NEODIDAGMATA XXIII
POZNAŃ 1997

II. PRACE BADAWCZE

STANISŁAW UBERMANOWICZ
Uniwersytet im. Adama Mickiewicza
w Poznaniu

SPECYFIKA SKALI WAŻONYCH OCEN

ABSTRACT. Ubermanowicz Stanisław, *Specyfika Skali Ważonych Ocen* (Specificity of Weighted Rating Scale), „Neodidagmata" XXIII, Poznań 1997, Adam Mickiewicz University Press, pp. 63-81. ISBN 83-232-0848-4. ISSN 0077-653X.

The author describes the essence and characteristics of the novel surveying evaluation scale of the educational process. Balancing between quality, intensity and conclusiveness of expressions is a selective clue to choose indicators i.e. stimulus ascertainments to the survey. On the basis of empirical studies of "Informatics Culture" (computer experience and cognizance) the author discusses quantitative parameters of the scale and the way of representing them in practice. Special attention was paid to a contextual representation of results, to the vector method of the appraisal value of alteration in time; and to the issues of inter-group comparability of the process effects.

Stanisław Ubermanowicz, Zakład Technologii Kształcenia Wydziału Studiów Edukacyjnych, Uniwersytet im. Adama Mickiewicza, ul. Słowackiego 20, 60-823 Poznań, Polska-Poland.
e-mail: uberman@hum.amu.edu.pl

Dużą popularnością cieszą się wciąż narzędzia badawcze wzorowane na metodzie Likerta. Dobór bodźców, kategorie wypowiedzi i zasady interpretacji wyników stały się przedmiotem wielu analiz metodologicznych i publikacji. Przypomnę, że istotą tych narzędzi jest swoista zgodność semantyczna stwierdzeń, które wyłania się z wielu intuicyjnych na podstawie kryteriów empirycznych. W konsekwencji wzrasta pewność, że każdym bodźcem-stwierdzeniem badamy jedynie różne aspekty tego samego *indicatum*, dzięki czemu składowe cząstkowe wolno scalać w rzetelniejszy wynik ogólny. Cecha ta uzasadnia używaną nazwę – Skale Sumowanych Ocen. Jednakże zsumowany wynik ogólny to zbyt mało, stąd mój zamysł przedstawienia metody obliczeń komplementarnego pakietu parametrów i czynników, jakie warto wyznaczać i upowszechniać w efekcie badań za pomocą zmodyfikowanych skal ocen.

Oczywiście wzory i parametry liczbowe byłyby puste bez odniesienia do konkretnego obiektu pomiaru. Mówi się wprawdzie, że skale te mierzą postawy, lecz przecież postawa może być wskaźnikiem pośrednim czegoś szerszego, np. kultury osobistej. Spłyceniem jest opinia, że bada się jedynie deklaracje i zachowania respondentów wobec ankiety. Nie ulega wątpliwości, że reakcja na bodziec zależy od splotu wielu czynników, m.in. wewnętrznych predyspozycji i głębokości namysłu. Jednak jest też zawsze pochodną informacji niesionej przez bodziec, stąd właśnie starannie dobrana treść stwierdzeń wyznacza to, co mierzymy daną skalą. Możliwość pomiaru różnorodnych zmiennych cząstkowych ilustrują w niniejszym tekście kategorie składników i czynników wyznaczanych Kwestionariuszem Kultury Informatycznej[1]. Składniki i czynniki są więc traktowane jako przekroje przez zmienne, natomiast parametry – jako przekroje przez zdarzenia losowe (...jak zabawnie nazywają respondentów statystycy).

W efekcie poszukiwań powstał **model innowacyjnej skali** przydatnej w ewaluacji procesów edukacyjnych. Swoistymi wyróżnikami proponowanej metody są wielorakie mechanizmy bilansowania wskaźników oraz komponentów *indicatum*. Podstawową procedurą wyłaniania bodźców do tych skal jest technika mini-maksowa, polegająca na szukaniu optymalnego kompromisu między grupową zgodnością wypowiedzi a głębią uproblemowienia stwierdzeń. Pomaga w tym dwuwymiarowość skali i wskaźnik rozziewu. Niezbędnym atrybutem jest też rozwarstwienie zmiennej ogólnej na jej składowe. Dzięki temu można wykrywać składniki źle wypadające podczas pomiaru i optymalizować je, jeśli proces ten (np. nauczanie) trwa nadal. Podczas wnioskowania bilansuje się więc nie tylko parametry liczbowe, np. jakość z intensywnością, odchylenie dodatnie z ujemnym, wzrost ze spadkiem, ale też porównywane są pary kontrolne zmiennych cząstkowych, np. *Doznanie* i *Poznanie*. Innego rodzaju równoważenie odbywa się w trakcie analizy danych. Polega ono na rozwarstwieniu zasobów na charakterystyczne dla badanej zbiorowości podzbiory (np. klasy licealne o odmiennych profilach nauczania, grupy studentów z różnorodnych kierunków), a następnie na scaleniu ich w harmonijnie dobrane przekroje, rezentatywne dla danej populacji.

Wobec narzędzi badawczych, które spełniają omawiane tutaj kryteria, proponuję używanie nazwy Skale Ważonych Ocen. W artykule wyjaśniam istotę owego „ważenia", znaczenie parametrów ilościowych dwuwymiarowej skali oraz sposób ich interpretacji w praktyce. Wprowadzam czytelnika w nową metodologię i terminologię. Szczególną uwagę poświęcam kontekstowej reprezentacji wyników, a ponadto technice szacowania jakości zmian zachodzących w czasie oraz problematyce międzygrupowej porównywalności efektów procesu lub eksperymentu.

[1] S. Ubermanowicz, M. Paprzycki: *Między stylistyką a statystyką w teście kultury informatycznej*, „Neodidagmata" XXII, Wyd. Naukowe UAM, Poznań 1996.

DWUWYMIAROWOŚĆ SKALI

Charakteryzowane tu skale wyróżnia dokładna symetria kategorii wypowiedzi – skrajne opcje muszą być semantycznie przeciwstawne, a centrum neutralne. Jeśli zamierza się wyodrębnić owe trzy stany zmiennej ciągłej, to konieczne jest jej próbkowanie z gęstością przynajmniej dwukrotnie większą. Nieodzowne jest więc badanie pewnych obszarów krytycznych wypowiedzi w strefie wątpliwości „chyba" oraz granicznych, wytyczonych słowem „absolutnie". Na tych przesłankach opiera się zmodyfikowana (względem pierwotnej *strongly agree...*) i preferowana przeze mnie 7-punktowa kategoryzacja wypowiedzi z przypisanymi im wagami:

absolutnie nie	nie	chyba nie	brak zdania	chyba tak	tak	absolutnie tak
–3	**–2**	**–1**	**0**	**+1**	**+2**	**+3**

Oczywiście wagi takie nadaje się stwierdzeniom spolaryzowanych dodatnio, wobec których oczekujemy, że zostaną zaakceptowane. Natomiast stwierdzeniom spolaryzowanym ujemnie, których prawdziwość respondenci powinni zakwestionować, przypisuje się wagi o znakach przeciwnych. Nadawanie odpowiednich wag, a tym samym depolaryzacja danych, odbywa się w pierwszym etapie przetwarzania wyników surowych, które wygodniej jest wprowadzać do komputera w postaci całkowitych, dodatnich liczb porządkowych (rang). Proponowany ciąg wag {–3...+3} oddaje trafniej specyfikę badanych cech w Skalach Ważonych Ocen, dlatego zbiór wypowiedzi opisanych właśnie takimi wagami stanowić będzie zawsze zmienną losową X.

Z symetrii kategorii wypowiedzi bierze się intuicyjnie wyczuwana dwoistość skali. Oto najprościej jest nie mieć zdania, co w dużym uproszczeniu oznacza zerową intensywność procesów decyzyjnych. W miarę podejmowania coraz bardziej pewnych decyzji wzrasta stopień intensywności wypowiedzi. Dominują decyzje oparte na przesłankach racjonalnych. Przesłanki mogą być różne, stąd każdy ma prawo do własnego zdania i w konsekwencji oba skrajnie zdecydowane wybory w wymiarze intensywności są równocenne. Jednak demokratycznym odniesieniem wypowiedzi są ukształtowane w danym środowisku standardy ewaluatywne. Można łatwo stwierdzić, czy dana jednostka akceptuje wartości uznawane przez większość, a co cenniejsze – czy przypadkiem nie myli się w kwestiach elementarnych. W tych normowanych środowiskowo i w miarę oczywistych zagadnieniach daje się więc ocenić także jakość wypowiedzi. Każda wypowiedź respondenta ma bowiem równocześnie wymiar intensywności i jakości. Ową dwuwymiarowość najlepiej oddaje izofora: – *Głośno krzyczy, lecz czy ma rację?*

Przyjęty model kategorii wypowiedzi ma mocne strony. Umożliwia odnalezienie względnego poziomu zerowego jakości danej cechy, jako środka między strefą niezgodności <–3, 0> a strefą zgodności <0, +3> ze standardem

ewaluatywnym. Łatwo uzmysłowić sobie ruchomość tego zera, gdyby np. zmierzyć kwestionariuszem *przydatność kolei* w różnych państwach, zwłaszcza tam, gdzie kolei w ogóle nie ma. Pozornie zero intensywności wypowiedzi jest mocniej osadzone, lecz przecież „brak zdania" mógł wybrać ktoś, kto miał silne argumenty i za, i przeciw. Ponadto respondent mógł nie zakreślić żadnej odpowiedzi. Traktujmy więc oba zera równorzędnie i przyjmujmy poziomy wyznaczane względem nich jako parametry ruchome, zależne od kryteriów. Fakt względności punktów zerowych nie umniejsza ich roli w osadzaniu (zaczepianiu) wskaźników statystycznych.

Wyprzedzając podane w następnym rozdziale uzasadnienie dopuszczalności addytywnych działań na skalach ważonych, rozpocznę prezentację parametrów mających głęboki sens w praktycznej interpretacji wyników pomiaru. Otóż klasyczny moment zwykły wyznacza w naszej skali średnią **jakość** – *quality* [Q] wypowiedzi, natomiast mniej znany moment absolutny wyznacza jej średnią **intensywność** – *intensity* [I]:

$$Q = \frac{1}{n}\sum_{i=1}^{n}(X_i) \qquad I = \frac{1}{n}\sum_{i=1}^{n}(|X_i|) \tag{1.}$$

gdzie X_i – waga (lub średnia z wag) wypowiedzi i-tego respondenta
n – liczebność próby (klasy, grupy lub całej zbiorowości).

Parametry te oznaczają wartości przeciętne, przy czym mogą być one liczone zarówno dla zmiennej globalnej, jak i dla dowolnego splotu jej składników bądź czynników. Przy obliczaniu intensywności jedyną odmianą jest działanie na wartościach bezwzględnych z wag. Trzeba jednak podkreślić, że parametry te mają adekwatną interpretację tylko w odniesieniu do zmiennej losowej wyznaczanej wagami z ciągu {–3...+3} o symetrycznych modułach. Ponadto wszystkie wzory dotyczą wyników uogólnionych z próby, a nie osiągnięć indywidualnych osób.

Co z przypadkami wyboru wypowiedzi równocześnie „tak" i „nie"? Gdyby sprzeczne wypowiedzi pojawiały się częściej wobec danej kwestii, świadczyłoby to o błędnej konstrukcji bodźca-stwierdzenia. Jeśli natomiast takie zdarzenia występują sporadyczne, to wynikają one raczej z uproblemowienia i z wpływu wielu odmiennych czynników warunkujących podjęcie decyzji. W mojej praktyce na blisko 6 tysięcy wypełnionych ankiet zdarzyło się to zaledwie 4 razy. W takich przypadkach analiza dwuwymiarowa jest nieodzowna – przyjmuje się indywidualną jakość wypowiedzi [Q_i] za równą średniej z wag nadanych bipolarnie obu opcjom, a jej intynsywność [I_i] jako średnią z wartości bezwzględnych owych wag. Przykładowo, w używanej tu skali wag wybór łączny „tak" (+2) i „nie" (–2) daje $Q_i = 0$ oraz $I_i = 2$, stąd takie dane podstawia się do wzorów jako wyjątek dla i-tego elementu, zamiast (X_i) oraz ($|X_i|$).

Ważnym parametrem jest odchylenie standardowe – *standard deviation* [Sd], jednakże ten moment centralny rzędu drugiego nie jest tutaj zbyt przydatny. Liczy się go bowiem z całkowitej wariancji, a następnie umieszcza z obu stron wartości średniej, zakładając symetrię rozkładu. Tymczasem empiryczne rozkłady z próby, wynikające z pomiarów skalami typu Likerta, mają specyficzną cechę. Oto wskutek naturalnego w tej metodzie dążenia wypowiedzi w stronę wyników pozytywnych, najczęściej mamy do czynienia z rozkładem lewoskośnym. Wszak z tej strony względem średniej respondenci mają większą swobodę wyboru. W tej sytuacji obliczenie odchylenia standardowego i dodanie go do wartości średniej prowadziłoby do paradoksalnej sugestii, jakoby rozproszenie wyników sięgało poza kres górny skali.

W Skalach Ważonych Ocen proponuję używanie wskaźników bardziej adekwatnych do analizy dyspersji. Są to **odchylenia kierunkowe** – *directional deviations* [Dd], opisane przez dwa parametry, wyznaczające osobno dewiacje dodatnie i ujemne. Liczy się je odrębnie dla osiągnięć ponadprzeciętnych {i} oraz poniżej przeciętnych {j}, lecz dodatkowo z uwzględnieniem w obu liczebnościach {h; l} przypadków zgodności ze średnią (gdy $X_i = Q$). Odchylenie dodatnie [+d] obejmuje więc wszystkie wyniki nie gorsze od średniej, a odchylenie ujemne [–d] wszystkie nie lepsze od średniej:

$$+d = +\sqrt{\frac{1}{h}\sum_{i:\{X_i \geq Q\}}^{h}(X_i - Q)^2} \qquad -d = -\sqrt{\frac{1}{l}\sum_{j:\{X_j \leq Q\}}^{l}(X_j - Q)^2} \qquad (2.)$$

gdzie X_i; X_j – wagi (lub średnie z wag) wypowiedzi kolejnego respondenta
Q – średnia jakość wypowiedzi w próbie n-elementowej ze wzoru (1.)
h; l – liczebności podzbiorów {i} oraz {j} w próbie n-elementowej.

Odchylenia kierunkowe, jako pierwiastki z wariancji cząstkowych, mają tę właściwość, że ich średnia ważona, liczona według proporcji liczebności {h; l} do {n}, daje odchylenie standardowe. Są więc znacznie trafniejszą miarą cenionego w statystyce parametru, a ich obliczenie jest łatwe nawet w prostym arkuszu kalkulacyjnym, jeśli użyje się funkcji warunkowych.

Często zamiast „odchylenie" używa się nazwy „rozrzut". Nie jest to słuszne, gdyż odchylenie (dewiacja) wyraża pewną odległość od bazy, natomiast rozrzut (dyspersja) to pewien zakres rozpostarcia. Właśnie liczenie odstępu od wzorca jest istotą parametrów zwanych momentami centralnymi, oznaczanymi tutaj małymi literami ze znakiem + lub –. W Skalach Ważonych Ocen trzymajmy się zasady następującej: odchylenia kierunkowe wyznaczają standard odległości od średniej, natomiast rozrzut to standard zakresu rozproszenia. Standard oznacza tu swoiste uśrednienie, stąd rozrzut jest mniejszy od całkowitego rozstępu wyników. Zgodnie z powyższą zasadą **rozrzut** – *dispersion* [D] jest sumą modułów obu odchyleń kierunkowych:

$$D = |+d| + |-d| \quad (3.)$$

gdzie +d; -d – odchylenia kierunkowe obliczone ze wzorów (2.)

Parametr ten stanowi odniesienie do standaryzacji innych wskaźników, występuje wówczas w mianowniku wzorów. Z tego powodu trzeba przyjąć i sprawdzać to, że w żadnej badanej grupie, wobec żadnego bodźca-stwierdzenia nie ma całkowitej jednomyślności wypowiedzi, a więc rozrzut jest większy od zera.

Skoro asymetria rozkładu jest dość częstym zjawiskiem, to warto jeszcze wyznaczyć parametr charakteryzujący stopień tegoż odkształcenia. Zamiast klasycznego momentu centralnego rzędu trzeciego, dla omawianej tu Skali polecam znacznie łatwiejszą do obliczenia, standaryzowaną miarę **skośności** – *skewness* [±s], będącą ilorazem różnicy między modułami odchyleń kierunkowych względem ich sumy, czyli rozrzutu:

$$\pm s = (|+d| - |-d|) : D \quad (4.)$$

gdzie +d; -d – odchylenia kierunkowe obliczone ze wzorów (2.)
D > 0 – rozrzut obliczony ze wzoru (3.)

Zerowa wartość współczynnika skośności świadczy o symetrii rozkładu, dodatnia o asymetrii prawostronnej, a ujemna – o asymetrii lewostronnej. Znajomość tego parametru z zakresu <–1, +1> jest ważna w podejmowaniu decyzji, czy np. uprawnione jest oparcie testów istotności na rozkładzie normalnym, czy raczej trzeba sięgnąć po graniczne rozkłady statystyk.

Skoncentrujmy się teraz na kluczowej w Skalach Ważonych Ocen metodzie doboru bodźców-stwierdzeń. Jako uzupełnienie znanych technik wyłaniania stwierdzeń na mocy ich dyskryminacyjności i wewnętrznej trafności, proponuję dodatkowe kryterium sprawdzania ich konkluzywności (mocy wnioskotwórczej). Polega to na mini-maksowej technice balansowania między optymalnym rozrzutem [D] a minimalnym **rozziewem** – *bias* [–b] będącym różnicą jakości i intensywności wypowiedzi:

$$-b = Q - I \quad (5.)$$

gdzie Q – średnia jakość, I – średnia intensywność wypowiedzi z wzorów (1.)

Rozziew jest tu miarą niezgodności opinii jednostek z ogółem w kwestii konkretnego standardu ewaluatywnego i najlepiej, aby był zerowy. Rozrzut natomiast oznacza standardową rozbieżność zdań w tej samej sprawie i najlepiej, jeśli jest tak duży, że wypowiedzi wypełniają całą strefę zgodności <0, +3>. Wynika to z potrzeby uproblemowienia bodźców, gdyż szkoda wysiłku na pytania trywialne. Jednak zbytni wzrost rozrzutu zwiększa rozziew ponad dopuszczalne granice. Metody osiągania najlepszego kompromisu (mini-maksa) są żmudne. W praktyce trzeba wielokrotnie przeprowadzać proces optymalizacyjno-weryfikacyjny stwierdzeń, zmieniając

początkowo zdania, a później nawet pojedyncze słowa. Każda modyfikacja bodźca wpływa na wypowiedzi, a te są wskaźnikami badanej zmiennej. Decyzje w sprawie uznania bodźców-stwierdzeń za odpowiednie podejmuje się ostatecznie na podstawie łącznych kryteriów wartości semantycznej zdań na tle parametrów rozrzutu i rozziewu. Jako przykład do analizy przedstawiam wskaźniki empiryczne z pomiaru kwestionariuszem, który przeszedł już trzy pełne cykle optymalizacji i weryfikacji.

Tabela 1

Przykładowe parametry wskaźników wypowiedzi w pomiarze Kultury Informatycznej

Wskaźnik	kod	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Jakość	Q	1,7	1,3	1,1	1,2	1,2	1,7	0,7	0,8	1,8	0,9	1,2	1,8	0,5	1,6	2,0	1,6	1,4	1,2	1,1	1,7	1,3	0,9	0,7	1,6
Rozrzut	D	2,8	2,4	2,0	2,9	2,2	2,7	2,9	3,0	2,6	3,3	2,1	2,4	3,3	2,6	2,8	2,3	2,3	2,9	2,5	2,3	2,6	2,9	3,1	2,8
Intensywn.	I	1,9	1,5	1,3	1,7	1,4	1,9	1,4	1,5	2,0	1,6	1,5	1,9	1,5	1,8	2,1	1,7	1,6	1,6	1,5	1,8	1,6	1,4	1,4	1,9
Rozziew	I–b I	0,2	0,2	0,2	0,5	0,2	0,2	0,7	0,7	0,2	0,7	0,3	0,1	1,0	0,2	0,1	0,1	0,2	0,4	0,4	0,1	0,3	0,5	0,7	0,3

Zauważmy (tabela 1), że rozrzuty z próby n = 534 wypowiedzi wobec 24 stwierdzeń kwestionariusza wynoszą D = 2,0 ÷ 3,3. Teoretycznie w tych zakresach lokuje się ok. 68% ogółu, tj. ponad 2/3 respondentów. Odnosząc te parametry do całej skali, która ma tylko 3 przedziały strefy zgodności, można stwierdzić, że dyspersja minimalna o rozmiarze 2 działek jest wystarczająca. Zresztą bodziec nr 3 brzmi: – *Umiem poradzić sobie z komputerem*, a parametry dotyczą pomiaru na końcu zajęć z informatyki, więc jedynie dziwić może niski poziom jakości Q. Zrozumiałe, że rozproszenie wyników z początku zajęć bezsensownie nakazywałoby odrzucenie tego bodźca. Konkluzywność jest większa w miarę wzrostu rozrzutu, lecz optymalnie do 3 działek. Kilka wskaźników przekroczyło tę granicę. Bodziec nr 10 brzmi: – *Jest mi źle w grupie, gdy inni lepiej znają komputery*. W tej formie stwierdzamy niepokojące fakty z zajęć grupowych, więc nie warto zabiegać o zgodność wypowiedzi, tym bardziej, że komplementarny bodziec nr 9 ujawnia spokój przy pracy indywidualnej. Jest to przykład kontrolnej pary wskaźników. Bodziec nr 13 brzmi: – *Komputery powinny być używane na wszystkich zajęciach.* Ma on nadmierny rozziew, dlatego podjąłem decyzję o zmianie treści na: – *Komputery powinny być używane tylko na zajęciach z informatyki*, a empiria zweryfikuje pomysł. Bodziec nr 23 brzmi: – *Pewnego dnia komputery zniewolą ludzi.* Zdumiewające, jak mało optymizmu, lecz trzeba zauważyć, że jest to ekstraspekcja i ów wgląd w innych świadczy o zatroskaniu manią komputerową. Pozostawiam bez zmian ten wskaźnik, gdyż komplementarny, introspekcyjny bodziec nr 24 wykazuje osobistą odporność na uzależnienie się. Wysoką wartość rozziewu (0,7) mają jeszcze bodźce nr 7 i 8 – oba próbkują uzdolnienia. Fakt, że studenci tak nisko oceniają swe możliwości nie skłoni

mnie jednak do zmiany treści stwierdzeń. Zamierzam natomiast sprawdzić empirycznie, jak wpłynie dodanie słówka *Zbyt...* na początku bodźca nr 22: *...długotrwała praca przy komputerze jest szkodliwa.*

PRZEJRZYSTOŚĆ WYNIKÓW

Nagminnie ferowane w sprawozdaniach z badań rozkłady procentowe traktować należy raczej jako publikację danych, a nie jako wyniki. Podanie informacji tylko o tym, ile procent badanych wybrało jaką opcję jest zrzuceniem na czytelnika obowiązku przetwarzania danych. Aby nie poprzestać na krytyce, przedstawiam propozycję prostego zabiegu ułatwiającego interpretację wyników pomiaru Skalą Ważonych Ocen. Chodzi o odniesienie parametrów do powszechnego schematu poznawczego, jakim jest zakorzeniony w nas skrypt ewaluatywny tradycyjnych ocen szkolnych {2, 3, 4, 5}. Kategorie tej starszej skali {ndst, dst, db, bdb} są bardziej jednorodne w porównaniu z nowszą skalą, lepiej pasują liczbowo w kontekście progu wymagań na zaliczenie w szkolnictwie wyższym i dlatego są tam nadal stosowane.

Przyjrzyjmy się uważnie przedstawionej tu propozycji konwersji opcji wypowiedzi ze Skali Ważonych Ocen na kategorie i wartości liczbowe stopni szkolnych:

chyba nie	–	brak zdania	–	chyba tak	–	tak	–	absolutnie tak
... niedostateczny		dostateczny		dobry		bardzo dobry		
...1,5 {2! 2 2+}		2,5 {3– 3 3+}		3,5 {4– 4 4+}		4,5 {5– 5 5!}		5,5

Zauważmy, że próbki wypowiedzi są zmienną dyskretną, kategorie stopni wytyczają rozłączne przedziały osiągniętych poziomów, natomiast reprezentacje liczbowo-symboliczne tworzą dodatkowe podprzedziały, by wreszcie wskutek liczenia średnich przejść w zmienną ciągłą. Takie zjawisko jest typowe w procesie odtwarzania z próbek cechy pierwotnie ciągłej.

Pojawienie się ocen 2! i 5! nie zdziwi tych, którzy dawniej spotykali się z intuicją nauczycieli podpowiadającą, że zdarzają się przypadki nie mieszczące się w standardach edukacyjnych (stąd dzisiejsze 1 i 6). Podobnie w wynikach badań może się zdarzyć, że pojedynczy wskaźnik wypadnie poza kresem ocen. Przy normalizowanym narzędziu jest to rzadkie, toteż dla tych wyjątków nie warto rozciągać skali kosztem wyrazistości wyników. Warto natomiast zgodzić się na to, aby każdy z przedziałów ocen był jednakowej rozpiętości plus/minus pół stopnia. W ten sposób powstaje teoretyczny, a wykorzystywany do obliczeń efektów względnych, absolutny kres górny osiągnięć {5,5}.

Węzłem gordyjskim wypowiedzi w kwestionariuszu jest „brak zdania". Odpowiada to sytuacji szkolnej, gdy musimy podjąć trudną decyzję: zaliczyć

czy nie (średnia 2,5). Dowolne odchylenie skali w lewo oznacza poziom niedostateczny i jest wyrazem rozbieżności między wartościami uznawanymi przez jednostkę wobec standardów ewaluatywnych całej grupy. Kolejny punkt krytyczny to strefa wątpliwości „chyba tak" (średnia 3,5), która wytycza dokładnie środek czterostopniowej skali ocen szkolnych i jest progiem przejścia na poziom dobry. Graniczna jest także sama wypowiedź „tak" (średnia 4,5), rozdzielająca poziom dobry od bardzo dobrego. Zdobycie najlepszej oceny jest więc uwarunkowane pewnością siebie w zwrocie „absolutnie tak".

Konwersja na oceny szkolne przydaje się w praktyce, lecz czy jest uprawniona teoretycznie? Oznacza przecież dodanie do każdego z momentów pewnej wartości stałej (*const* = 2,5). Skoro operujemy tutaj wagami i momentami, to trzeba określić, z jakim typem skal mamy do czynienia. Śmiem twierdzić, że w publikacjach klasyfikujących skale pomija się jedną z ważniejszych, odrębnych kategorii. Nazwijmy ją „skalą ważoną" i zdefiniujmy jej istotę. Ze **skalą ważoną** mamy do czynienia wówczas, jeśli jedna z kategorii próbkujących wyznacza zerową wartość badanej cechy, a pozostałe próbki wskazują na monotoniczność, choć nie da się określić podziałki jednostkowej. Ponadto w każdej z głównych działek skali, wytyczonych gęstością próbkowania, daje się empirycznie wyodrębnić co najmniej dwa sąsiednie podprzedziały, wobec których test potwierdza istotność różnic estymat i jednokierunkowość przyrostów, bez możliwości wykazania, że przyrosty badanej cechy są jednakowe.

Właściwości takie – zerową wagę i monotoniczne przedziały – posiada Skala Ważonych Ocen i dzięki nim można szacować estymatory, wyznaczać poziomy oraz przesuwać je o wartość stałą, nawet jeśli badana cecha ma charakterystykę krzywoliniową. Przesunięcie oznacza jedynie zmianę kryteriów oceniania. Dopuszcza się także obliczanie średniej ruchomej, lecz nie wolno twierdzić, że ocena dobra {4} jest dwa razy lepsza od niedostatecznej {2}. Reasumując: konwersja obejmuje przejście z 7-punktowej skali wypowiedzi na 6-przedziałową skalę wartości o środkach 0, 1, 2, 3, 4, 5 i rozpiętości działek ±0,5, z czego do reprezentacji wyników uogólnionych wykorzystuje się cztery górne przedziały w kontekście tradycyjnych poziomów osiągnięć szkolnych {ndst, dst, db, bdb}.

W celu zapewnienia całkowitej jednoznaczności przy prezentacji wyników proponuję odmienne nazewnictwo i oznaczenia dla parametrów przed i po konwersji. Jakość i intensywność po przesunięciu przyjmują adekwatnie nazwy **wartość** – *value* [V] oraz **ważkość** – *gravity* [G]:

$$V = 2{,}5 + Q \qquad G = 2{,}5 + I \tag{6.}$$

gdzie Q – średnia jakość, I – średnia intensywność wypowiedzi z wzorów (1.)

Konwersja w praktyce jest dodaniem stałej 2,5 do uprzednio wyliczonych średnich, stąd ważkość osiąga poziomy z zakresu <2,5; 5,5>. Wartość wypowiedzi teoretycznie ma zakres <–0,5; +5,5>, lecz w praktyce pomiaru zoptymalizowanym kwestionariuszem cząstkowe zmienne prawie zawsze mieszczą się w przedziale <2; 5>. Parametr V jest bazowym wyznacznikiem poziomu badanego standardu ewaluatywnego [M], oznaczającego *wartość przeciętną indicatum z próby.*

Empirycznym wskaźnikiem najbardziej zbliżonym do teoretycznej estymaty (*wartości oczekiwanej indicatum dla populacji*) jest **norma** – *norm* [N], o ile oczywiście dysponujemy reprezentatywnym i obszernym zasobem danych. Wbrew pejoratywnym skojarzeniom nazwy, nie jest to jakiś narzucony pułap do osiągnięcia, lecz specyficzna baza normalizacji narzędzia, którą warto umieszczać w nagłówkach tabel z wynikami. Parametr ten oblicza się jako przeciętną ze średnich wartości osiągniętych przez poszczególne podgrupy w pomiarze końcowym:

$$N = \frac{1}{k} \sum_{i=1}^{k} (V_i'') \qquad (7.)$$

gdzie V_i'' – średnia wartość w i-tej klasie (grupie) w pomiarze końcowym
k – ilość wydrębnionych klas lub grup.

Istotną cechą normy w porównaniu ze średnią wartością ogólną całej próby badawczej jest to, że niweluje ona nieco wpływ zróżnicowanych liczebności w poszczególnych podgrupach. Przykładowo: chociaż dysponuję wynikami wieluset studentów pedagogiki, a z innych kierunków (w tym technicznych) otrzymuję zwykle reprezentacje mniejsze niż stuosobowe, to jednak sposób obliczania normy powoduje, że nie jest ona zdominowana przez osiągnięcia tej najliczniejszej grupy.

Jak w owej przestrzeni intuicyjnie wyczuwanych liczbowych wartości ocen rozpościera się zróżnicowanie wypowiedzi i rozkład gęstości wyników? Oczywiście dokładnie tak samo, jak przed konwersją, gdyż odchylenia są momentami centralnymi, liczonymi wokół średniej. Wyznaczmy wobec tego trzy łatwo interpretowalne poziomy wartości: **wysoki** – *high* [H], **średni** – *mid* [M] i **niski** – *low* [L], wykorzystując zdefiniowane wcześniej odchylenia kierunkowe i wartość V:

$$H = V + |+d| \qquad M = V \qquad L = V - |-d| \qquad (8.)$$

gdzie +d; –d – odchylenia kierunkowe obliczone ze wzorów (2.)
V – średnia wartość wypowiedzi z wzoru (6.)

Teoretycznie w bliskiej odległości wokół średniej M mieszczą się wyniki 1/3 respondentów. W dwu pozostałych, szerszych, asymetrycznych otoczeniach poziomów wysokiego H i niskiego L także po 1/3. Inaczej ujmując

zjawisko rozkładu gęstości – w zakresie rozrzutu D mieści się ok. 2/3 respondentów, a poza nim po ok. 1/6 z każdej strony. Jest wysoce prawdopodobne, że w klasie 30-osobowej wynik poniżej poziomu L uzyska 5 uczniów. Poziomy HML mogą więc być osiami lub kresami podziałów na podgrupy. Ze względu na duże znaczenie interpretacyjne, w publikacjach wyników pomiaru Skalami Ważonych Ocen starajmy się zamieszczać wartości wszystkich trzech poziomów, chyba że rozkład jest symetryczny i wystarczy podać odchylenie standardowe.

Tabela 2

Przykładowe parametry składników zmiennej w pomiarze Kultury Informatycznej

Składniki	Kod	Ocena	Aplauz	Ambicje	Intencje	Spokój	Odpór	Osąd	Wgląd	Obycie	Pewność	Zdolność	Ogłada
Norma	N	4,2	4,3	4,1	4,1	3,9	3,7	3,7	4,0	3,9	3,9	3,5	3,6
V+d	H	5,0	4,7	4,5	4,8	4,6	4,5	4,8	4,6	4,5	4,5	4,1	4,5
Wartość	M	4,4	4,3	4,0	4,1	3,7	3,6	4,1	3,7	3,6	3,5	3,3	3,5
V–d	L	3,2	3,5	2,9	3,3	2,8	1,8	3,3	2,9	2,5	2,6	2,3	2,7
Skośność	±c	–0,33	–0,33	–0,38	–0,07	0,00	–0,33	–0,07	+0,06	–0,10	+0,05	–0,11	+0,11

Zauważmy (tabela 2), jak łatwo jest po konwersji na poziomy ocen szkolnych wnioskować o tym, które składniki procesu winien nauczyciel poprawić. Mimo wysokiej *Oceny* odbytych zajęć, mimo pozytywnego *Osądu* komputera, inne parametry wymagają jeszcze sporego wysiłku. Zwłaszcza *Obycie*, *Pewność siebie* i *Zdolności* są stanowczo zbyt nikłe. Słabość komponentu poznawczego ujawnia także *Wgląd* w realne możliwości komputera, wobec zbyt optymistycznego względem normy, a osadzonego raczej na wiedzy potocznej *Osądu*. Nad emocjami też warto pracować, gdyż poziom L = 1,8 *Odporu* na uzależnienie się jest alarmujący. Oczywiście porównanie wyników innych klas tegoż nauczyciela pozwala dopiero stwierdzić, czy nie jest to wyłącznie specyfika uczniów. Przy okazji zwróćmy uwagę na znaczną skośność rozkładów empirycznych składników. Zmienna globalna *Kultura* ma jednak zawsze o wiele lepszą symetrię niż jej komponenty.

Pozostaje jeszcze kwestia **rozdzielczości skali**, którą wyznaczają podprzedziały ufności liczone dla każdej z próbkowanych działek. Jest to cecha płynna, której nie da się oderwać od wyników empirycznych, uzyskanych konkretnym narzędziem. Okazuje się, że rozdzielczość jest nierównomierna wzdłuż skali, przy czym im wyniki są lepsze, tym większą ufnością można je obdarzyć. Zależy ona także od spójności badanych grup. W praktyce pomiaru Kultury Informatycznej graniczne przedziały ufności wyników ogólnych wahały się znacznie: od ±0,04 dla studentów kierunku Informatyka, aż do ±0,23 w grupie o największym zróżnicowaniu doświadczeń komputerowych. Z liczb tych wynika, że stosowane w tabelach precyzje zaokrągleń mają jedynie znaczenie orientacyjne. Niemniej empiria wykazuje, że dla wię-

kszości wyników rozdzielczość (podziałka skali) może być co najmniej trzykrotnie lepsza niż jeden cały stopień (działka jednostkowa), co potwierdza trafność używania w szkołach ocen z plusami i minusami.

WARTOŚCIOWOŚĆ PRZEMIAN

Ruchomość parametrów powoduje, że tak naprawdę w pełni cenny jest dopiero pomiar różnicowy. Oznacza to konieczność dwukrotnego pomiaru tym samym narzędziem. Z wielu typów eksperymentów wybierzmy tylko dwa przykłady oparte na analizie czynników modyfikujących wtórną wypowiedź tego samego respondenta. Popularną techniką jest manipulacja wysublimowaną zmienną interweniującą, jeśli więc wykryliśmy różnicę, to zapewne pod wpływem owej zmiennej. Z reguły jednak mamy splot nie dających się wyizolować czynników. Pozostaje zatem możliwość użycia takich bodźców-stwierdzeń, aby wypowiedzi były wskaźnikami czynników, a pośrednio także zmiennej globalnej. W eksperymencie można stymulować zmiany w wypowiedziach poprzez modyfikację wybranych cech badanego obiektu, uzyskując komplementarne pary wyników niemal natychmiast. W wyniku analizy otrzymujemy wówczas informację o tym, jaka z cech była wyżej oceniana. Tutaj zajmiemy się szczegółowo drugim typem pomiaru. Chodzi o obserwację i ocenę procesu przemian zachodzących w czasie. Dopiero po zakończeniu procesu dowiadujemy się, na ile wartościowe były zmiany poszczególnych czynników lub składników *indicatum*. Choć obie techniki są podobne, to jednak do tej pierwszej trzeba używać innego nazewnictwa parametrów (o tym nieco w ostatnim rozdziale).

Proces przemian jest pewną klasą abstrakcji. Nie daje się wyrazić parametrem, natomiast można z niego wyodrębnić opisujący zjawisko zasób parametrów powiązanych. Z tego względu **zmiana** – *alteration* [±A] w ujęciu parametrycznym jest trójką zaczepionych wektorów tendencji do przemian w czasie, zachodzących na każdym z trzech poziomów wartości:

$$\pm A = \{\pm H, \pm M, \pm L\} \qquad (9.)$$

gdzie $\pm H = H'' - H'$ $\pm M = M'' - M'$ $\pm L = L'' - L'$ ze wzoru (8.)

Przyjmijmy konwencję oznaczania parametrów obu pomiarów indeksami „prim" i „bis" [M'; M''], natomiast ich różnice znakiem „plus/minus" [±M]. W praktyce zmiana jest zbiorem różnic między parametrami HML w pomiarze końcowym wobec pomiaru początkowego, odnoszonych jednakże zawsze podczas wnioskowania do któregoś z poziomów momentów zwykłych, np. do średniej własnych osiągnięć końcowych danej grupy [M''] lub do normy [N], a nade wszystko – wzajemnie do siebie. Wyjaśnijmy od razu, dlaczego potrzeba łącznej analizy wskaźników zmian jest tak bardzo ważna. Otóż ich wartości liczbowe określają jedynie szacunkowo nasilenie zjawisk,

natomiast stopień zrównoważenia trójki wektorów precyzyjnie wyznacza trafność badanego procesu.

W badaniu statystycznej zależności między dwoma pomiarami posłużmy się parametrem podobnym do współczynnika korelacji. Odróżnia go jednak swoiste uporządkowanie, polegające na przyrównywaniu zawsze obu wyników tej samej osoby (stąd prefiks *auto-*). Dla współczynnika autokorelacji przyjmijmy tu krótką nazwę **korelat** – *correlate* [±c] i obliczajmy go jako nieobciążony liczebnościami iloraz kowariancji wypowiedzi tych samych respondentów w stosunku do średniej geometrycznej z wariancji początkowej i końcowej:

$$\pm c = \sum_{i=1}^{n} \{ (X_i' - Q') (X_i'' - Q'') \} : \sqrt{\sum_{i=1}^{n} (X_i' - Q')^2 \cdot \sum_{i=1}^{n} (X_i'' - Q'')^2} \qquad (10.)$$

gdzie $X_i'; X_i''$ – wagi (lub średnie z wag) wypowiedzi i-tego respondenta w obu pomiarach

$Q'; Q''$ – średnie (dla całej próby) jakości wypowiedzi w pierwszym i drugim pomiarze

n – liczebność próby (klasy, grupy lub całej zbiorowości)

Autokorelacja przyjmuje wartości z zakresu <–1, +1> a jej interpretacja jest wyrazistsza niż korelacji. Dodatnie wartości wskazują, w jakim stopniu ci sami respondenci, którzy na wejściu osiągnęli wynik powyżej/poniżej średniej osiągnęli adekwatnie lepszy/gorszy wynik na wyjściu. Ujemne wartości wskazują, że role się odmieniły, tj. respondenci zamienili się miejscami w rankingu wyników końcowych. Korelat bliski zeru świadczy o braku związku między osiągnięciami w obu pomiarach, lecz ponadto może wystąpić w przypadku bardzo dużych przyrostów wyników, czego nie można interpretować jako niekorzystnego przemieszania.

Do wykrywania subtelnych przesunięć z jednej do drugiej połówki Skali Ważonych Ocen, a więc do badania zmian wewnętrznej zgodności w grupach, wykorzystajmy charakterystyczną dla skali metodę bilansowania momentów zwykłych z absolutnymi. **Bilans** – *balance* [±b] oznaczać tu będzie różnicę między modułami rozziewu w pierwszym i drugim pomiarze:

$$\pm b = | -b' | - | -b'' | \qquad (11.)$$

gdzie –b' oraz –b'' – rozziewy w pomiarze początkowym i końcowym ze wzoru (5.)

Parametr ten interpretuje się podobnie jak bilans w księgowości. Jego zerowa wartość oznacza zrównoważenie procesu i nie ma co oczekiwać, że będzie dodatni. „Superata" jest możliwa jedynie wówczas, gdy początkowo w słabszej grupie był rozziew, który udało się zniwelować. Wartość ujemna bilansu świadczy natomiast o niekorzystnym zjawisku rozwarstwienia grupy, w której część respondentów przesunęła się do strefy niezgodności z opinią większości. W odróżnieniu od parametrów uśrednionych, bilans jest bardzo czułym wskaźnikiem pojedynczych przypadków.

Mimo zastrzeżeń co do wartości zbiorczego parametru charakteryzującego zmiany, wyznaczmy pewien zgrubny wskaźnik, przyjmując **trend** – *trend* [±T] jako wektor tendencji do przemian uśredniony z wszystkich trzech poziomów wartości:

$$\pm T = \left\{(\pm H) + (\pm M) + (\pm L)\right\} \frac{1}{3} \qquad (12.)$$

gdzie ±H = H'' – H' ±M = M'' – M' ±L = L'' – L' ze wzoru (8.)

Dopuszczalność dodawania różnic z poziomów HML bierze się stąd, że dystrybuanty rozkładu normalnego właśnie w zakresie przedziałów wyznaczonych odchyleniami kierunkowymi <–d, +d> mają charakterystyki w miarę liniowe. Trend jest uzupełniającym wskaźnikiem jakości przemian wespół z różnicą poziomów średnich [±M]. W zasadzie największą korzyść przynosi wyznaczenie zwrotu wektora trendu, a więc informacja o dodatniej bądź ujemnej jego wartości.

Tabela 3

Przykładowe parametry czynników i wyniku ogólnego pomiaru Kultury Informatycznej

Czynnik	Kod	Opinie	Poglądy	Motywacje	Wprawa	Emocje	Ogłada	Doznanie	Poznanie	Kultura
X'' – X'	±H	0,00	–0,02	–0,04	0,25	–0,16	0,05	–0,09	–0,01	–0,03
Zmiana	±M	0,20	0,13	–0,08	0,55	–0,08	0,12	0,01	0,27	0,14
	±L	0,20	0,38	–0,13	0,65	–0,11	0,11	0,06	0,31	0,27
Korelat	±c	0,30	0,35	0,38	0,47	0,49	0,13	0,53	0,47	0,53
Bilans	±b	0,01	0,09	0,00	0,29	–0,04	0,01	0,00	0,00	0,00
Trend	±T	0,13	0,16	–0,08	0,49	–0,12	0,09	0,00	0,19	0,13

Zauważmy (tabela 3), że z różnic rozpatrywanych w oderwaniu od poziomów bezwzględnych można wnioskować jedynie o zjawiskach lokalnych – długości wektorów mają znaczenie umowne, lecz znaki jednoznacznie określają tendencje. W przypadku tej konkretnej klasy uczniowie nabrali *Wprawy*, a jednak u wszystkich obniżyły się *Motywacje* i *Emocje*. Zjawiska są niekorzystne na poziomie H. Splot czynników spowodował, że zarówno *Doznanie*, jak i *Poznanie* uczniów H było ujemne i w konsekwencji spadła zmienna globalna – *Kultura*. Można domniemywać po wzroście *Poznania* uczniów L, że program zajęć był dopasowany do tego niższego poziomu, stąd aspiracje lepszych nie zostały zaspokojone. Potwierdza to wzrost *Opinii* o kursie i komputerach tylko u uczniów z poziomów ML. Gdyby program zajęć był zbyt trudny, to parametry poziomów HML odmieniłyby się diametralnie, natomiast trafnie dobrany poziom zajęć zrównoważyłby wektory. Współczynniki autokorelacji są w tej klasie niższe od typowych, co świadczy o wewnętrznych roszadach. Warto zwrócić uwagę na zerowe wartości bilansu w dwóch najbardziej uogólnionych czynnikach i w zmiennej globalnej. Jest to typowe zjawisko wygładzania, towarzyszące uśrednianiu wyniku całkowitego.

PORÓWNYWALNOŚĆ EFEKTÓW

Cenna byłaby możliwość przyrównywania osiągnięć różnych klas czy grup. Problem w tym, że np. przejście z poziomu 3 na 4 nie jest równoważne przejściu z 4 na 5. Blisko jest przecież kres owego wzrostu. Poprzednie parametry wystarczają tylko do porównań lokalnych, gdyż oparte są na bazie ruchomych poziomów. Wnioskowanie ogólniejsze jest jednak możliwe dzięki metodzie względnego odniesienia wektorów rzeczywistych do wektorów potencjalnie absolutnych. Kolejne proponowane parametry oparte są na podejściu różnicowo-proporcjonalnym od samego początku obliczeń.

Pojęciem kluczowym oderwania się od wszelkich ruchomych poziomów jest fluktuacja. Nie da się jej wyrazić jednym parametrem, lecz można ją opisać pewnym zasobem wektorów. W ogólnym ujęciu **fluktuacja** – *fluctuation* [±F] jest tutaj indeksem swobodnych wektorów wyznaczających wszystkie indywidualne tendencje do zmian wypowiedzi w czasie:

$$\pm F = \{X_i'' - X_i'\} \qquad (13.)$$

gdzie X_i'' – waga (lub średnia z wag) wypowiedzi i-tego respondenta w drugim pomiarze

X_i' – waga (lub średnia z wag) wypowiedzi i-tego respondenta w pierwszym pomiarze.

W praktyce oznacza to zbiór różnic między wypowiedziami w pomiarze końcowym wobec pomiaru początkowego, liczonych osobno dla każdego z respondentów, bez odnoszenia do momentów zwykłych. Porównajmy to ujęcie z definicją zmiany. Otóż przy wyznaczaniu *Zmiany* różnice między dwoma pomiarami liczone są z wcześniej uśrednionych poziomów grupowych, natomiast *Fluktuacja* jest tu jedynie zasobem różnic indywidualnych, z których to pierwotnych elementów dopiero w ujęciu względnym oblicza się współczynniki standardowe wzrostu i spadku.

Łączny wektor fluktuacji nie ma zaczepienia do jakiegoś bezwzględnego poziomu ocen. Orientacyjnym punktem odniesienia mogłaby być hipotetyczna oś fluktuacji, będąca poziomem osiągniętym prawdopodobnie gdzieś w połowie okresu badań. Do precyzyjnych obliczeń efektów jakiegoś procesu lub eksperymentu niezbędne jest jednak odniesienie do wektora absolutnego. Najlepszą bazą okazuje się zbiór możliwych maksymalnych osiągnięć. W proponowanej metodzie następuje przejście od zbioru wektorów swobodnych do pary wektorów kierunkowych, mających cechy współczynników standardowych i będących podstawą wskaźników uogólniających.

Ustalmy, że w Skalach Ważonych Ocen nasilenie fluktuacji opisują dwa wektory kierunkowe: **wzrost** – *increase* [+e] i **spadek** – *decrease* [–e], wyznaczające osobno względne tendencje dodatnie i ujemne. Liczy się je odrębnie dla zysków {i} oraz strat {j}, lecz z uwzględnieniem w obu liczebnościach {h; l} przypadków braku zmian (gdy X'' = X'). Wzrost obejmuje więc wszystkie

wyniki nie gorsze, a spadek wszystkie nie lepsze od stanu początkowego, liczone w proporcji do zakresu możliwych maksymalnych zmian:

$$+e = \sum_{i=1}^{h} (X_i'' - X_i') : \sum_{i=1}^{h} (X_{max} - X_i') \qquad -e = \sum_{j=1}^{l} (X_j'' - X_j') : \sum_{j=1}^{l} (X_{max} - X_j'') \quad (14.)$$

dla i: $\{X_i'' \geq X_i'\}$ dla j: $\{X_j'' \leq X_j'\}$

gdzie $X_i''; X_j''$ – waga (lub średnia z wag) wypowiedzi respondenta w drugim pomiarze
$X_i'; X_j'$ – waga (lub średnia z wag) wypowiedzi respondenta w pierwszym pomiarze
X_{max} – najwyższa waga przypisana do kresu górnego skali, tutaj $x_{max} = +3$
h; l – liczebności podzbiorów {i} oraz {j} w próbie n-elementowej.

Obydwa parametry są standardowymi amplitudami zachodzących przemian, przy czym wzrost jest praktyczną miarą oddziaływania czynników pozytywnych, a spadek – wskaźnikiem występowania zjawisk negatywnych. Jako współczynniki względne mogą być wyrażane w procentach. W praktyce zdarzają się takie sukcesy, że w danej klasie spadek jest zerowy, lecz tylko w pojedynczej zmiennej cząstkowej. Charakterystyczne jest bowiem to, że identyczne oddziaływania są akceptowane przez jednych, a odrzucane przez innych. Ponadto sposób obliczania tych parametrów powoduje, że są one czułe nawet na odosobnione przypadki tendencji niekorzystnych.

Przez analogię do znaczenia, jakie spełniają odchylenia kierunkowe w analizie pomiaru jednokrotnego, do wymiarowania zmian mierzonych testem różnicowym ważne są kierunkowe wektory tendencji. Suma modułów wzrostu i spadku wyznacza **zasięg** – *extent* [E] fluktuacji:

$$E = |+e| + |-e| \quad (15.)$$

Parametr ten określa standardowy zakres amplitud zachodzących przemian i jest w badanym procesie praktyczną miarą intensywności oddziaływań splotu wszystkich czynników, zarówno sprzyjających, jak i niekorzystnych. W pomiarze dydaktycznym wielkość zasięgu nosi wyraźne piętno siły wpływu nauczyciela.

Najbardziej wartościowym wskaźnikiem uogólniającym wyniki dwukrotnego pomiaru za pomocą Skali Ważonych Ocen jest **efekt** – *effect* [±E], obliczany z różnicy między modułem wzrostu a modułem spadku:

$$\pm E = |+e| - |-e| \quad (16.)$$

Ten bardzo czuły parametr wyznacza kierunek i standardową wartość tendencji przemian w badanym procesie. Jest praktyczną miarą jakości oddziaływania splotu wszystkich czynników, ze wskazaniem tych, które dominowały (zwykle pozytywne). W pomiarze dydaktycznym wskaźnik *Efektu* jest wypadkową akceptacji bądź odrzucania zarówno treści, jak i metod stoso-

wanych przez nauczyciela. Jako współczynnik względny z zakresu <–1, +1> może służyć do porównywania osiągnięć między dowolnymi grupami i populacjami, w przeciwieństwie do wskaźnika *Trendu*. W pełni uzasadnione jest wyrażanie efektu także w procentach (±E · 100%).

Tabela 4

Przykładowe parametry osiągnięć różnych klas w pomiarze Kultury Informatycznej

Klasa	Kod	A	B	C	D	K	L	M	N	R	S
Trend	±T	0,50	0,45	0,37	0,64	0,28	0,13	0,29	0,12	0,11	0,04
Wzrost	+e	0,27	0,22	0,22	0,34	0,21	0,18	0,19	0,21	0,21	0,19
Spadek	–e	–0,06	–0,04	–0,04	–0,17	–0,10	–0,12	–0,14	–0,16	0,19	–0,20
Efekt	±E	0,21	0,18	0,18	0,17	0,11	0,06	0,05	0,05	0,02	–0,01

W praktyce pomiaru Kultury Informatycznej (tabela 4) w klasach A-D o najlepszych wynikach ogólnych efekt ± E osiągał wartości 0,17 ÷ 0,21 przy wzrostach +e = 0,22 ÷ 0,34; w klasach słabszych (K-N) ±E = 0,05 ÷ 0,11 przy spadkach –e = 0,10 ÷ 0,16, a w najsłabszej (S) ±E = –0,01 przy +e = 0,19 i –e = 0,20. Trend ±T przyjmował tam wartości z zakresu 0,37 ÷ 0,64; 0,12 ÷ 0,29 oraz 0,04. Nie ma przy tym prostych zależności między uśredniającym *Trendem* a *Efektem* czułym na przypadki. Jak widać na przykładzie klasy S, nawet znak tych wskaźników może różnić się w przypadku wartości bliskich zeru. W klasie tej zaszło zjawisko ilustrujące cechy obu parametrów. Oto średni i słabsi respondenci uzyskali tendencję wzrostową (+L; +M), a najlepsi odnieśli się krytycznie do procesu (–H). Wskutek zmian dodatnich na dwu poziomach trend ma wartość +0,04, lecz efekt wykazał obniżkę –0,01, gdyż oceny krytyczne były silniejsze od przychylnych.

TRAFNOŚĆ PROCESÓW

Bilansowanie jakości z intensywnością oraz konkluzywnością wypowiedzi jest kluczem selekcyjnym doboru bodźców-stwierdzeń w kwestionariuszu. To dodatkowe kryterium jest istotnym wyróżnikiem Skali Ważonych Ocen. Fakt, iż jest to kryterium wewnętrzne, nie umniejsza jego rangi, a wręcz przeciwnie – zwykle jest odniesieniem cenniejszym od zewnętrznej metody sędziów kompetentnych. Nie chodzi tu jednakże wyłącznie o samą optymalizację narzędzia, lecz także o praktyczną przydatność proponowanej metody. Jeśli szanujemy wypowiedzi respondentów, np. naszych uczniów, to zjawisko wzajemnej zgodności wobec poruszanych kwestii winno być dla nas wiążące, zwłaszcza wtedy, gdy jako nauczyciele staramy się na nich oddziaływać.

Wśród wielu możliwych realizacji narzędzi wskażę na dwa odmienne przykłady zastosowań skal ważonych. Pierwszy dotyczył eksperymentu rów-

noległego, polegającego na porównaniu matury nowatorskiej względem tradycyjnej[2]. Jedyną zmienną modyfikującą była forma egzaminu z matematyki – pierwszego dnia testowa, a drugiego problemowa. Obie formy oceniali uczniowie i nauczyciele. W pomiarze tym nie chodziło o zmiany w czasie, lecz o różnice wypowiedzi dokładnie tych samych osób wobec obu form. Z tego względu, przy zachowaniu omówionej tu metody pomiaru, uległy modyfikacji jedynie niektóre nazwy. Parametr *Zmiana* zastąpiły *Różnice*, w miejsce *Trendu* weszła *Preferencja* danej formy matury, a *Wzrost/Spadek* zostały zastąpione przez opcje *Za/Przeciw testowi*. Respondenci oceniali czynniki merytoryczno-przedmiotowe i formalno-organizacyjne. Znamienne, że te pierwsze wypadły dobrze, a drugie znacznie gorzej. Wiadomo już, co trzeba poprawić, aby przybliżyć wartości uznawane przez cztery różne gremia kompetentnych sędziów: decydentów od spraw formalnych, metodyków od meritum, uczniów zdających oraz nauczycieli poprawiających prace. Cieszy fakt, że metodycy matematyki doceniają rolę badania postaw **podmiotów** matury.

Drugi przykład to pomiar trafności kształcenia informatycznego, niezbędny do rzetelnej samooceny i doskonalenia[3]. Nauczyciel jest w stanie poszukać wyjaśnień niezliczonych zjawisk zachodzących w jego klasie. Aby jednak dostrzec pewne symptomy nie wystarcza intuicja, niezbędne jest czułe narzędzie pomiaru różnicowego zmian zachodzących w czasie. Dopasowanie zajęć do poziomu możliwości i oczekiwań uczniów warunkuje efektywność działań. Zbyt wysoki stopień trudności, ignorowanie progów percepcji ujawnia się nieproporcjonalnie dużym spadkiem wartości zmiennej *Poznanie* na poziomie L. Dyskomfort z powodu nadmiernego, nieustannego stresu powoduje spadek zmiennej *Doznanie* na wszystkich trzech poziomach HML, co oznacza, że uczniowie nie wytrzymują presji. Z kolei trywialność treści i brak spełnienia aspiracji tych najlepszych objawia się obniżką poziomu H, mimo że poziom L może zwyżkować. Takie zjawiska występują bardzo intensywnie w klasach o silnym zróżnicowaniu, nie tylko na zajęciach z Informatyki. A przecież o jakości procesu decyduje równomierność zmian na wszystkich trzech poziomach. Konieczna jest zbieżność nauczycielskich propozycji z uczniowskimi oczekiwaniami, nie tylko w zakresie metod, lecz także w zakresie uznawanych wartości i treści nauczania.

W psychologicznych definicjach akcentuje się względną stabilność postaw, tymczasem z empirii wyłania się wyraźnie chaos nastawień wobec nauki w szkole. Dochodzi do paradoksu, kiedy to raczej wiedza pozostaje na stałym poziomie. Problem balansowania między doznawaniem a pozna-

[2] S. Ubermanowicz, *Wyniki pomiaru wartości próbnej matury z matematyki*, [w:] *Nowa matura – matematyka*, Kuratorium Oświaty, Poznań 1997.

[3] S. Ubermanowicz, *Jakość zajęć z informatyki*, [w:] *Media a edukacja*, Wyd. eMPI^2, Poznań 1997.

waniem przez uczniów **kulturowego universum** jest wyzwaniem dla współczesnego nauczyciela. Zamiast martwić się nikłymi efektami realizacji faktograficznych programów nauczania, lepiej skupić się na strukturach i treściach międzyprzedmiotowych, próbując własnych metod harmonizowania emocji, subtelnego stymulowania motywacji, kształtowania roztropnej świadomości, a zwłaszcza wyrobienia zdolności do generowania autowiedzy. Każdy nauczyciel stosuje swoiste technologie edukacyjne, po czym za efekty swej pracy wystawia stopnie... uczniom. A przecież warto, aby badając proces kształcenia pomiarem różnicowym za pomocą Skal Ważonych Ocen, mógł co pewien czas ocenić także sam siebie.