HTML spowoduje nieprawidłowe wyświetlanie tekstu w zaproszeniach wysłanych do Badanych.

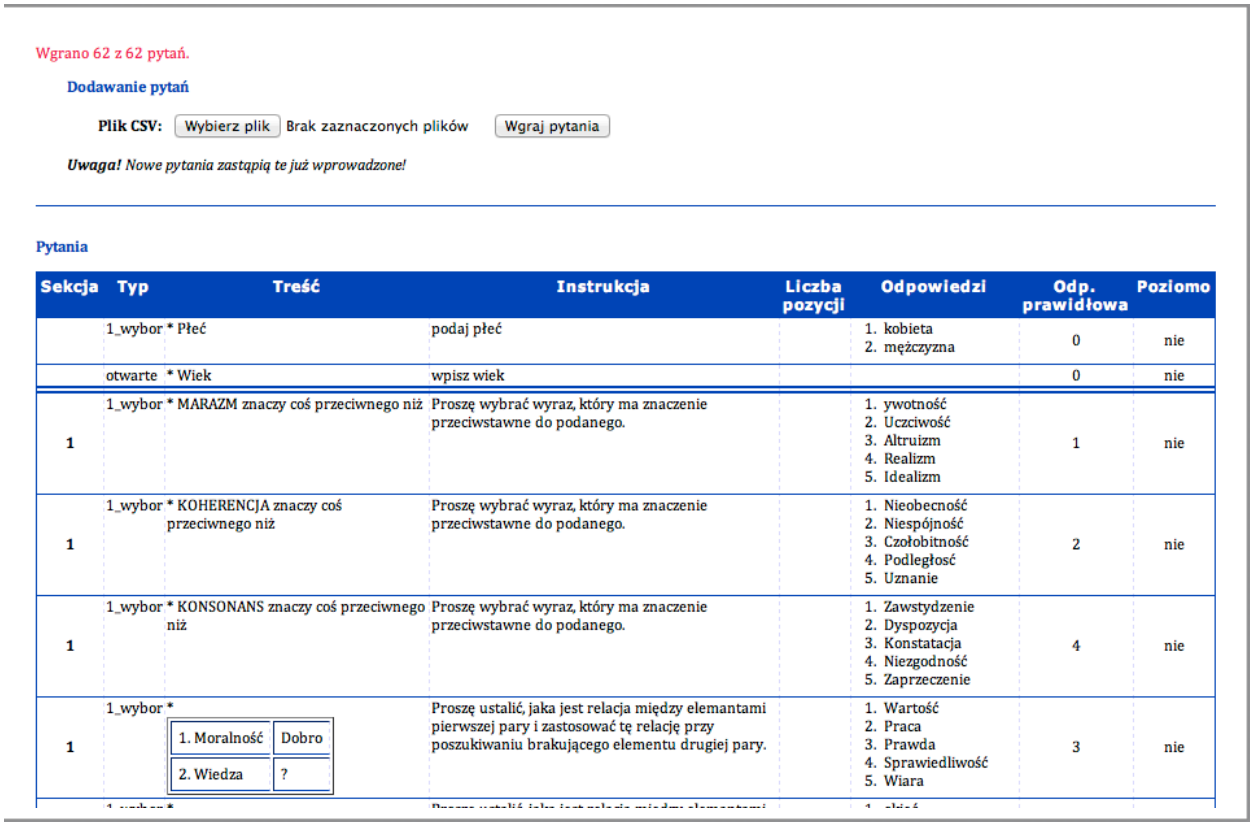
Po wprowadzeniu wymaganych informacji klikamy przycisk Dodaj.



Następnie klikamy przycisk Wybierz plik. Wybieramy plik z pytaniami w formacie CSV i naciskamy Otwórz, po czym naciskamy przycisk Wgraj pytania.



Zostaną wyświetlone wszystkie pytania, które zostały wprowadzone. Należy sprawdzić, czy wyświetlone pytania są prawidłowe, po czym nacisnąć przycisk Zatwierdź to badanie.



Wgrano 62 z 62 pytań.

Dodawanie pytań

Plik CSV: Wybierz plik Brak zaznaczonych plików Wgraj pytania

Uwaga! *Nowe pytania zastąpią te już wprowadzone!*

Pytania

Sekcja	Typ	Treść	Instrukcja	Liczba pozycji	Odpowiedzi	Odp. prawidłowa	Poziomo
	1_wybor	* Płeć	podaj płeć		1. kobieta 2. mężczyzna	0	nie
	otwarte	* Wiek	wpisz wiek			0	nie
1	1_wybor	* MARAZM znaczy coś przeciwnego niż	Proszę wybrać wyraz, który ma znaczenie przeciwstawne do podanego.		1. ywotność 2. Uczciwość 3. Altruizm 4. Realizm 5. Idealizm	1	nie
1	1_wybor	* KOHERENCJA znaczy coś przeciwnego niż	Proszę wybrać wyraz, który ma znaczenie przeciwstawne do podanego.		1. Nieobecność 2. Niespójność 3. Czołobitność 4. Podległość 5. Uznanie	2	nie
1	1_wybor	* KONSONANS znaczy coś przeciwnego niż	Proszę wybrać wyraz, który ma znaczenie przeciwstawne do podanego.		1. Zawstydzenie 2. Dyspozycja 3. Konstatacja 4. Niezgodność 5. Zaprzeczenie	4	nie
1	1_wybor	* 1. Moralność: Dobro; 2. Wiedza: ?	Proszę ustalić, jaka jest relacja między elemantami pierwszej pary i zastosować tę relację przy poszukiwaniu brakującego elementu drugiej pary.		1. Wartość 2. Praca 3. Prawda 4. Sprawiedliwość 5. Wiara	3	nie
1	1_wybor	* Jeśli ma miejsce zjawisko A, następuje po nim zjawisko B. Zjawisko B nie nastąpiło. Miało miejsce zjawisko A.	Zakładając, że dwa pierwsze zdania są prawdziwe, proszę ocenić prawdziwość zdania trzeciego jako wyprowadzonego z nich wniosku.		1. Tak 2. Nie 3. Nie wiadomo	2	tak
1	1_wybor	* Jeśli ktoś jest chory na limbozę, to w jego krwi wzrasta liczba krwinek beta. Osoba A nie jest chora na limbozę. Liczba krwinek beta we krwi osoby A nie jest podwyższona.	Zakładając, że dwa pierwsze zdania są prawdziwe, proszę ocenić prawdziwość zdania trzeciego jako wyprowadzonego z nich wniosku.		1. Tak 2. Nie 3. Nie wiadomo	3	tak
Sekcja	**Typ**	**Treść**	**Instrukcja**	**Liczba pozycji**	**Odpowiedzi**	**Odp. prawidłowa**	**Poziomo**

✔ Zatwierdź to badanie

Lista twoich badań

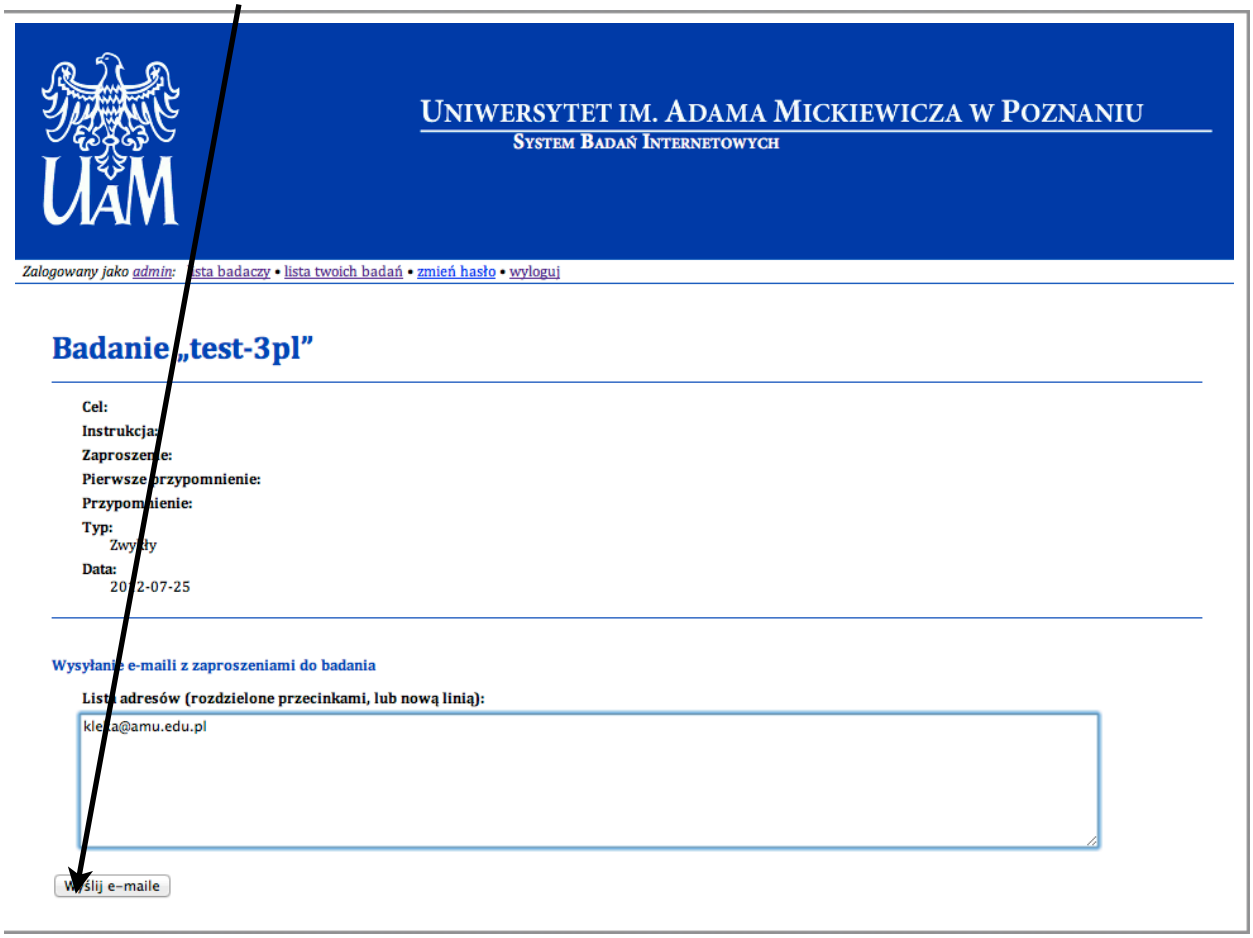
Można usunąć wprowadzone badanie, jeśli nie zostało zatwierdzone. Aby usunąć badanie klikamy Usuń

Można również edytować dane badanie, jeśli nie jest zatwierdzone. Klikamy w nazwę badania, które chcemy edytować. Na kolejnym ekranie klikamy przycisk Edytuj informacje o badaniu.

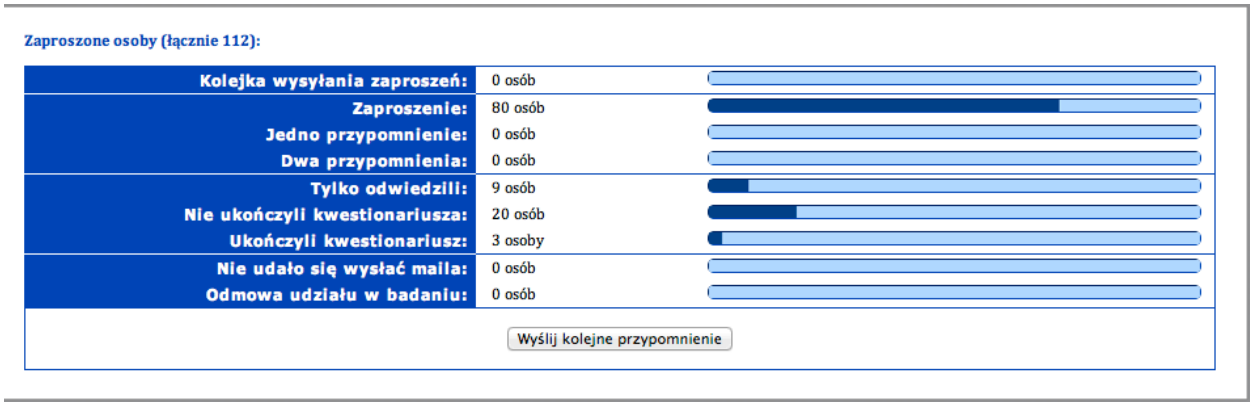
Po dokonaniu zmian naciskamy przycisk Zapisz.

Wprowadzanie adresów e-mail Badanych

Po zatwierdzeniu badania wprowadzamy listę adresów e-mail Badanych i naciskamy przycisk Wyślij e-maile.

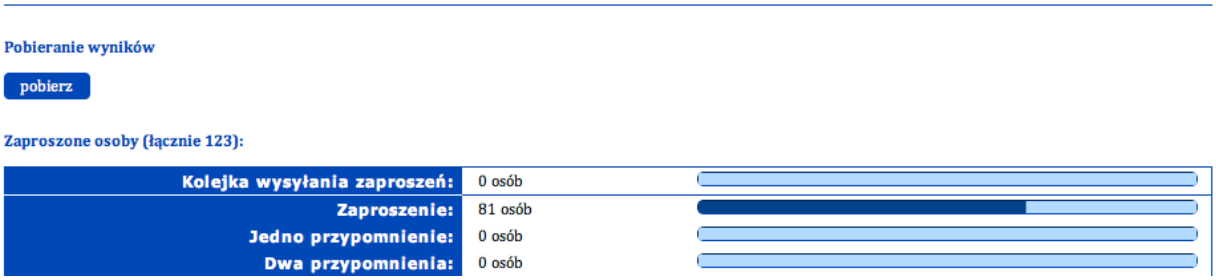


Po wysłaniu e-maili pojawi się tabelka ze statystykami badania. Przy każdym otworzeniu strony danego badania zostaną wyświetlone aktualne statystyki.

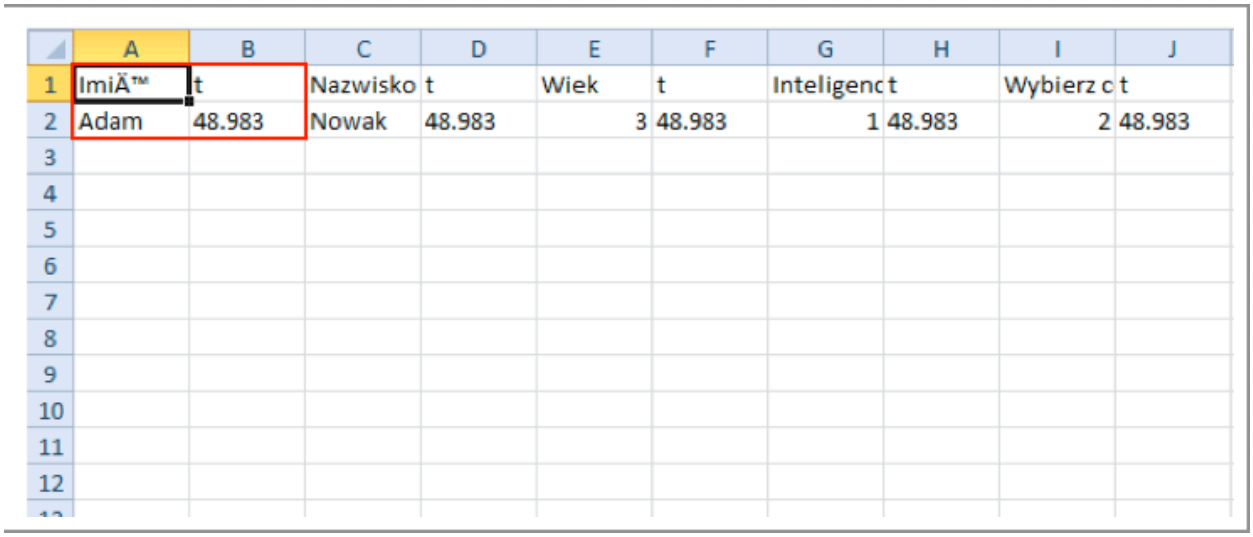


Pobieranie wyników

Jeśli przynajmniej jeden Badany ukończył udzielanie odpowiedzi na wszystkie pytania w badaniu, ponad tabelką ze statystykami pojawi się przycisk do pobrania wyników.



Klikamy przycisk pobierz i zapisujemy plik na dysku twardy komputera. W otrzymanym pliku CSV mamy w kolumnach kolejno treść pytania i czas odpowiadania na nie wyrażony w sekundach.



	A	B	C	D	E	F	G	H	I	J
1	ImiÄ™	t	Nazwisko	t	Wiek	t	Inteligenc	t	Wybierz c	t
2	Adam	48.983	Nowak	48.983	3	48.983	1	48.983	2	48.983
3										
4										
5										
6										
7										
8										
9										
10										
11										
12										

Zmiana hasła

Aby zmienić hasło naciskamy zmień hasło. Po wprowadzeniu zmian naciskamy Zapisz.