PRZEGLĄD PSYCHOLOGICZNY, 2013, TOM 56, Nr 1, 29-43

Kontrowersje wokół rzetelności jako pojęcia psychometrycznego

Elżbieta Aranowska
Instytut Psychologii
Uniwersytetu Kardynała Stefana Wyszyńskiego
Jolanta Rytel*
Instytut Psychologii
Uniwersytetu Kardynała Stefana Wyszyńskiego

CONTROVERSIES
OVER THE PSYCHOMETRIC CONCEPT OF RELIABILITY

Abstract. Reliability is an essential standard when determining whether or not data acquired from a psychological test is trustworthy. In the article the analysis of the concept of psychological test reliability is discussed from the perspective of within-person score variability as well as from the perspective of outcome measurement approach. The problems concerning reliability estimation are specified for different score structures. The article explains when the negative value of estimates of reliability is expected to occur in the course of analysis. New estimates of reliability (inter-class correlation coefficients) are also described and interpreted. Finally, the legitimacy of some of statutory directives concerning psychological diagnosis based on psychological tests is discussed.

Od kilku już dekad przytacza się we wszystkich podręcznikach z zakresu psychometrii stwierdzenie, że rzetelność dotyczy dokładności mierzenia i interpretowana jest jako „stałość pomiarów, gdy procedura badania testem jest powtarzana dla jednostek lub dla grup badanych osób" (*Standardy dla testów*, 2007, s. 56). Takie rozumienie pojęcia rzetelności powinno prowadzić do dwu różnych koncepcji: jednej – szczegółowej dla jednostki, i drugiej – ogólnej dla konkretnej populacji. Należy – i to z wysokim prawdopodobieństwem – do-

* Adres do korespondencji: Jolanta Rytel, Instytut Psychologii, Uniwersytet Kardynala Stefana Wyszyńskiego, ul. Wójcickiego 1/3, bud. 14, 01-938 Warszawa; e-mail: j.rytel@uksw.edu.pl

mniemywać, że „przyłożenie" testu psychologicznego do jednej osoby uruchamia zupełnie inne sytuacyjne reakcje behawioralne niż w przypadku innej osoby. Oczywiście takich reakcji nie spodziewamy się przy mierzeniu sprężystości różnych sprężyn czy wysokości różnych stołów, ale pewnie – w jakimś niewielkim zakresie – już się ich (czyli indywidualnych reakcji: wciągania brzucha, stawania na palcach, garbienia się, pochylania głowy itp.) spodziewamy przy mierzeniu wysokości różnych osób. I o ile najważniejsze czynniki niesprzyjające dokładności mierzenia, np. wysokości osoby, można łatwo wychwycić w procedurze pomiaru i je kontrolować, o tyle nie sposób ujawnić ich wszystkich w procedurze mierzenia własności psychicznych. Badacz czy diagnosta stara się kontrolować zaledwie kilka najważniejszych z nich. Z reguły jednak są to czynniki zewnętrzne, wspólne – jako zmienne zniekształcające pomiar – dla całych populacji i mające charakter błędów częściowo systematycznych. Rzadko kiedy kontroli podlegają indywidualne czynniki wewnętrzne, takie jak poziom koncentracji na zadaniu, poziom agrawacji czy symulacji, ciśnienie tętnicze, sposób reakcji na szkła kontaktowe, specyficzny styl odpowiadania, złe samopoczucie itd., mimo że stałość reakcji – i w konsekwencji wyników mierzenia (pomiarów) – a zatem dokładność pomiaru, wydaje się głównie zdeterminowana przez specyficzne własności jednostki, podobnie jak chwilowy poziom nasilenia konkretnej własności psychicznej jednostki.

Należy więc sądzić, że nie wszystkie jednostki (osoby) w podobny sposób „poddają się" procedurze pomiaru danym narzędziem. Mówiąc inaczej, mimo iż jedni podlegają pełnej procedurze mierzenia tak jak inni, należałoby mieć mniejsze zaufanie do efektów mierzenia własności tych drugich. Jednych „łatwiej" zmierzyć (wynik jest bardziej jednoznaczny, charakterystyczny dla nich), innych – „trudniej" (wynik jest mniej jednoznaczny). W języku statystycznym należałoby stwierdzić, że zmienność wartości pomiarowych „wewnątrz osób badanych" może być bardzo różna, utrudniając zarówno ocenę samego narzędzia pomiaru, jak i interpretację wyniku. Ten wyraźnie kazuistyczny charakter procedury mierzenia nie znajduje odbicia w klasycznej teorii pomiaru (CT) ani w teorii odpowiedzi na pozycje (IRT). Można go częściowo kontrolować, odwołując się do nowej teorii rzetelności, tzw. teorii generalizowalności wyników (Aranowska, 2005), na poziomie skumulowanej zmienności wewnątrz wszystkich osób badanych, ale nie jednostki.

Standardy dla testów (2007), nakazując szacowanie rzetelności narzędzi psychologicznych, równocześnie nie podają sugestii, kiedy rzetelność – zgodnie z powszechną praktyką – może podlegać ocenie raz, a kiedy należy tę ocenę wyznaczać po wielokroć, gdy wiadomo, jak bardzo zależy ona od jednorodności analizowanej grupy. Wyjątek stanowi tu standard 2.11 (oraz 2.12, jako szczególny przypadek 2.11), który nakłada na wydawcę testu obowiązek obwieszczania różnych ocen rzetelności dla różnych populacji, gdy zaistnieją „ogólnie akceptowane przyczyny teoretyczne lub empiryczne" (*Standardy dla testów*, 2007, s. 71) tych rozbieżności. W sposób niejawny próbuje się tu przemycić dyrektywę o potrzebie permanentnej oceny psychometrycznej testu przy każ-

dym jego zastosowaniu, z czym nie sposób się nie zgodzić. W konsekwencji, w tym przynajmniej względzie, panowałaby zbieżność procedur stosowanych przy ocenie trafności i rzetelności. Wynika stąd, że pojęcie „rzetelność testu", taktowane jako uniwersalne i niezależne od szeroko rozumianej sytuacji badawczej, jest nieadekwatne, test bowiem może być rzetelny zaledwie warunkowo dla odpowiedniej populacji i odpowiedniego celu badania. Powyższe stanowi nową, ważną konkluzję dla praktyki psychometrycznej.

Drugim ważnym aspektem rzetelności w sensie dokładności mierzenia jest sposób ujmowania wyniku pomiaru, czyli jego struktury. W klasycznej teorii testów – od ujęcia Lorda i Novicka (1968) – jest to wynik przeciętny z teoretycznie nieskończonej liczby pomiarów, przeprowadzanych tym samym narzędziem z uwzględnieniem błędu. W aktualnym rozumieniu wyniku pomiaru w psychologii, a także w nowszych ujęciach formalnych (psychometrii), co najdobitniej zostało uwidocznione w teorii generalizowalności wyników, błąd rozkłada się na specyficzne części, wynikające z „zauważonych" i wprowadzonych przez badacza źródeł tego błędu (ang. *facet*). Addytywne składowe błędu odpowiadają efektom (głównym i interakcyjnym) kontrolowanych – w czasie procedury mierzenia – czynników ją zakłócających. Ten właśnie wynik przeciętny, uzyskany wyłącznie z pomiarów powtarzanych, oczyszczony ze świadomie kontrolowanych błędów na poziomie jednostki, oznaczałby jej najbardziej charakterystyczny poziom pomiaru i rozumiany byłby jako uogólniony, zgeneralizowany wynik pomiaru (czy – za Nowakowską, 1975 – generyczny). Niestety, nie istnieje teoria statystyczna umożliwiająca szacowanie takich wyników pojedynczego obiektu bez odwoływania się do wyników pomiaru innych osób z populacji, do której badany także należy. W konsekwencji zakłada się, że struktura wyniku (pomiaru) badanego składa się z sumy wielu elementów oddających wielkość oddziaływań na badanego kontrolowanych w rzeczywistości czynników dodanych do wartości oczekiwanej mierzonej cechy (własności) całej populacji oraz do ciągle jeszcze nie zidentyfikowanej cząstki pomiaru, (nieprawdziwie) nazwanej błędem pomiaru jednostki. W tym ostatnim elemencie bowiem zawiera się to, co indywidualne dla badanego, jego specyficzna reakcja na inne niż kontrolowane przez badacza źródła wpływów. Ostatecznie za wynik pomiaru $x_{ij...r}$ przyjmuje się:

$$x_{ij...r} = \mu + \alpha_i + \beta_i + ... + (\alpha\beta)_{ij} + ... + \varepsilon_{ij...r} \quad (1)$$

gdzie: μ – średnia populacyjna, α_i – efekt główny zadziałania specyficznego poziomu (i-tego) źródła błędu A, β_j – efekt główny zadziałania specyficznego poziomu (j-tego) źródła błędu B, $(\alpha\beta)_{ij}$ – efekt interakcji pomiędzy specyficznymi poziomami obu źródeł błędu, $\varepsilon_{ij...r}$ – błąd pomiaru.

Uznając powyższe założenie za zasadne, oszacowanie wyspecyfikowanych wyżej efektów głównych i interakcyjnych możliwe jest poprzez zastosowanie wieloczynnikowej analizy wariancji z całkowicie powtarzanymi pomiarami (po

źródłach błędów, czyli pozycjach, sytuacjach, ...), w której zawsze równocześnie występuje ukryty dodatkowy czynnik – Osoby (badane). Miarą rzetelności w tym wypadku jest – jak wiadomo – wewnątrzklasowy współczynnik korelacji dla pełnej struktury wyniku.

PROBLEMY Z SZACOWANIEM RZETELNOŚCI DLA RÓŻNYCH STRUKTUR WYNIKU

Jednakże struktura wyniku nie zawsze jest pełna. Jeśli badacz tylko jednokrotnie dokonuje pomiaru danej własności za pomocą odpowiedniego narzędzia u osoby badanej, struktura wyniku z konieczności nie będzie uwzględniała interakcji pomiędzy wprowadzonymi poziomami różnych źródeł błędów. W takiej sytuacji – ograniczając się na przykład do jednego czynnika, jaki stanowią pozycje testowe – struktura wyniku (pozbawiona składnika interakcji) może zostać przedstawiona w postaci trzech modeli wyniku pomiaru:

(2) Model mieszany: $x_{ij} = \mu + \alpha_j + \pi_i + \varepsilon_{ij}$,

gdzie α_j – efekt główny (stały) zadziałania j-tej pozycji narzędzia, $j = 1, 2, \ldots, k$, π_i – efekt główny (losowy) związany z i-tą osobą badaną, $i = 1, 2, \ldots, n$.

(3) Model stały: $x_{ij} = \mu + \alpha_j + \varepsilon_{ij}$;

(4) Model losowy: $x_{ij} = \mu + \pi_i + \varepsilon_{ij}$.

W klasycznej teorii testów rzetelność rozumiana jako zgodność wewnętrzna szacowana była wzorem Kudera-Richardsona 20 dla odpowiedzi dwukategorialnych (diagnostycznych lub nie) bądź też dla odpowiedzi wielokategorialnych (przy założeniu, że skala stanowi continuum) – wzorem α Cronbacha. Prace Cyrila Hoyta, wprowadzające możliwość traktowania zbioru odpowiedzi wszystkich osób badanych jako danych stanowiących najprostszy schemat analizy wariancji, doprowadziły do utożsamiania procedury wyznaczania rzetelności z obliczaniem odwrotności statystyki F_O służącej do weryfikowania hipotezy o braku różnic pomiędzy osobami badanymi.

(5) $$\hat{r}_{tt} = 1 - \frac{1}{F_O} = 1 - \frac{MS_{\text{reszta}}}{MS_{\text{osoby}}}.$$

Niewątpliwą zasługą Hoyta było zwrócenie uwagi na fakt, iż uzyskana w ten sposób wartość oszacowania będzie dokładnie równa wartości otrzymanej przy zastosowaniu wzoru – odpowiednio do rodzaju danych – KR-20 lub α Cronbacha. Równocześnie unaocznilo to wiele nadużyć klasycznej teorii testów:

(a) taki jak przedstawiony powyżej sposób szacowania rzetelności jest prawdziwy, a tym samym uprawniony, wyłącznie przy przyjęciu przez badacza mieszanej struktury wyniku (por. wzór 2). Oznacza to, że przyjmując inny model struktury wyniku, na gruncie klasycznej teorii testów, badacz nie miał żadnej możliwości wyznaczenia wartości rzetelności dokonanego pomiaru (por. Aranowska, 2005). Zatem szacował coś, czego nie można i nie wolno traktować jako oceny rzetelności. Praktyka tego rodzaju datuje się od dziesięcioleci i nadal ma się bardzo dobrze;

(b) z perspektywy statystycznej szacowanie zgodności wewnętrznej za pomocą wzoru KR-20 lub α Cronbacha zawyża ocenę rzetelności. Obydwa estymatory to estymatory obciążone (Winer, 1971);

(c) posługiwanie się dowolną metodą statystyczną jest uprawnione jedynie wtedy, gdy spełnione zostały jej założenia. Do założeń analizy wariancji jednoczynnikowej z całkowicie powtarzanymi pomiarami (na której to statystyce bazuje wzór 5) należą: normalność wielowymiarowego rozkładu prawdopodobieństwa pozycji narzędzia oraz symetryczność macierzy kowariancji. Konieczność sprawdzania tych założeń nigdy nie była *explicite* akcentowana w literaturze psychometrycznej. Okazuje się, że nawet w przypadku najprostszej metody parametrycznej, jaką jest test t Studenta, pogwałcenie założeń (o normalności rozkładu prawdopodobieństwa i jednorodności wariancji) prowadzi do znaczących zmian rozkładu prawdopodobieństwa wartości statystyki testu (Ramsey, 1980; Sawilowsky, Blair, 1992; Zimmerman, 1998).

Problemy omówione w punkcie a i b zostały rozwiązane przez Winera (1971), który przedstawił nieobciążone postaci estymatorów rzetelności, a nadto wyznaczył estymatory dla losowych modeli struktury wyników pomiaru (por. Aranowska, 2005).

Powyższe ujmowanie rzetelności w sensie zgodności wewnętrznej nastręcza dodatkowych problemów. Analizując wzór (5) widać, że wartość rzetelności będzie najwyższa, czyli będzie dążyć do jedności, gdy wariancja reszt, MS_{reszta}, będzie dążyć do zera i/lub wariancja przeciętnych (średnich) wyników osób, MS_{osoby}, dążyć do nieskończoności. Inaczej mówiąc, rozważmy dwie sytuacje: pierwszą, kiedy wariancja resztowa przyjmie wartość równą zero, oraz drugą, gdy wariancja wyników osób badanych jest maksymalna, rośnie do nieskończoności.

W sytuacji pierwszej rzetelność narzędzia będzie maksymalna, gdy rozbieżność pomiędzy reakcją osoby badanej na daną pozycję testową (wyrażoną wynikiem pomiaru specyficznej reakcji osoby na tę pozycję – x_{ij}) a typowym dla osoby badanej sposobem reagowania na wszystkie pozycje testowe (wyrażonym jako średnia reakcji osoby badanej na wszystkie pozycje testu – $\overline{x_i}$) jest taka sama jak rozbieżność typowej, średniej reakcji na tę pozycję wszystkich badanych ($\overline{x_j}$) ze średnią reakcją wszystkich badanych na wszystkie pozycje testowe ($\overline{x}$), co można przedstawić jako:

$$x_{ij} - \overline{x_i} = \overline{x_j} - \overline{x} \,. \tag{6}$$

Jest to prawda dla każdej osoby i każdej pozycji, co oznacza brak zmienności wynikającej z indywidualnego dla osób badanych sposobu traktowania pozycji testowych. Innymi słowy, reakcja osoby badanej nie jest w żaden sposób dla niej specyficzna. Tym samym wariancję reszt można przedstawić następująco:

$$MS_{\text{reszta}} = \frac{1}{(k-1)(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{k} \left(x_{ij} - \overline{x_i} - \overline{x_j} + \overline{x} \right)^2 . \tag{7}$$

Niezależnie od różnic indywidualnych między osobami, pojedyncze pozycje testowe oddziaływują na nie w identyczny sposób, czego konsekwencją jest stałość różnic reakcji na pary pozycji wśród wszystkich osób badanych. Jak łatwo zauważyć, brak wariancji reszt – gdy przyjmuje ona wartość równą zero – wynika z własności samego narzędzia, którego pozycje stanowią jednorodny zbiór bodźców. Reakcja na pozycje osób badanych zależy tylko od siły bodźca, jakim jest pozycja. Nie uwzględnia natomiast szczególnych własności osoby badanej, specyficznego sposobu jej reagowania w konkretnej sytuacji diagnostycznej, ani też wpływu wielu innych – niekontrolowanych w tej sytuacji – bodźców, które mogą spowodować, że w kolejnym pomiarze reakcja osoby badanej na tę samą pozycję będzie odmienna. Jak wspomniano wyżej, to możliwość dokonania pomiaru dynamiki procesu odpowiadania na te same bodźce winna leżeć u podłoża szacowania rzetelności pomiaru. Wymaga to jednak przeprowadzenia operacji mierzenia więcej niż jednokrotnie.

W sytuacji drugiej rzetelność narzędzia będzie tym większa, im bardziej typowy, uśredniony sposób reagowania osób badanych na pozycje testu ($\overline{x_i}$) będzie się różnił, czyli im bardziej osoby badane będą – w sensie psychicznym – od siebie różne. Rzetelność może być wyznaczona tylko wtedy, gdy mianownik wzoru (5) różni się od zera, innymi słowy, gdy przynajmniej dwie osoby różnią się typowym, uśrednionym sposobem reagowania.

$$MS_{\text{osoby}} = \frac{k}{n-1} \sum_{i=1}^{n} \left(\overline{x_i} - \overline{x} \right)^2 . \tag{8}$$

Rozważając obie z tych sytuacji jednocześnie (wartość MS_{reszta} maleje i wartość MS_{osoby} rośnie), zauważyć można swoisty paradoks. O ile minimalizowanie wartości licznika zakłada jednolitość reagowania osób badanych na pozycje narzędzia, o tyle maksymalizacja wartości mianownika odwrotnie – wymaga jak największych różnic w indywidualnych sposobach reagowania na narzędzie. Trudno sobie wyobrazić występowanie obu tych sytuacji równocześnie.

Inny problem stanowi brak precyzyjnie teoretycznie wyznaczonych granic rzetelności. Przedstawiony wyżej wzór (5) umożliwia otrzymanie wartości z przedziału $(-\infty, 1>$. Dzieje się tak, gdy licznik wzoru (5) jest większy od mianownika, czyli wtedy, gdy:

(9) $$\frac{1}{(k-1)(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{k}\left(x_{ij}-\overline{x_i}-\overline{x_j}+\overline{x}\right)^2 > \frac{k}{n-1}\sum_{i=1}^{n}\left(\overline{x_i}-\overline{x}\right)^2 .$$

Mnożąc dwustronnie przez $(k-1)(n-1)$, otrzymamy:

(10) $$\sum_{i=1}^{n}\sum_{j=1}^{k}\left(x_{ij}-\overline{x_i}-\overline{x_j}+\overline{x}\right)^2 > k(k-1)\sum_{i=1}^{n}\left(\overline{x_i}-\overline{x}\right)^2 .$$

Przekształcając lewą stronę nierówności, otrzymamy dwie sumy, mianowicie:

(11) $$\sum_{i=1}^{n}\sum_{j=1}^{k}\left[\left(x_{ij}-\overline{x_j}\right)-\left(\overline{x_i}-\overline{x}\right)\right]^2 =$$

$$\sum_{i=1}^{n}\sum_{j=1}^{k}\left[\left(x_{ij}-\overline{x_j}\right)^2-2\left(x_{ij}-\overline{x_j}\right)\left(\overline{x_i}-\overline{x}\right)+\left(\overline{x_i}-\overline{x}\right)^2\right] =$$

$$\sum_{i=1}^{n}\sum_{j=1}^{k}\left(x_{ij}-\overline{x_j}\right)^2-\underbrace{2\sum_{i=1}^{n}\sum_{j=1}^{k}\left(x_{ij}-\overline{x_j}\right)\left(\overline{x_i}-\overline{x}\right)}_{=0}+\sum_{i=1}^{n}\sum_{j=1}^{k}\left(\overline{x_i}-\overline{x}\right)^2 =$$

$$\sum_{i=1}^{n}\sum_{j=1}^{k}\left(x_{ij}-\overline{x_j}\right)^2+k\sum_{i=1}^{n}\left(\overline{x_i}-\overline{x}\right)^2 .$$

Przenosząc ostatni element wzoru (11) na prawą stronę nierówności (10), otrzyma się:

(12) $$\sum_{i=1}^{n}\sum_{j=1}^{k}\left(x_{ij}-\overline{x_j}\right)^2 > \left[k(k-1)-k\right]\sum_{i=1}^{n}\left(\overline{x_i}-\overline{x}\right)^2 .$$

Mnożąc obydwie strony nierówności przez odwrotność stałych prawej strony, uzyska się:

(13) $$\frac{1}{k(k-2)}\sum_{i=1}^{n}\sum_{j=1}^{k}\left(x_{ij}-\overline{x_j}\right)^2 > \sum_{i=1}^{n}\left(\overline{x_i}-\overline{x}\right)^2 .$$

Ocena rzetelności, wyznaczona za pomocą wzoru (5), będzie wartością ujemną, gdy rozbieżności pomiędzy specyficzną reakcją osoby badanej na j-tą pozycję testową a typowym, uśrednionym sposobem odpowiadania na nią całej

grupy dla wszystkich osób i wszystkich pozycji narzędzia, ważona odpowiednią stałą (związaną z odwrotnością liczby pozycji narzędzia), jest większa niż zmienność indywidualnych sposobów odpowiadania na wszystkie pozycje testowe – por. wzór (6). Gdy ważona zmienność indywidualnych sposobów reagowania na poszczególne pozycje testowe przekracza zmienność wynikającą z różnic w reagowaniu pojedynczych osób na cały test, rzetelność będzie ujemna (por. też Aranowska, 1996, 2005). Sytuacja ta została zilustrowana w tabeli 1.

Tabela 1.

Przykładowe wyniki trzech osób, badanych trzema pozycjami narzędzia

Osoby	Pozycje			Wynik ogólny	$\overline{x_i}$
	I	II	III		
1	3	2	7	12	4
2	7	3	2	12	4
3	2	7	6	15	5
Σ	12	12	15	39	12
$\overline{x_j}$	4	4	5	13	$\overline{x} = 4^1/_3$

Łatwo oszacować rzetelność narzędzia składającego się z trzech pozycji. Wykorzystując wzór (5), uzyska się wartość α = -9,00. Jak widać, zmienność wyników ogólnych otrzymanych przez osoby badane, a tym samym ich średnich, jest nieznaczna (MS_{osoby} = 3/2 x 6/9 = 1,00). Natomiast zmienność wyników w kolumnach jest duża, co oznacza znaczną rozbieżność specyficznych dla osób badanych sposobów reagowania na pozycje testu względem typowej, uśrednionej reakcji wszystkich osób na daną pozycję (MS_{reszta} = 10,00). Wartość licznika ułamka we wzorze (5) jest 10 razy większa od wartości mianownika. Lewa strona wzoru (13) wynosi zatem (1/(3 x 1)) x 10 = (1/3) x 10 = = 3,3(3), a ponieważ prawa jest równa 1,00, spełniona jest nierówność (13) i α przybiera wartość ujemną.

W celu lepszej ilustracji omawianego problemu, niżej zostanie zaprezentowany kolejny przykład, w którym reakcje osób badanych na pozycje testowe są bardziej ujednolicone, a ich wyniki ogólne – podobnie jak w przykładzie pierwszym – mało zróżnicowane.

Tabela 2.

Przykładowe wyniki trzech osób, badanych trzema pozycjami narzędzia

Osoby	Pozycje			Wynik ogólny	$\overline{x_i}$
	I	II	III		
1	5	3	4	12	4
2	5	3	4	12	4
3	8	3	7	18	6
Σ	18	9	15	42	14
$\overline{x_j}$	6	3	5	14	$\overline{x} = 4^2/_3$

W tym przypadku ocena rzetelności jest sensowna, $\hat{\alpha} = 0{,}75$. Zmienność wyników ogólnych otrzymanych przez osoby badane, a tym samym ich średnich, jest tym razem nieznacznie większa (MS_{osoby} = 3/2 x 24/9 = 4). Przeciwnie niż w przykładzie 1, zmienność wyników w kolumnach jest nieznaczna, co oznacza, że reakcje osób badanych na poszczególnie pozycje narzędzia są względnie jednorodne (MS_{reszta} = ¼ x 36/9 = 1,00). Wartość licznika ułamka we wzorze (5) tym razem jest czterokrotnie mniejsza od wartości mianownika, ocena rzetelności będzie zatem dodatnia.

Warto podkreślić, że im więcej pozycji k, tym mniejsza zmienność odpowiedzi badanych w ramach pojedynczych pozycji, przy równoczesnej małej zmienności między osobami wystarczy, by rzetelność była ujemna – por. wzór (13). Od dawna próbowano rozszerzać klasyczną koncepcję rzetelności, tak by uzyskiwać miary unormowane (Krus, Helmstadter, 1993). Ponieważ próby te okazały się mało efektywne, nie uwzględnia się ich w analizach prezentowanych w tym artykule.

Przypomnijmy jeszcze raz, że w klasycznej teorii testów ocena rzetelności narzędzi rośnie, gdy coraz większym różnicom indywidualnym między badanymi towarzyszą coraz mniejsze zmiany w ich reakcjach na pojedyncze pozycje testowe. Zmiany te – co bardzo ważne – nie muszą stanowić (choć mogą) wspólnego wzorca reagowania, wyrażonego na przykład przez wysoką korelację między pozycjami.

Analizowanie tych zmian w modelach osadzonych na gruncie klasycznej teorii rzetelności nie jest możliwe. Równocześnie ujawnienie wzorca tych zmian jest kluczowe, ponieważ stanowi on efekt oddziaływania różnych źródeł zakłóceń procedury mierzenia, a tym samym różnych błędów pomiaru. Z tego powodu winien on stanowić podstawę rozumienia dokładności pomiaru, a co za tym idzie – podstawę określania miar rzetelności. Założenie o jego istnieniu jest – jako zasadnicze – przyjmowane w nowej koncepcji rzetelności wyrażonej przez tzw. teorię generalizowalności wyniku pomiaru. Jak już podkreślano, dokładność pomiaru powinna odnosić się do przebiegu procesu reakcji bada-

nego na kolejne bodźce testu, czyli – mówiąc językiem statystycznym – bazować na zmienności wewnątrz osób badanych.

W przykładach prezentowanych powyżej źródło błędu pomiaru stanowić może zmienność wynikająca z operowania tą konkretną próbką pozycji, a także zmienność pojawiająca się we wszystkich planach badawczych – niekoniecznie traktowana przez badacza jako źródło błędu – wynikająca z tej konkretnej próby osób. Ignorowanie tej ostatniej zmienności prowadziłoby badacza do rozumienia struktury wyniku zgodnie z modelem stałym, przedstawionym we wzorze (3). Zmienność między osobami w tym przypadku należałoby zaliczyć do błędu pomiaru.

TEORIA GENERALIZOWALNOŚCI WYNIKU

Istotą nowej koncepcji oceny rzetelności testów psychologicznych jest indywidualizacja założeń badaczy co do źródeł błędów pomiarów i ocena tych błędów poprzez analizę wyników badań z adekwatnych (do liczby uznanych źródeł), wieloczynnikowych schematów analiz wariancji z powtarzanymi pomiarami (Cronbach i in., 1972; Shavelson, Webb, Rowley, 1989). Takiego wielostronnego szacunku błędów nie można wykonać, dokonując jednego pomiaru badanych tak, jak rutynowo czyniono do tej pory. Wymaga to od badacza inwencji co do tego, jak uzyskać wielokrotne pomiary (w różnych sytuacjach) bez generowania nowych źródeł błędów (np. wynikających z procesu uczenia się itd.). Rozwiązanie tego problemu jest na tyle trudne, że ciągle jeszcze w pewnych dziedzinach psychologii plany badań ograniczają się do jedno- i dwu- (jak w prezentowanych wyżej przykładach) lub maksymalnie trójczynnikowych schematów analiz wariancji z całkowicie powtarzanymi pomiarami.

Przy dwuczynnikowym modelu analizy wariancji z powtarzanymi pomiarami dla mieszanego modelu struktury wyniku postaci (1) właściwym współczynnikiem rzetelności narzędzia jest wewnątrzklasowy współczynnik korelacji, tzw. miara ρ_2 (por. Aranowska, 2005, s. 111):

$$(14) \qquad r_{tt} = \rho_2 = \frac{MS_{\text{osoby}} - MS_{\text{reszta}}}{MS_{\text{osoby}} + (k-1)MS_{\text{reszta}}}$$

Współczynnik ten można uważać za złożoną korelację ze wszystkich możliwych do utworzenia par pozycji testu, wyrażającą uśrednioną relację reakcji badanych na kolejne pozycje w trakcie rozwiązywania testu. Inaczej – wyrażającą stopień nasilenia określonego, stałego sposobu zmian reagowania badanych na kolejne zadania testowe czy też – wyrażającą ich zgodność reagowania.

Dla danych z przykładu (1) wartość wewnątrzklasowego współczynnika korelacji będzie – na podstawie (14) – ujemna, ponieważ MS_{osoby} jest mniejsza od MS_{reszta}. Istotnie:

r_{tt} = (1 – 10)/(1 + 2 x 10) = -9/21 = -0,42857.

Dla danych z przykladu (2):

r_{tt} = (4 – 1)/(4 + 2 x 1) = 3/6 = 0,50 i jest, zgodnie z interpretacją wartości korelacji, przeciętnym nasileniem zgodnego sposobu zmian reagowania badanych.

Interpretacja wartości wewnątrzklasowego współczynnika korelacji jest bardzo konserwatywna, mimo iż jego wartość jest zawsze mniejsza od wartości klasycznego już współczynnika zgodności wewnętrznej testu, jakim jest α Cronbacha (dowód – zob. Aranowska, 2005). Zgodny sposób reagowania, wywołany własnościami narzędzia (tu – pozycji), powinien być wyrażony przez korelację rzędu przynajmniej 0,7.

Warto podkreślić, iż w przypadku mieszanego modelu struktury wyniku, opisanego wyżej, wewnątrzklasowy współczynnik korelacji, ρ_2, stanowi dokładnie średnią korelację par pozycji wyznaczoną z ilorazu średniej kowariancji i uśrednionej wariancji pojedynczych zadań. [Czytelnik łatwo sprawdzi, że w przykładzie 2 wariancje kolejnych pozycji wynoszą odpowiednio 3, 0 oraz 3, natomiast cov_{12} = 0, cov_{13} = 3, zaś cov_{23} = 0.] Ostatecznie iloraz uzyska wartość [(0 + 3 + 0)/3]/[(3 + 0 + 3)/3] = 1/2 = 0,50. Natomiast w przykładzie (1) wariancje kolejnych pozycji są identyczne i wynoszą odpowiednio 7, a cov_{12} = = -3,5, cov_{13} = -6,5, zaś cov_{23} = 1. Ostatecznie iloraz stanowi wartość [(-3,5 – 6,5 + 1)/3]/[(7 + 7 + 7)/3] = -3/7 = -0,42857. Wartości współczynnika generalizowalności dla losowego modelu struktury wyniku w opisanych wyżej przykładach byłyby oczywiście inne od pokazanych, ale ich interpretacje byłyby podobne: jako przeciętnych korelacji wszystkich par pozycji.

W literaturze przedmiotu znaleźć można poważne opinie specjalistów z zakresu psychometrii podważające użyteczność klasycznych współczynników zgodności wewnętrznej jako miar rzetelności testów psychologicznych (Bentler, 2009; Green, Yang, 2009a, 2009b; Sijtsma, 2009a, 2009b). Mimo iż autorzy dyskutują między sobą nad nowym sposobem ujmowania rzetelności testów, nie ulega wątpliwości, że kwestionują przydatność α Cronbacha i szukają rozwiązania poprzez stosowanie modeli równań strukturalnych, uwzględniających zarówno pojedyncze pozycje testu, jak i jego wynik ogólny (Sijtsma, 2009b). Koncepcja ta jest ogólniejsza niż koncepcja generalizowalności. Obie koncepcje łączy postulat dotyczący priorytetu analizy zmian odpowiedzi jednostek badanych na kolejne pozycje testu, uwidoczniony w operowaniu korelacjami między pozycjami. Nowość tego ujęcia polega na poszukiwaniu takich zbiorów pozycji, na które grupa reaguje jednorodnie, i sprawdzeniu tej tezy poprzez weryfikację odpowiednio skonstruowanego modelu.

DYSKUSJA

Trzy grupy wskaźników tradycyjnie uznawało się za miary rzetelności: otrzymane na podstawie dwukrotnego badania równoległymi wersjami testu, otrzymane na podstawie dwukrotnego badania tym samym testem oraz uzyskane w wyniku jednokrotnego badania „i oparte na wielkościach korelacji między wynikami dla poszczególnych pozycji testu czy skal (współczynniki

zgodności wewnętrznej). [...] trzy wymienione kategorie można uznać za specyficzne przypadki ogólniejszej klasyfikacji: klasyfikacji współczynników generalizowalności" (*Standardy dla testów*, 2007, s. 60). Niestety, trudno się zgodzić zarówno z ostatnią konkluzją autorów *Standardów dla testów*, jak i z wcześniejszym podziałem miar.

Z interpretacji miar generalizowalności przedstawionej wyżej wynika wprost, że klasyczne współczynniki zgodności nie są i nie mogą być specyficznymi przypadkami tych miar. Gdyby tak było, wartości tych współczynników w podobnych sytuacjach badawczych byłyby identyczne, a – jak pokazano wyżej – tak nie jest. Tylko oceny rzetelności na podstawie dwukrotnego badania tym samym testem bądź wersją równoległą, rozumiane jako wartości współczynników korelacji *r* Pearsona, mogą być traktowane jako pseudomiary generalizowalności wyniku pomiaru. Klasyczne współczynniki zgodności wewnętrznej nie porównują korelacji, lecz wariancje (pozycji czy skal oraz wyników ogólnych testu).

Koncepcja generalizowalności jest nową, niezależną próbą oceny dynamiki zmian pomiaru w zmieniających się sytuacjach mierzenia, a takie ujęcie – zgodnie z intuicją – odpowiada wiarygodności uzyskiwanych wyników przy użyciu konkretnej aparatury, a zatem dotyczy rzetelności narzędzia pomiarowego, jego odporności na konkretne zakłócenia. W tym sensie ocena rzetelności jest warunkowa, zrelatywizowana do szeroko rozumianej sytuacji. Nie ma jednej wartości rzetelności na stałe przyporządkowanej testowi. Ta kontekstowość stanowi o niezaprzeczalnej ważności nowego ujęcia.

Posługując się danym testem, badacz podaje z reguły wartości klasycznych współczynników rzetelności z podręcznika dla testu, zazwyczaj jednak nie wyznacza ich dla badanej przez siebie populacji, uważając, że narzędzie ma stałą, przysługującą mu na zawsze rzetelność. Ukazując wyżej, jak bardzo wartości nowych miar zależą od wkładu indywidualnej zmienności jednostki, rysuje się potrzeba szacowania rzetelności za każdym razem, kiedy test jest używany, praktycznie w każdym badaniu.

Informacje o zachowaniu się narzędzia w różnych sytuacjach i dla odmiennych populacji zbierane przez laboratoria czy pracownie testów umożliwiłyby ogólną orientację na temat działania narzędzia i pozwoliłyby wyznaczyć warunki skrajne (sytuacje), w których test nie powinien być stosowany. Innymi słowy, weryfikacja twierdzenia o zadowalającej rzetelności testu powinna przypominać uzasadnienie każdej innej tezy empirycznej. Z tej perspektywy obowiązkiem badacza sięgającego po dane narzędzie jest przyczynianie się do ustalenia zakresu jego stosowalności.

Powyższa konkluzja w równej mierze dotyczy psychologa diagnosty, korzystającego z wyników narzędzi psychologicznych przy formułowaniu ocen różnych aspektów funkcjonowania jednostki. „Należy jednak pamiętać o tym, że odpowiedzialność za prawidłowe stosowanie i interpretowanie jego [testu – E. A., J. R.] wyników w zasadzie spoczywa na użytkowniku testu. Akceptując tę odpowiedzialność, użytkownik testu musi mieć odpowiednią wiedzę o wła-

ściwym zastosowaniu testu oraz o populacji, dla której ten test jest przeznaczony" (*Standardy dla testów*, 2007, s. 195-196).

Testy wykorzystywane są coraz częściej i na coraz większą skalę w niemal wszystkich dziedzinach życia. Powszechnie stosuje się je do oceny funkcjonowania psychologicznego jednostki i w procesie planowania, realizacji oraz oceny skutków działań interwencyjnych, do podejmowania decyzji o charakterze prawnym czy urzędowym, na gruncie edukacyjnym i zawodowym, a także w procesie planowania i wspierania rozwoju własnego jednostki. Interpretacja wyniku testu stanowi zaledwie jedną z informacji uwzględnianych przez psychologa przy formułowaniu orzeczenia. Mimo to często traktowana jest jako informacja najważniejsza „ze względu na przekonanie o ich [testów – E. A., J. R.] obiektywności i matematycznej precyzji" (*Standardy dla testów*, 2007, s. 194). Z tego powodu obowiązkiem użytkownika testu jest zadbać o jej jak największą wiarygodność. „Tworząc opinię diagnostyczną, należy dokonać interpretacji wyników testowych, odwołując się do ich rzetelności i trafności. [...] Od psychologów oczekuje się, aby wyciągając wnioski oparte na wynikach testowych, brali pod uwagę wszelkie dane o rzetelności i trafności wyników, które można traktować jako uzasadnienie lub brak uzasadnienia tych wniosków i odpowiednio ograniczać przedstawioną opinię" (tamże, s. 222-223). Opinie takie mają bardzo duże znaczenie dla dalszego życia diagnozowanej jednostki.

Orzeczenia wydawane przez psychologa niosą bardzo poważne skutki społeczne, polegające na przykład na odmówieniu jednostce możliwości wykonywania zawodu, możliwości ubiegania się o przyjęcie do określonej szkoły czy organizacji, wreszcie – na wykluczeniu samego diagnosty z grona specjalistów. Takie orzeczenia z mocy ustawy są konieczne chociażby do uzyskania prawa jazdy specjalnej kategorii, do zakwalifikowania się do zdawania egzaminu, wymaganego przy staraniu się o zdobycie licencji detektywa czy wykonywania zawodu policjanta, ale też niezbędne są dla uzyskania „zwolnień warunkowych, wyroków sądowych, przyznania opieki rodzicielskiej, kompetencji do stawania przed sądem oraz rozstrzygania sporów dotyczących naruszenia dóbr osobistych" (tamże, s. 207).

W Rozporządzeniu Ministra Zdrowia w sprawie badań psychologicznych kierowców i osób ubiegających się o uprawnienia do kierowania pojazdami oraz wykonujących pracę na stanowisku kierowcy (Dz. U. z dnia 26 kwietnia 2005 r., § 6.2) można znaleźć informację, iż badanie psychologiczne w szczególności obejmuje ocenę cech osobowości, sprawności intelektualnej oraz sprawności psychofizycznej. Wszystkie wymienione oceny mogą być sformułowane na podstawie wyników odpowiednich testów psychologicznych, a wydający je psycholog komunikacji z zasady już nie zastanawia się nad trafnością czy rzetelnością stosowanego narzędzia. Równocześnie konsultant krajowy w dziedzinie medycyny pracy dodaje wytyczne w celu wyjaśnienia wątpliwości przy przeprowadzaniu badań kierowców (Wągrowska-Koski, 2008, s. 9): „Zdolność do wzięcia udziału w ruchu drogowym u kierowcy określamy jako tzw. gotowość do prowadzenia pojazdu. Na gotowość tę składają się: zdolność

prawidłowego, logicznego myślenia, pamięć, krytycyzm, odpowiednia przerzutowość i pojemność uwagi, wyobraźnia, zdolność do podejmowania decyzji, a przede wszystkim także ogólna dobra kondycja i prawidłowy stan zdrowia. Pewną rolę odgrywają także naturalne skłonności charakterologiczne, jak ustępliwość i opanowanie emocjonalne. Wszystkie te cechy powinny występować w nadmiarze, tak aby mogły stanowić pewną rezerwę w razie konieczności podejmowania kolejnych i szybkich decyzji podczas prowadzenia pojazdów". Waga odpowiedzialności diagnosty jest wysoka. Orzeczenie otwiera możliwość prowadzenia pojazdu przez pięć lat (do 60. roku życia). Związana jest nie tylko z poprawnością diagnozy aktualnej, ale również z prognozą zachowania. Fałszywa diagnoza wyspecyfikowanych predyspozycji wiąże się z poważnymi skutkami, łącznie z wykluczeniem diagnosty z grona psychologów transportu (czyli usunięciem z listy marszałka województwa zawierającej nazwiska psychologów uprawnionych do przeprowadzania badań psychologicznych).

W innym Rozporządzeniu Ministra Zdrowia, w sprawie badań lekarskich i psychologicznych osób ubiegających się lub posiadających licencję detektywa (Dz. U. z dnia 15 września 2003 r., § 9.1), stwierdza się, że badanie psychologiczne obejmuje ocenę poziomu umysłowego, ocenę osobowości (ze szczególnym uwzględnieniem funkcjonowania w sytuacjach trudnych) oraz ocenę dojrzałości społecznej i emocjonalnej. Zakres ten może zostać poszerzony na wniosek psychologa. Takich regulacji prawnych, zgodnie z którymi ocena psychologiczna warunkuje możliwość uzyskania bądź kontynuowania wybranego zawodu, jest znacznie więcej.

Praktyka używania narzędzi psychologicznych zwykle jest zrutynizowana. Wiara urzędników państwowych w nieomylność diagnozy psychologicznej na podstawie testów jest tak samo niepokojąco wysoka, jak niektórych ich użytkowników. Niestety, to ci drudzy muszą przyjąć na siebie obowiązek permanentnego śledzenia informacji o zmieniających się własnościach narzędzi psychologicznych (z populacji na populację, z wiekiem, ze względu na płeć, wraz z moderującym wpływem innych zmiennych itd.) oraz interweniowania w odnośnych gremiach przy ewentualnej niemożności wywiązania się z podjętych zobowiązań (wystawiania orzeczeń). Niska rzetelność automatycznie uniemożliwia ocenę trafności. Diagnoza przeprowadzona narzędziem nietrafnym jest nie tylko diagnozą niepoprawną, lecz także dającą podstawy do zakwestionowania wiarygodności całego procesu oceny.

BIBLIOGRAFIA

Aranowska, E. (1996). *Metodologiczne problemy zastosowań modeli statystycznych w psychologii. Teoria i praktyka*. Warszawa: Studio 1.

Aranowska, E. (2005). *Pomiar w psychologii*. Warszawa: Wydawnictwo Naukowe Scholar.

Bentler, P. M. (2009). Alpha, dimension-free, and model-based internal consistency reliability. *Psychometrika*, 74, 1, 137-143.

Cronbach, L. J., Gleser, G. C., Nanda, H., Rajaratnam, N. (1972). *The dependability of behavioral measurements: The theory of generalizability for scores and profiles*. New York: Wiley.

Green, S. B., Yang, Y. (2009a). Reliability of summed item scores using structural equation modeling: An alternative to coefficient Alpha. *Psychometrika*, 74, 1, 155-167.

Green, S. B., Yang, Y. (2009b). Commentary of coefficient Alpha: A cautionary tale. *Psychometrika*, 74, 1, 121-135.

Krus, D. J., Helmstadter, G. C. (1993). The problem of negative reliabilities. *Educational and Psychological Measurement*, 53, 3, 643-650.

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

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

Ramsey, P. H. (1980). Exact Type I error rates for robustness of Student's "t" test with unequal variances. *Journal of Educational Statistics*, 5, 337-349.

Sawilowsky, S. S., Blair, R. C. (1992). A more realistic look at the robustness and Type II error properties of the "t" test to departures from population normality. *Psychological Bulletin*, 111, 352-360.

Shavelson, R. J., Webb, N. M., Rowley, G. L. (1989). Generalizability theory. *American Psychologist*, 44, 6, 922-932.

Sijtsma, K. (2009a). On the use, the misuse, and the very limited usefulness of Cronbach's Alpha. *Psychometrika*, 74, 1, 107-120.

Sijtsma, K. (2009b). Reliability beyond theory and into practice. *Psychometrika*, 74, 1, 169-173.

Standardy dla testów stosowanych w psychologii i pedagogice (2007). American Educational Research Association, American Psychological Association, National Council on Measurement in Education. Tłum. E. Hornowska. Gdańsk: Gdańskie Wydawnictwo Psychologiczne.

Wągrowska-Koski, E. (2008). Wskazówki do przeprowadzenia badań profilaktycznych kierowców. *Praca i Zdrowie*, 2, 1-12.

Winer, B. J. (1971). *Statistical principles of experimental design*. New York: McGraw-Hill.

Zimmerman, D. W. (1998). Invalidation of parametric and nonparametric statistical tests by concurrent violation of two assumptions. *Journal of Experimental Education*, 67, 55-68.