

ARTYKUŁY

ETYCZNY WYMIAR DECYZJI EKSPERCKICH PODEJMOWANYCH NA PODSTAWIE WYNIKÓW TESTÓW PSYCHOLOGICZNYCH CZĘŚĆ I

Jerzy Brzeziński

Instytut Psychologii
Uniwersytet im. Adama Mickiewicza
Poznań

ETHICAL DIMENSION OF THE EXPERTISE DECISIONS BASED ON RESULTS OF PSYCHOLOGICAL TESTS. PART I

Summary. The article, which is a first part of a bigger study, concerns the ethical context of decisions taken towards people on the base of their results in psychological tests. The Author singled out two kinds of determinants of such decisions: (1) internal and (2) external ones. Internal determinants – only these are mentioned here – influence the psychometrical goodness of the test, and the quality of quantitative (statistical) and qualitative (psychological) interpretation of the test result. The range of the issues mentioned in the article was determined by three perspectives, equivalent to three theories: psychological, psychometrical and statistical one. The Author focused his attention on the dominant role of theoretical perspective. He also claims that the ethicality of the expert using psychological test goes beyond the traditional concept of ethical evaluation of psychologist's activity towards the individuals that come under his various influences. Ethically biased expert's activity may be connected with gross lacks in his methodological awareness and with uncritical approach to the test, which sometimes is a psychological test in nothing but name. The Author recommends respecting in this scope the instructions formulated in Standards (AERA/APA/NCME, 1999).

W procedurach diagnostycznych stosowanych przez różne instytucje ważną rolę odgrywają eksperci (rzeczoznawcy). To ich zgodna (w zadowalającym stopniu) opinia, odnosząca się do różnych aspektów szeroko pojmowanych kompetencji osób podlegających ich ocenie stanowi jeden z istotnych elementów procedury kwalifikacyjnej stosowanej przez te instytucje. Wiadomo, że nawet w małym i jednorodnym zespole ekspertów niekiedy dość trudno wskazać kryteria,

Adres do korespondencji: Jerzy Brzeziński, Instytut Psychologii UAM,
ul. Szamarzewskiego 89, 60-568 Poznań.

formułowanych *implicite* opinii, zwłaszcza gdy nie mają one wyrazistej ilościowej natury. Jeżeli formułowane przez ekspertów interpretacje wyników zastosowanych przez nich narzędzi mają mieć walor *intersubiektywności*, to bardzo ważne staje się pytanie nie tylko o *dokładność* (precyzję) zastosowanych narzędzi, ale też o trafność owej interpretacji. Ponadto, tak mierniam, *obiektywizm* wydawanych przez ekspertów ocen wzrośnie, jeżeli będą się oni odwoływać do dobrze *zdefiniowanych* i *skwantyfikowanych* kryteriów. Te, jak pokazuje kilkudziesięcioletnie doświadczenie, najlepiej pozyskać, odwołując się do zestawu (baterii) testów psychologicznych.

Poprawnie skonstruowany i zastosowany test psychologiczny (a więc taki, który stosowany jest zgodnie ze *standardami* przedstawionymi w: AERA/APA/NCME, 1999) może być zastosowany jako w maksymalnym stopniu spełniające wymóg *intersubiektywności* narzędzie. Podobnie traktować należy baterię testów.

Dobrze poznana jest problematyka ilościowego ujęcia *zgodności* (rzetelności) sędziów w odniesieniu do wydawanych przez nich ekspertyz. Dość powszechnie stosowaną miarą jest jakiś *współczynnik zgodności sędziów* – może to być *współczynnik W-Kendalla*, gdy oceniane obiekty można rangować (por. np. Brzeziński, 2005a, s. 500–505), albo współczynnik rzetelności sędziów obliczany na podstawie danych, do których zastosowano model *ANOVA* z powtarzaniem pomiarów zmiennej zależnej (por. Winer, Brown, Michels, 1991; Brzeziński, 2000). Analizując rzetelność sędziów zwraca się uwagę także na ich *stabilność*, czyli niezmienność w jakimś okresie czasu formułowanych przez nich ocen. Na problematykę trafności, formułowanych przez rzeczoznawców interpretacji wyników użytych narzędzi, kładzie się relatywnie mniejszy nacisk. Zakłada się też – przynajmniej milcząco – że są oni *etyczni*, czyli nie kierują się w swojej pracy eksperckiej uprzedzeniami i nie są podatni na żadne zewnętrzne wpływy, które mogłyby wypaczyć ich sądy, czyniąc je *stronniczymi* (ang. *bias*).

Poza *rzetelnością* i *trafnością* ważna jest, jak wspomnieliśmy, strona *etyczna* – była o niej mowa w moim wcześniejszym artykule, w którym zebrałem doświadczenia z pracy w zespołach eksperckich dwóch podstawowych polskich instytucji dokonujących ewaluacji prowadzonych przez uczelnie kierunków studiów: Uniwersyteckiej Komisji Akredytacyjnej i Państwowej Komisji Akredytacyjnej (por. Brzeziński, 2003a). Rzecz jasna, owa etyczność jest też w znaczącym stopniu warunkowana przez rzetelność i trafność formułowanych ocen. Musimy bowiem pamiętać, że sama wysoka rzetelność narzędzia nie gwarantuje równie wysokiej trafności zastosowanej, w danym kontekście, interpretacji (tu: psychologicznej) jego wyniku. Może przecież być tak, że sędziowie posługują się rzetelnym kryterium (np. wynikiem kwestionariusza osobowości czy testu inteligencji), ale to wcale nie musi oznaczać, iż zastosowana przez nich interpretacja wyniku będzie trafna. Rzetelny (w sensie gwarantujący precyzyjny, stabilny pomiar) chronometr nie będzie dostarczał podstaw do zadowalająco trafnych interpretacji uzyskanego przez daną osobę czasu wykonania jakiegoś standardowego zadania w kontekście dotyczącym *osiągnięć szkolnych* uczniów. Jednakże związek *czasu reakcji* czy tzw. *czasu inspekcji IT* (ang. *Inspection Time*) z poziomem *inteligencji* był i jest badany przez psychologów i – w ograniczonym zakresie – może być wykorzystywany w cząstkowych pomiarach związanych z badaniem pewnych aspektów inteligencji. Także sędziowie mogą, w sposób wysoce zgodny, formułować swoje sądy w oparciu o dane kryterium (wysoka rzetelność), ale niekoniecznie interpretacja związków ocenianego zachowania z kryterium musi być trafna (np. *rasa* jako predyktor sukcesów zawodowych – por. kontrowersyjny wniosek książki R. Hernsteina i Ch. Murraya z 1994 roku: *The bell curve* mówiący o tym, że Afroamerykanie są mniej inteligentni niż biali Amerykanie; różnica = -15 jednostek IQ, czyli 1 odchylenie standardowe)?

Jak się wydaje, dobrym punktem wyjścia do dyskusji nad problemem *etyczności* ocen formułowanych przez ekspertów, posilkujących się w procesie podejmowania decyzji wynikami testów psychologicznych jest zwrócenie uwagi na kilka kluczowych zagadnień związanych z ich kompetencjami w zakresie *stosowania* testów psychologicznych i *interpretowania* ich wyników.

Zacznijmy od przytoczenia ważnego wskazania odnoszącego się do tego, *kto* może odpowiedzialnie posługiwać się testami psychologicznymi. Odpowiedzialnie, czyli w sposób zgodny z intencją *twórcy testu* (ang. *test developer*) i z pełną świadomością ograniczeń interpretacyjnych wyniku testowego. W stuletniej już historii testu psychologicznego¹ powstało bardzo dużo testów, które wymagają zróżnicowanych kwalifikacji od ich użytkowników. Można – za APA (2000) – wyróżnić trzy grupy testów psychologicznych, angażujące coraz większe kwalifikacje ich *użytkowników* (ang. *test users*):

- A – proste testy (takie jak, na przykład, testy kompetencji zawodowych czy testy osiągnięć), które mogą być stosowane przez nie-psychologów,
- B – bardziej złożone testy, takie jak: testy inteligencji ogólnej (np. *test Ravena*) czy testy zainteresowań zawodowych, które mogą być stosowane przez osoby posiadające podstawową wiedzę i także kwalifikacje w zakresie konstruowania testów i ich stosowania, analizy statystycznej i odpowiednich teorii psychologicznych,
- C – indywidualnie stosowane testy inteligencji (np. kolejne edycje: *S-B*, *WAIS* i *WISC*) i testy osobowości (np. *MMPI*, *NEO-FII*, *16 PF*) oraz techniki projekcyjne (np. *Rorschach*, *TAT*, *PZN*), które mogą być stosowane przez ekspertów „with at least a Master’s degree in psychology” i z przynajmniej roczną praktyką diagnostyczną, odbytą pod kierunkiem doświadczonego psychologa.

Podobny, aczkolwiek bardziej rozbudowany, podział testów na cztery grupy: A, B, C, D został także przyjęty przez grono ekspertów *Pracowni Testów Psychologicznych* działającej przy *Polskim Towarzystwie Psychologicznym*:

- “A – testy zastrzeżone wyłącznie dla psychologów i wymagające specjalnego przeszkolenia w zakresie ich stosowania,
- B – testy zastrzeżone wyłącznie dla psychologów bez dodatkowych warunków,
- C – testy dopuszczone do stosowania przez nie-psychologów z wyższym wykształceniem po przeszkoleniu,
- D – testy dopuszczone do stosowania przez nie-psychologów z pokrewnym wykształceniem do badań naukowych lub działalności zawodowej korespondującej z rodzajem metody”².

Moim zdaniem, to ważne, aby do świadomości potencjalnych użytkowników testów psychologicznych oraz odbiorców ekspertyz wystawianych przez specjalistów docierały informacje o *minimalnym* poziomie kwalifikacji użytkownika zakładanym przez twórcę i producenta testu. Zbyt dużo bowiem pseudospecjalistów sięga – bez specjalnego przygotowania – po testy psychologiczne. Oczywiście psycholog zobowiązany jest przez kodeksy zawodowe i inne

¹ Rok 1905, w którym A. Binet i T. Simon ogłosili swój pierwszy test inteligencji można uznać za początek „ery” testów w psychologii (więcej na temat ewolucji testów inteligencji w: Boake, 2002; Brzeziński, 2003b).

² Na podstawie: Protokół IV spotkania zespołu ds. określenia zasad posługiwania się metodami i narzędziami psychologicznymi z 27 marca 2002 roku (maszynopis).

zalecenia zawodowe (np. ITC, 2000; APA, 2002; PTP, 1994)³, aby chronić testy przed osobami, które nie są specjalistami, ale, tak naprawdę, są to tylko zalecenia. Jeżeli jednak (tu, w Polsce) odbiorcy ekspertyz psychologicznych będą zwracali uwagę na to czy są one sporządzone przez *magistrów psychologii* – bo zgodnie z obowiązującą od 1 stycznia 2006 roku *Ustawą z dnia 8 czerwca 2001 r. o zawodzie psychologa i samorządzie zawodowym psychologów* (DzU 2001 r. Nr 73, poz. 763) do wykonywania zawodu psychologa uprawniona jest jedynie osoba, która legitymuje się dyplomem magistra psychologii (por. Art. 8 ust. 1 pkt 1 cytowanej ustawy) – to, działając zgodnie z obowiązującym prawem, przyczynimy się do wyeliminowania z rynku usług psychologicznych (tu: eksperckich) osób, które nie są *profesjonalistami*.

Założmy – aby nie prowadzić zbyt szeroko zakrojonej analizy – iż eksperci spełniają kryteria dostępu do testów psychologicznych z grupy C (według standardów APA, 2000) czy z grupy A i B (według roboczej sugestii zespołu ekspertów *Pracowni Testów Psychologicznych PTP*). Są zatem psychologami, którzy mają doświadczenie w zakresie operowania w indywidualnym badaniu diagnostycznym testami typu *WAIS* czy *MMPI* oraz technikami projekcyjnymi, takimi jak: *Rorschach* czy *TAT*.

Problem etyczności postępowania eksperta, posiłkującego się w formułowaniu określonej ekspertyzy testem psychologicznym (baterią testów) pojmuję dość szeroko. To znaczy, że za etyczne/nieetyczne uznaję nie tylko:

- odstępstwa od zaleceń kodeksu etycznego (np. APA, 2002; PTP, 1994) w zakresie postępowania psychologa wobec *osób biorących udział w badaniu testowym* (ang. *test takers*), ale też
- postępowanie „eksperta”, który nie ma dostatecznej *wiedzy fachowej*: psychologicznej, psychometrycznej, metodologicznej i statystycznej, która wymagana jest dla poprawnej, *ilościowej i jakościowej* interpretacji wyników testowych,
- nie ma dostatecznych *umiejętności praktycznych* w stosowaniu testu psychologicznego.

Ekspert (przypominam, że w odniesieniu do pewnych testów psychologicznych może nim być jedynie psycholog) odwołujący się do wyniku testowego musi się, przede wszystkim, wyka-

- zać stosownymi kompetencjami w zakresie:
- teorii psychologicznej,
- teorii psychometrycznej,
- teorii statystycznej.

³ I tak: Standard 9.11 *Maintaining Test Security (Ethical Principles of Psychologists and Code of Conduct*, APA, 2002, s. 1072): „(...) Psychologists make reasonable efforts to maintain the integrity of test materials and other assessment techniques consistent with law and contractual obligations, and in a manner that permits adherence to this Ethics Code”; Art. 12 *Kodeksu etyczno-zawodowego psychologa* (PTP, 1994): „(...) Nie udostępnia specyficznych technik diagnozy psychologicznej (w tym testów psychologicznych – przyp. J. B.) osobom nieprzygotowanym do ich kompetentnego stosowania. Psycholog przeciwstawia się podejmowaniu działalności psychologicznej, a zwłaszcza stosowaniu specyficznych technik diagnostycznych i terapeutycznych przez osoby nieposiadające kwalifikacji psychologicznych”; Pkt 1.4: *Ensure that test materials are kept securely (International Guidelines for Test Use, 1. Take responsibility for ethical test use*, ITC, 2000): „1.4.3. Protect the integrity of test by not coaching individuals on actual test materials or other practice materials that might unfairly influence their test performance”, 1.4.4. *Ensure that test techniques are not described publicly in such a way that their usefulness is impaired*”.

Ponadto musi wziąć pod uwagę *prawa* i *obowiązki* osoby badanej. Patrząc bowiem całościowo na trafność i rzetelność ekspertyzy „testowej”, nie można się ograniczać tylko do analizowania postępowania samego eksperta – aczkolwiek to ono musi przede wszystkim przykuwać naszą uwagę – ale też trzeba wziąć pod uwagę i to, że osoba badana ma nie tylko swoje prawa (i o nich nie może zapominać ekspert), ale ma też obowiązki (o czym nie zawsze pamiętamy). Najdogłębniej zostały owe „prawa” i „obowiązki” ujęte w dokumencie przygotowanym przez zespół roboczy *American Psychological Association* (APA, 1998): *Rights and responsibilities of test takers: Guidelines and expectations*.

W niniejszym artykule, który stanowi pierwszą część obszerniejszego opracowania⁴ skupię się tylko na analizie kontekstu etycznego związanego z *perspektywą teoretyczną* testu psychologicznego – psychologiczną, psychometryczną i etyczną.

Teoria psychologiczna

Za przełomową w historii psychometrii datę należy uznać rok 1955, w którym ukazał się jeden z najważniejszych artykułów odnoszących się do tworzenia i stosowania testów psychologicznych. Była to praca L. J. Cronbacha i P. Meehla poświęcona nowej odmianie trafności – *trafności teoretycznej* (ang. *construct validity*; por. Cronbach, Meehl, 1955/2005). Zgodnie z koncepcją obu badaczy, o wartości testu, a dokładniej zastosowanej przez użytkownika interpretacji wyniku testowego, przesądza leżąca u jego podstaw teoria psychologiczna. Interpretacja wyniku testu psychologicznego odnosi się bowiem do *terminów teoretycznych (konstruktów)*, które mają swój sens właśnie na gruncie owej, dobrze *uprawdopodobnionej* teorii. W tym sensie, można powiedzieć o interpretacji wyniku uzyskanego przez daną osobę w kwestionariuszu *BDI A* Becka ogniskującej się na ocenie poziomu depresji, iż ma ona satysfakcjonującą trafność teoretyczną, gdyż została poprawnie wywiedziona, jak i sam kwestionariusz, z dobrze empirycznie potwierdzonej *poznawczej teorii depresji A. Becka*. Zinterpretowany w terminach tej teorii wynik może stanowić podstawę dla adekwatnej oceny postępów, jakie czyni pacjent poddany psychoterapii, a także dla dokonywania zadowalających prognoz co do społecznego funkcjonowania pacjenta po pomyślnym zakończeniu leczenia.

Trzeba zauważyć, że nie wszystkie testy psychologiczne, zwłaszcza te, które powstały w pierwszej połowie XX wieku, były konstruowane w taki sposób, że pierwotna względem nich była określona teoria psychologiczna. Takie „duże” testy, jak *Skale Inteligencji D. Wechslera (WAIS, WISC, WPPSI)* czy kwestionariusz *MMPI* powstawały w ten sposób, że ich twórcy, kierując się dużym doświadczeniem praktycznym (np. klinicznym: D. Wechsler czy S. R. Hathaway i J. C. McKinley), zainteresowani byli tym, aby uzyskać narzędzie, za pomocą którego można

⁴ Analizie kontekstu etycznego, odnoszącego się do pozostałych uwarunkowań ekspertyzy odwołującej się do wyniku testowego będzie poświęcony drugi artykuł, który także ukaże się w *Polskim Forum Psychologicznym*.

będzie dobrze różnicować osoby o różnym natężeniu badanej cechy. Niejako wtórnie, dokonano rekonstrukcji założeń teoretycznych *implicite* „tkwiących” w danym teście⁵.

Empirycznie sprawdzona teoria psychologiczna „legitymizuje” dany zbiór obrazków, pytań, przymiotników czy zadań jako naukowo (na gruncie psychologii) uzasadniony test psychologiczny. Poza kontekstem tej teorii będzie to tylko naukowo „pusty” zbiór pozycji. Dlatego, na przykład, nie można uznać zestawionej przez L. Szondię serii fotografii przedstawiających pacjentów psychotycznych jako testu psychologicznego, który pozwalałby na interpretację wyników spełniającą kryteria podane w cytowanym artykule L. J. Cronbacha i P. Meehla.

Zatem, to *teoria psychologiczna* (dobrze empirycznie uzasadniona!) jest pierwszym i podstawowym kryterium oceny tego, czy dane narzędzie będzie można włączyć do baterii testów, która będzie dostarczała zbioru wyników (ujętych na przykład pod postacią *profilu psychometrycznego* – na wzór profilu *WAIS* czy *MMPI*). Ten zaś profil będzie miał swoją interpretację psychologiczną, która pozwoli na podjęcie racjonalnych decyzji wobec poddanych badaniu osób. Odejście od tego kryterium będzie równoznaczne z odejściem od nauki na rzecz, na przykład, astrologii czy innych sztuk tajemnych.

Teoria psychometryczna i teoria statystyczna

Wszystkie, powszechnie stosowane przez psychologów w badaniach typu *diagnostycznego* testy psychologiczne zostały skonstruowane w psychometrycznych ramach modelu *wyniku prawdziwego* (ang. *true score model*; tzw. *klasyczna teoria testów*), którego pełny wykład dał w swoim fundamentalnym dziele: *Theory of mental tests*, H. Gulliksen (1950; por. jego omówienie w: Nowakowska, 1975). Jego rozwinięcie przyniosła praca F. M. Lorda i R. M. Novicka (1968): *Statistical theories of mental test scores*. Kolejnym etapem prac nad rozwojem teoretycznych podstaw testów psychologicznych były dwie teorie: *teoria wyników generycznych* (ang. *generic true scores*) i jej uogólnienie pod postacią *teorii uniwersalizacji* (ang. *generalizability theory*) przedstawione w monografii: *The dependability of behavioral measurements: Theory of generalizability for scores and profiles* (por. Cronbach, Rajaratnam, Nanda, Gleser, 1972). Najnowszy model psychometryczny, to *Item Response Theory* (w skrócie: *IRT*; polska nazwa: *teoria odpowiedzi na pozycje testu*). Czytelnik może zapoznać się z nim z dostępnego w języku polskim *Wprowadzenia do teorii odpowiedzi na pozycje testu* Ch. L. Hulina, F. Drasgowa i Ch. K. Parsonsa (2005).

Zarówno dla osoby tworzącej test, jak i dla psychologa ów test stosującego i interpretującego uzyskany wynik testowy, a także wysnuwającego z niego określone wnioski praktyczne

⁵ Jak napisano w podręczniku do *WAIS-III* (por. Kaufman, Lichtenberger, 1999, s. 3): „The development of Wechsler's tests was not based on theory (except perhaps on Spearman's (1927) *g*, or general intelligence theory) but instead on practical and clinical perspectives. (...) Wechsler's view of IQ tests was that they were a way to peer into an individual's personality. Years after the development of the Wechsler scales, extensive theoretical speculations have been made about the nature and meaning of these tests and their scores, but originally the tests were developed without regard to theory” (podkr. moje). Czytelnik może zajrzeć do pracy E. Hornowskiej (1998), która dokonała rekonstrukcji założeń teoretycznych Skali Inteligencji D. Wechslera w terminach Cattella i Horna *teorii płynnej i skryształizowanej inteligencji* oraz Guilforda *modelu struktury intelektu*. Także w podręczniku do *WAIS-III* wymienia się te próby szukania teoretycznych podstaw *Skali Inteligencji WAIS-III* (por. Kaufman, Lichtenberger, 1999, rozdz. 4: *How to interpret the WAIS-III*, s. 77–163).

(np. uznającą badaną osobę za spełniającą kryterium selekcji) ważne jest, aby test spełniał założenia objęte jego modelem psychometrycznym. Z tego modelu wyprowadza się bowiem kryteria psychometrycznej *dobroci* testu. W modelu H. Gulliksen'a – a on, przypominam, jest zakładany przez bez mała wszystkie znane w świecie testy, np. *WAIS-III* czy *NEO-FFI* – ważne jest, aby test cechowała wysoka *rzetelność* (ang. *reliability*) i jak najniższa wartość *błędu standardowego pomiaru SEM* (ang. *standard error of measurement*). Bez znajomości wartości tych dwóch parametrów, a zwłaszcza *SEM* nie jest możliwe ilościowe zinterpretowanie wyniku testowego. Mówiąc dokładniej, niemożliwe jest oszacowanie nieznanego i wprost niedostępnego psychologowi wartości *wyniku prawdziwego*⁶.

Psycholog – i tu wkraczamy na pole teorii statystycznej – odwołujący się do wyniku testowego (dokładniej: *wyniku otrzymanego*) powinien posłużyć się *estymacją przedziałową wyniku prawdziwego*, wychodząc jedynie od poznanej, w efekcie przeprowadzonego badania testowego, wartości *wyniku otrzymanego*. Zwracają na to uwagę autorzy *Standards...* (AERA/APA/NCME, 1999, por. Standard 2.2., s. 31). W rozdziale: *Reliability and errors of measurement* (op. cit., s. 29) zwraca się uwagę psychologa na to, iż w procedurze interpretacji wyników osoby badanej znacznie ważniejsza jest znajomość *standardowego błędu pomiaru* aniżeli *współczynnika rzetelności* testu:

„Standard 2.2.

Należy przedstawić dane dotyczące standardowego błędu pomiaru, tak całkowitego jak i warunkowego jeżeli ma on znaczenie, zarówno w terminach wyników surowych (czy wyrażonych na wyjściowej skali), jak i na skali wyników przeliczonych, zalecanej dla określonej interpretacji tych wyników.

Komentarz: Najczęściej wykorzystywane wyniki przeliczone to wyniki standardowe, równoważniki wieku czy klasy oraz wyniki centylowe. Ponieważ surowe wyniki testowe w testach zorientowanych na normy są rzadko bezpośrednio interpretowane, to standardowe błędy pomiaru wyrażone w jednostkach przeliczonych są bardziej użyteczne w typowych wykorzystaniach testu. Przedział ufności dla wyniku prawdziwego, wyniku uniwersalnego czy wyniku centylowego osoby badanej pełni niemal te same funkcje co standardowy błąd pomiaru i może być wykorzystywany jako alternatywna forma przedstawiania danych na temat rzetelności. Dane na temat standardowego błędu pomiaru są szczególnie istotne w takich sytuacjach, w których decyzje muszą być podjęte natychmiast, a dodatkowe źródła informacji są ograniczone” (tł. E. Hornowska).

Posługiwanie się, w pracy eksperckiej, wprost *wynikami otrzymanymi*, a nie budowanymi wokół nich *przedziałami ufności* prowadzić może do podejmowania wobec osób badanych niekorzystnych dla nich decyzji. Dość często psychologowie, występując przed różnymi organami decyzyjnymi (np. sądami) informują o wartościach wyników „otrzymanych”, a nie o dolnej

⁶ Podstawowe równanie H. Gulliksen'a *klasycznej teorii testów* wygląda następująco: *wynik otrzymany w teście przez osobę badaną* (ang. *obtained test score*), X_t = *nieobserwowalny wynik prawdziwy* (ang. *true score*), X_u + *składnik błędu* (ang. *error component*), X_e . Chcąc tedy poznać prawdopodobną wartość *wyniku prawdziwego* psycholog musi wyjść od wyliczonej, na podstawie przeprowadzonego badania empirycznego, wartości współczynnika *rzetelności* testu. Znając wartość współczynnika *rzetelności* (aktualną dla danej populacji, do której adresowany jest test!) psycholog może wyliczyć wartość *SEM*, daną wzorem: $SEM = SD_t \sqrt{1 - r_{tt}}$, gdzie: SD_t – *odchylenie standardowe* (ang. *standard deviation*) *wyników otrzymanych* testu w próbie standaryzacyjnej, r_{tt} – *współczynnik rzetelności* testu. J. Brzeziński (2005a) podaje sześć metod empirycznego szacowania wartości współczynnika r_{tt} .

i górnej granicy przedziału ufności, który – z poziomem ufności $(1 - \alpha)100\%$ – zawiera szacowaną wartość „wyniku prawdziwego”. Przyjmijmy, że, przykładowo, osoba badana w wieku 30 lat uzyskała w WAIS-R(PL) IQ (w *Skali Pełnej*) = 69 (górną granicą przedziału upośledzenia umysłowego dla kategorii F70: *upośledzenie łagodne*: 50-69)⁷. Jeżeli przyjmiemy także, dość powszechnie akceptowany 90% poziom ufności, to, w przybliżeniu, możemy powiedzieć, iż wynik prawdziwy osoby badanej lokuje się w przedziale wartości IQ: 69 ± 7 , czyli: 62-76⁸. To zaś oznacza, że równie prawdopodobny jest każdy wynik IQ z tego przedziału, a osobę badaną (kierując się tylko wartością IQ) można sklasyfikować jako upośledzoną umysłowo w stopniu *łagodnym*, jak i lokującą się w przedziale *pogranicza upośledzenia*. To zaś nie jest obojętne dla danej osoby, wobec której mogą zapaść brzemienne dla dalszych jej losów życiowych (szkoła, zakład pracy, renta) decyzje. W tych decyzjach będzie miał swój udział (z czego nie zawsze zdaje sobie sprawę!) psycholog i musi on sobie zdawać sprawę z ich etycznych konsekwencji. Te zaś są pochodną jego przygotowania psychometrycznego.

Znajomość *rzetelności* testu, oszacowanej za pomocą jednej z sześciu metod (por. Brzeziński, 2005), niezbędna jest dla wyznaczenia dokładności pomiaru danej cechy (w oparciu o znaną wartość *SEM*). Jednakże rzetelność testu nic nie mówi psychologowi o tym, czy stanowi on adekwatną miarę danego *konstraktu*. Jak wiadomo można z bardzo dużą dokładnością mierzyć *czas reakcji* (*CR*), ale zwiększanie dokładności jego pomiaru (np. do rzędu 0,001 s) wcale nie oznacza, że *CR* jest adekwatną miarą inteligencji (oczekiwał tego na przykład J. McKean Cattell w swoich, zresztą nieudanych, próbach konstrukcji pierwszego „testu umysłowego”). Psycholog musi zatem zadać pytanie o to, co tak naprawdę „jego” test mierzy? Jest to pytanie o *trafność* testu (ang. *test validity*).

Współczesne rozumienie trafności, przyjęte m. in. przez autorów *Standards...* (AERA/APA/NCME, 1999), akcentuje to, że należy mówić o *trafności interpretacji wyniku testowego*, a nie o trafności testu jako takiego:

„Pojęcie trafności odnosi się do stopnia, w jakim dane empiryczne oraz teoria uzasadniają interpretację wyników testowych w zakładanym kierunku. Trafność jest zatem najbardziej podstawowym etapem w procesie tworzenia i oceny testu. Proces walidacji obejmuje ciągle zbieranie danych w celu dostarczenia mocnych podstaw naukowych dla proponowanej interpretacji wyników testowych. Procesowi walidacji podlega zatem kierunek interpretacji wyników testowych, a nie test jako taki. Jeżeli wyniki testowe są interpretowane na więcej niż jeden sposób, to każdy kierunek zakładanej interpretacji powinien zostać poddany walidacji” (tl. E. Hornowska; podkr. moje).

Odnosząc się do wyżej przytoczonego stanowiska autorów *Standards...* w kwestii trafności chciałbym zwrócić uwagę Czytelnika na dwie sprawy. Po pierwsze, i najważniejsze, akcentuje się w nim – jak zresztą w całym tym opracowaniu – rolę teorii (tu: psychologicznej) w samym procesie tworzenia testu oraz w interpretacji wyniku testowego. Zatem, wartość testu psychologicznego jest w dużej mierze pochodną wartości (statusu metodologicznego) stworzonej (lub przyjętej, bo już taka funkcjonuje) przez psychologa teorii. Ona to nadaje sens psychologiczny wynikowi testowemu, którego precyzja określona jest przez wartość *SEM*.

⁷ Według *Klasyfikacji zaburzeń psychicznych i zaburzeń zachowania w Międzynarodowej Statystycznej Klasyfikacji Chorób i Problemów Zdrowotnych ICD-10* (tzw. Rewizja dziesiąta z 1993 roku).

⁸ Por. J. Brzeziński i in. (1996, Tabela 15, s. 41).

Po drugie, dane empiryczne, niezbędne w procesie walidacji mogą być zbierane na kilka różnych sposobów (AERA/APA/NCME, 1999; Brzeziński, 2005a, rozdz. 17; Brzeziński, 2005c, rozdz. 7–12). W odniesieniu do testów osiągnięć, ale też niektórych kwestionariuszy osobowości ważna jest analiza treści pozycji testu i jego związku z mierzonym *konstruktem*, zamierzoną przez autora interpretacją wyniku testowego i jego zastosowaniem w praktyce diagnostycznej. Taka analiza danych prowadzi do ustalenia, według tradycyjnej terminologii, *trafności treściowej* (ang. *content validity*); mówi się też o *treściowym aspekcie* trafności. Wynik testowy może też być konfrontowany z wynikami innych narzędzi (w tym i innych testów psychologicznych) czy zachowań określonych grup (np. odnoszących sukces w jakiejś dziedzinie) – ogólnie z danymi *kryterialnymi*. Tradycyjnie takie dane mogą być zbierane na dwa sposoby. Pierwszy, tradycyjnie utożsamiany z trafnością dignostyczną (ang. *concurrent validity*) polega na równoczesnym porównaniu wyniku testowego z kryterium. Z kolei drugi sposób polega na porównaniu wyniku testowego z danymi kryterialnymi zebranymi po upływie pewnego czasu. Inaczej mówiąc, idzie o dokonywanie *predykcji kryterium* na podstawie wyniku testowego. Tradycyjnie ten sposób zbierania danych kryterialnych i ich porównywania z wynikiem testowym określany był mianem *trafności prognostycznej* (ang. *predictive validity*). Ogólnie mówi się o *trafności kryterialnej* (ang. *criterion validity*) albo *zorientowanej na kryterium* (ang. *criterion oriented validity*). Mówi się też, odpowiednio, o diagnostycznym i prognostycznym aspekcie trafności. Od czasu pracy L. J. Cronbacha i P. Meehla zaczęto kłaść silny nacisk na jeszcze jeden aspekt trafności – *trafność teoretyczną* (ang. *construt validity*)⁹. W Standardzie 1.1. *Standards...* owa rola teorii (psychologicznej) została bardzo mocno zaakcentowana:

„Standard 1.1.

Należy przedstawić informacje uzasadniające każdą proponowaną interpretację i wykorzystanie wyników testowych, w tym dane empiryczne dotyczące trafności interpretacji wyników testowych oraz teorię leżącą u podstaw testu.

Komentarz: Uzasadnienie to powinno obejmować wskazanie, jakie dane są konieczne do uzasadnienia zamierzonej interpretacji wyników testowych. Wyniki zbiorcze powinny obejmować analizę logiczną wraz z danymi empirycznymi i uzasadniać sposób wykorzystania testu. Wymagane dane mogą pochodzić z badań lokalnych prowadzonych w obszarach, w których test ma być stosowany, z badań prowadzonych wcześniej, lub z syntetycznych analiz badań spełniających założone kryteria. Żaden z rodzajów danych nie jest bardziej preferowany od innych. To jakość i związek danych z zamierzonym wykorzystywaniem testu przemawiają za ich wagą. Prezentacja danych dotyczących trafności powinna uwzględniać wagę tych informacji z punktu widzenia tego, co znajduje się w literaturze przedmiotu, włączając w to także dane niespójne z punktu widzenia zamierzonej interpretacji testu. Na osobach tworzących test spoczywa odpowiedzialność za dostarczenie danych potwierdzających proponowany przez nich kierunek interpretacji, natomiast na osobach stosujących test spoczywa odpowiedzialność

⁹ D. Westen i R. Rosenthal (2003, s. 608) uznali, iż: „trafność teoretyczna jest jednym z najważniejszych pojęć w całej psychologii. Jest ona sercem jakichkolwiek badań w których badacze używają takiej miary jaką jest wskaźnik zmiennej, która nie jest bezpośrednio obserwowalna (np. inteligencja, agresja, pamięć robocza). Jeśli testowi psychologicznemu (albo, szerzej, procedurze psychologicznej, włączając manipulację eksperymentalną) brak trafności teoretycznej, to rezultaty otrzymane z pomocą takiego testu czy takiej procedury będą trudne do zinterpretowania”. Powiem więcej – nie przedstawiają one żadnej wartości naukowej. Podobnie F. M. Lord i M. R. Novick (1968, s. 278), twórcy drugiej generacji modelu psychometrycznego, traktują trafność teoretyczną jako „najważniejszą właściwość testu”.

za ocenę jakości danych walidacyjnych z punktu widzenia zastosowania testu w lokalnej sytuacji" (tł. E. Hornowska).

Jeżeli psycholog przystępuje do zbierania danych empirycznych uzasadniających dany sposób *interpretacji wyniku testowego*, to powinien on postępować w sposób opisany w kolejnej, bardzo ważnej w rozwoju psychometrii pracy, porównywalnej z cytowaną już pracą L. J. Cronbacha i P. Meehla. Uwaga ta dotyczy słynnego artykułu D. T. Campbella i D. W. Fiskego *Convergent and discriminant validation by the multitrait-multimethod matrix* (1959/2005), w której autorzy przedstawili nową metodę badania trafności testu – tzw. metodę analizy „macierzy wielu cech – wielu metod” (ang. *multitrait – multimethod matrix*). Ta metoda daje najpełniejszy wgląd w trafność interpretacji wyniku testowego, gdyż pozwala określić nie tylko zakres *zbieżności* wyniku nowego testu z wynikami innych miar podobnych konstruktów (*aspekt zbieżny*) oraz *aspekt różnicowy* wyniku testowego z wynikami innych testów mierzących inne konstrukty.

Na silny związek trafności z etyką postępowania psychologa konstruującego test psychologiczny i przystępującego do interpretacji i stosowania jego wyniku zwracał uwagę przed ćwierćwieczem wybitny psycholog S. Messick w pracy: *Trafność testu a etyka oceny (diagnozy)*: „Trafność testu jest więc zintegrowanym ewaluatywnym sądem co do adekwatności i odpowiedniości wniosków wyprowadzanych na podstawie wyników testowych. Ta ocena opiera się na czterech podstawach: (1) na indukcyjnym podsumowaniu zbieżnych i różnicowych dowodów empirycznych świadczących o tym, że wyniki testowe można interpretować w terminach znaczenia konkretnego konstruktów, (2) na ocenie opartej na wartościowaniu konsekwencji takiej interpretacji, (3) na racjonalnych uzasadnieniach i dowodach na rzecz odpowiedniości konstruktów i użyteczności wyników w konkretnych zastosowaniach testu oraz (4) na ocenie potencjalnych społecznych konsekwencji proponowanego wykorzystania testu oraz aktualnych konsekwencji wynikających z jego użycia” (Messick, 1980/2005, s. 480; tł. M. Zakrzewska).

Także i on zwracał uwagę na rolę teorii w procesie walidacji, a dokładniej na precyzyjne zdefiniowanie *konstruktów* i powiązanie z nim wyniku testowego poddawanego określonej interpretacji psychologicznej. Trafność teoretyczna znalazła się w centrum modelu trafności testu – opracowanego przez S. Messicka (op. cit., Ryc. 12.2, s. 481) – wiążącym takie elementy, jak: *trafność teoretyczna, interpretacja testu, wartościowanie konsekwencji, decyzje, zastosowania testu*.

W kontekście tu prowadzonych rozważań warto jeszcze zwrócić uwagę – a co większość psychologów-praktyków zdaje się nie zauważać – na Standard 1.2. sformułowany w *Standards...* (AERA/APA/NCME, 1999):

„Standard 1.2.

Autor testu powinien wyraźnie określić, jaki jest pożądany kierunek interpretacji i sposób wykorzystywania wyników testowych. Populacja (lub populacje), dla których test ma być stosowany powinny być wyraźnie zdefiniowane, podobnie jak konstrukt mierzony przez test.

Komentarz: Twierdzenia dotyczące trafności powinny odnosić się do konkretnych sposobów interpretacji czy zastosowań wyników testowych. Posługiwanie się sformułowaniem: „trafność wyników testowych” nie jest poprawne. Żaden test nie jest trafny dla wszystkich celów i wszystkich sytuacji. Każde rekomendowane zastosowanie czy interpretacja testu wymaga potwierdzenia trafności. Należy określić jasnym i czytelnym językiem populację, dla której test jest przeznaczony, mierzony z założenia konstrukt oraz sposób i kontekst, w jakim wyniki testowe mają być wykorzystywane” (tł. E. Hornowska; podkr. moje).

Odpowiedzialne użycie testu, to takie, które związane jest z badaniem trafności każdego nowego – odbiegającego od dotychczasowych i uznanych za standardowe – zastosowania testu i każdej nowej, odbiegającej od dotychczas stosowanych, interpretacji psychologicznych wyniku testowego. Przykładowo, wykorzystanie Wechslerowskich *Skal Inteligencji WAIS-R* i *WAIS-III* o znanej trafności *teoretycznej* i trafności *treściowej* interpretacji *FIQ*, *VIQ* i *PIQ* oraz *wyników przeliczonych* (ang. *scaled scores*) testów składających się na tę baterię, w innych celach, aniżeli pierwotnie założony przez D. Wechslera cel, jakim było diagnozowanie poziomu i struktury inteligencji osoby badanej wymaga ponownego sprawdzenia trafności nowego wykorzystania wyników owych *Skal Inteligencji* i nowej interpretacji ich wyników. Taka konieczność występuje – moim zdaniem – gdy za pomocą profilu psychometrycznego *WAIS-R* czy *WAIS-III* chcemy oceniać stopień podobieństwa strukturalnego jakichś klinicznych grup kryterialnych do grupy tzw. „normy”.

Uważam, że problem etycznych konsekwencji wynikających z lekceważenia przez psychologów roli teorii psychologicznej w konstruowaniu i interpretowaniu testu nie znalazł jeszcze należytego naświetlenia w podręcznikach testowych i w kursach z zakresu psychometrii. Znacznie większą rolę przywiązuje się do precyzji wyniku testowego (badanie jego rzetelności i zwracanie uwagi na minimalizowanie *SEM*), aniżeli do trafności różnorodnych zastosowań testu oraz interpretacji wyniku testowego. Jakąś próbą dokonania pozytywnych zmian w tym obszarze diagnostyki psychometrycznej w Polsce było wydanie zbioru podstawowych, klasycznych prac psychometrycznych odnoszących się do zagadnień trafności i, w mniejszym zakresie, rzetelności (por. Brzeziński, 2005c) i przekonanie Gdańskiego Wydawnictwa Psychologicznego, aby zechciało wydać polskie tłumaczenie podstawowego dla każdego psychologa zbioru standardów: *Standards for educational and psychological testing*. Mam nadzieję, że wpłynie to na poziom świadomości diagnostycznej polskich psychologów-praktyków.

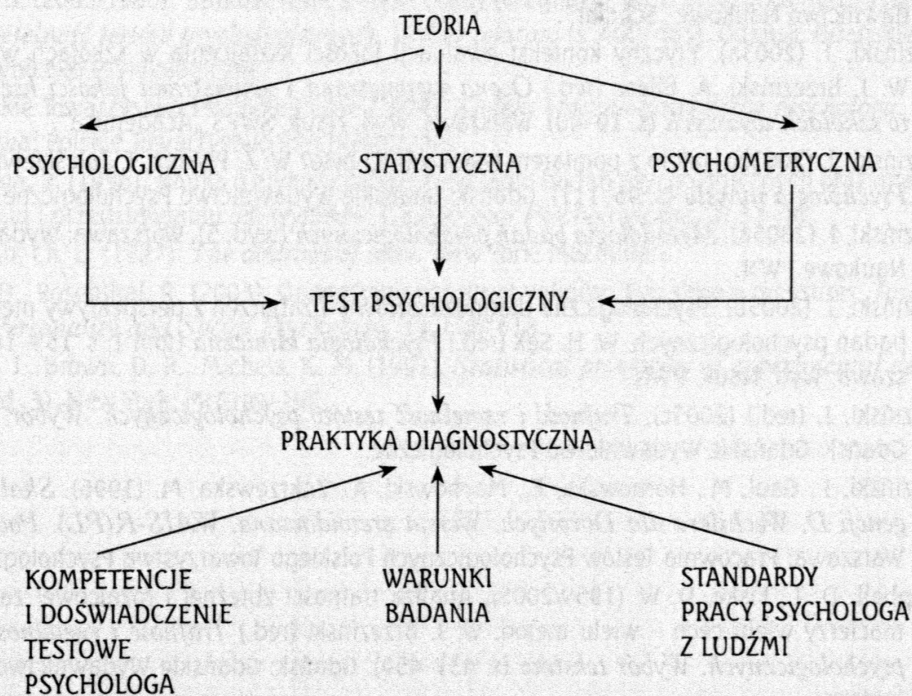
Podsumowanie

W wielkim skrócie można powiedzieć, że zakres poruszanych w tym artykule zagadnień został wyznaczony przez trzy perspektywy, równoznaczne z trzema *teoriami*: *psychologiczną*, *psychometryczną* i *statystyczną*. Jednakże nie są one równo znaczące. To perspektywa teorii psychologicznej jest tu nadrzędna. Wszak w języku określonej teorii psychologicznej wynik testowy „coś” znaczy, przyjmuje określony sens, podlega naukowo uzasadnionej interpretacji. I w tym sensie *teoria psychologiczna* zakreśla granice sensownego użycia testu psychologicznego.

Na rycinie 1 schematycznie zostały przedstawione podstawowe uwarunkowania diagnostycznego użycia testu psychologicznego, które mają dwojaki charakter. Są one:

- *wewnętrzne* względem samego testu i mające wpływ na jego *dobroć* psychometryczną i na *jakość* interpretacji *ilościowej* (statystycznej) i *jakościowej* (psychologicznej) wyniku testowego,
- *zewnętrzne*, obejmujące: (1) kompetencje zawodowe i doświadczenie użytkownika w zakresie stosowania testu, (2) warunki badania, a także (3) przyjęte przez psychologa standardy pracy z ludźmi (wszystko to, co zwykle się określać standardami etycznymi, a co ujęte jest w kodeksach etycznych) i mające, pośrednio, wpływ na końcowy efekt diagnostyczny.

W niniejszym artykule skupiłem się na zanalizowaniu wyłącznie *uwarunkowań wewnętrznych* oraz na ukazaniu ich wzajemnych powiązań. Starałem się też przekonać Czytelnika, iż etyczność postępowania psychologa-eksperta sięgającego po test psychologiczny i legitymizującego za pomocą środków naukowych – w oparciu o jego wynik – decyzje instytucji zamawiających ekspertyzę wykracza poza tradycyjne pojmowanie etycznej waloryzacji postępowania psychologa wobec osób podlegających jego oddziaływaniom (diagnostycznym, pomocowym, szkoleniowym itp.). Otóż – i wcale nie jest to wyłącznie moja teza¹⁰ – etycznie obciążone postępowanie eksperta może się też wiązać z rażącymi brakami w jego świadomości metodologicznej oraz z bezkrytycznym stosunkiem do narzędzia, które czasami tylko z nazwy jest „testem psychologicznym”. Zatem bycie psychologiem oznacza nie tylko dobrą orientację w różnych formach pomocy psychologicznej, ale oznacza – a osobiście uważam, że przede wszystkim – dobre opanowanie warsztatu badawczego psychologii (z psychometrią włącznie!). Nieumiejętne postępowanie diagnostyczne, które prowadzić będzie do obciążonych diagnoz jest w dużym stopniu pochodną złego wykształcenia psychologa w zakresie teorii psychologicznych oraz metodologii i psychometrii.



Rycina 1. Triada: TEORIA – TEST – PRAKTYKA

¹⁰ Por. R. Rosenthal (1996): „Nikogo nie dziwi fakt, że badania narażające uczestników na niebezpieczeństwo budzą sprzeciw moralny. Okazuje się jednak, że wątpliwości natury etycznej mogą wiązać się również z przeprowadzaniem badań całkowicie bezpiecznych. Ich wątpliwa wartość etyczna wynika bowiem z braków i niedociągnięć wykorzystanych w nich planów badawczych. (...) Etyczne uzasadnienie przeprowadzania jakichkolwiek badań jest tym mniejsze, im gorsza jest jakość wykorzystywanych w nich planów badawczych, im gorsza jest jakość przeprowadzanej analizy wyników i im gorszy jest sposób przedstawiania rezultatów badań”.

LITERATURA CYTOWANA

- AERA/APA/NCME, American Educational Research Association, American Psychological Association, National Council on Measurement in Education (1999). *Standards for educational and psychological testing* (wyd. 5). Washington, DC: Author (polskie wydanie w przygotowaniu – Gdańskie Wydawnictwo Psychologiczne, tł. E. Hornowska).
- APA, American Psychological Association (1998). *Rights and responsibilities of test takers: Guidelines and expectations*. Uzyskane: 04. 11. 2005 r. www.apa.org/science/ttrr.html
- APA, American Psychological Association (2000). *Report of the Task Force on Test User Qualifications. Practice and Science Directorates*. Uzyskane: 04. 11. 2005 r. <http://www.apa.org/science/tuq.pdf>
- APA, American Psychological Association (2002). Ethical principles of psychologist and code of conduct. *American Psychologist*, 57, 1060–1073.
- Boake, C. (2002). From the Binet-Simon to the Wechsler-Bellevue: Tracing the history of intelligence testing. *Journal of Clinical and Experimental Neuropsychology*, 24, 383–405.
- Brzeziński, J. (2000). *Badania eksperymentalne w psychologii i pedagogice*. Warszawa: Wydawnictwo Naukowe „Scholar”.
- Brzeziński, J. (2003a). Etyczny kontekst ewaluacji jakości kształcenia w szkołach wyższych. W: J. Brzeziński, A. Eliasz (red.) *Ocena wewnętrzna i zewnętrzna jakości kształcenia w szkołach wyższych* (s. 19–40). Warszawa: Wyd. Nauk. SWPS „Academica”.
- Brzeziński, J. (2003b). Jak to z pomiarem testowym IQ było? W: Z. Piskorz, T. Zaleśkiewicz (red.) *Psychologia umysłu* (s. 96–111). Gdańsk: Gdańskie Wydawnictwo Psychologiczne.
- Brzeziński, J. (2005a). *Metodologia badań psychologicznych* (wyd. 5). Warszawa: Wydawnictwo Naukowe PWN.
- Brzeziński, J. (2005b). Psychologiczna diagnoza zdrowia i zaburzeń z perspektywy metodologii badań psychologicznych. W: H. Sęk (red.) *Psychologia kliniczna* (tom 1, s. 153–180). Warszawa: Wyd. Nauk. PWN.
- Brzeziński, J. (red.) (2005c). *Trafność i rzetelność testów psychologicznych. Wybór tekstów*. Gdańsk: Gdańskie Wydawnictwo Psychologiczne.
- Brzeziński, J., Gaul, M., Hornowska, E., Machowski, A., Zakrzewska, M. (1996). *Skala Inteligencji D. Wechslera dla Dorosłych. Wersja zrewidowana. WAIS-R(PL). Podręcznik*. Warszawa: Pracownia Testów Psychologicznych Polskiego Towarzystwa Psychologicznego.
- Campbell, D. T., Fiske, D. W. (1959/2005). Analiza trafności zbieżnej i różnicowej za pomocą macierzy wielu cech – wielu metod. W: J. Brzeziński (red.) *Trafność i rzetelność testów psychologicznych. Wybór tekstów* (s. 431–459). Gdańsk: Gdańskie Wydawnictwo Psychologiczne.
- Cronbach, L. J., Meehl, P. (1955/2005). Trafność teoretyczna testów psychologicznych. W: J. Brzeziński (red.) *Trafność i rzetelność testów psychologicznych. Wybór tekstów* (s. 403–430). Gdańsk: Gdańskie Wydawnictwo Psychologiczne.
- Cronbach, L. J., Rajaratnam, L., Nanda, H., Gleser, G. C. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York, NY: J. Wiley.
- Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.

- Hornowska, E. (1998). Skala Inteligencji Davida Wechslera w świetle teorii inteligencji. W: J. Brzeziński, E. Hornowska (red.) *Skala Inteligencji WAIS-R Wechslera. Polska adaptacja, standaryzacja, normalizacja i wykorzystanie w diagnostyce psychologicznej* (wyd. 2, s. 21–48). Warszawa: Wyd. Nauk. PWN.
- Hulin, Ch. L., Drasgow, F., Parsons (1983). Wprowadzenie do teorii odpowiedzi na pozycje testu. W: J. Brzeziński (red.) *Trafność i rzetelność testów psychologicznych. Wybór tekstów* (s. 213–271). Gdańsk: Gdańskie Wydawnictwo Psychologiczne.
- ITC, International Test Commission. (2000). *International guidelines for test-use*. Uzyskane: 04. 11. 2005 r. http://www.intestcom.org/test_use_full.htm#The%20Guidelines
- Kaufman, A. S., Lichtenberger, E. O. (1999). *Essentials of WAIS-III assessment*. New York: Wiley.
- Nowakowska M. (1975). *Psychologia ilościowa z elementami naukometrii*. Warszawa: PWN.
- Lord F. M., Novick M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Messick, S. (2005/1980). Trafność testu a etyka oceny (diagnozy). W: J. Brzeziński (red.) *Trafność i rzetelność testów psychologicznych. Wybór tekstów* (s. 460–485). Gdańsk: Gdańskie Wydawnictwo Psychologiczne.
- PTP, Polskie Towarzystwo Psychologiczne (1994). *Kodeks etyczno-zawodowy psychologa*. Warszawa: Polskie Towarzystwo Psychologiczne.
- Rosenthala, R. (1996). Nauka a etyka w przeprowadzaniu badań psychologicznych oraz analizowaniu i przedstawianiu ich wyników. *Czasopismo Psychologiczne*, 2, 37–46.
- Spearman, Ch. E. (1927). *The abilities of man*. New York: Macmillan.
- Westen, D., Rosenthal, R. (2003). Quantifying construct validity: Two simple measures. *Journal of Personality and Social Psychology*, 84, 608–618.
- Winer, B. J., Brown, D. R., Michels, K. M. (1991). *Statistical principles in experimental design* (wyd. 3). New York: McGraw-Hill.