KONTROWERSJE WOKÓŁ POJĘCIA TRAFNOŚCI

JOLANTA **RYTEL***

Instytut Psychologii, Uniwersytet Kardynała Stefana Wyszyńskiego

STRESZCZENIE

W artykule zrelacjonowano zmieniające się na przestrzeni ponad 100 lat koncepcje pojęcia trafności pomiaru testowego. Aktualnie pojęcie trafności odnosi się do stopnia, w jakim dane empiryczne oraz teoria uzasadniają interpretację wyników testowych w zakładanym kierunku (American Educational Research Association, American Psychological Association, National Council on Measurement in Education [AERA, APA, NCME], 2007, s. 31). Przedstawiono 5 podstawowych źródeł danych dotyczących trafności oraz problemy związane z integracją dowodów na rzecz trafności w spójną argumentację. Podkreślono użyteczność zaproponowanego przez Kane'a podejścia do walidacji opartego na argumentacji, odwołującego się do logiki nieformalnej i struktury argumentu wprowadzonej przez Toulmina. Omówiono także różnice stanowisk zajmowanych przez badaczy w odniesieniu do 2 podstawowych kwestii: czemu przysługuje trafność i jaki jest właściwy sposób jej ustalania?

SŁOWA KLUCZOWE

TRAFNOŚĆ
POMIAR PSYCHOLOGICZNY
STANDARDY STOSOWANIA TESTÓW PSYCHOLOGICZNYCH

* ADRES DO KORESPONDENCJI: j.rytel@uksw.edu.pl

CONTROVERSIES OVER THE CONCEPT OF VALIDITY

ABSTRACT

In the article the key changes which have occurred in conceptualization of validity are considered. Validity is currently defined in term of the degree to which a proposed interpretation of test scores is justified by evidence and theory (American Educational Research Association, American Psychological Association, National Council on Measurement in Education [AERA, APA, NCME], 2007, p. 31). Five types of validity evidence are described, and problems with the integrations of various strands of evidence in sound validity argument are discussed. Usefulness of Kane's argument-based approach to validity referring to informal logic and the structure of the argument introduced by Toulmin is stressed in the article. There is also emphasized the lack of uncontroversial definition of validity.

KEYWORDS

validity, psychological measurement, testing standards

Rytel, J. (2021). Kontrowersje wokół pojęcia trafności. *Studia Psychologica: Theoria et Praxis*, *21*(1), 49–63. https://doi.org/10.21697/sp.2021.21.1.03

Przez ostatnie ponad sto lat poglądy na temat trafności zmieniały się kilkakrotnie, jednakże stanowisko, że sama trafność wyróżnia się wśród innych pojęć psychometrycznych, pozostało niezmienne. Trafność od zawsze uważano w psychometrii za własność najbardziej fundamentalną i najważniejszą (Angoff, 1988). Zgodnie z definicją podaną w *Standardach dla testów stosowanych w psychologii i pedagogice* „pojęcie trafności odnosi się do stopnia, w jakim dane empiryczne oraz teoria uzasadniają interpretację wyników testowych w zakładanym kierunku. Trafność jest zatem najbardziej podstawową kategorią w procesie tworzenia i oceny testu" (American Educational Research Association, American Psychological Association, National Council on Measurement in Education [AERA, APA, NCME], 2007, s. 31). Podstawowa rola pojęcia trafności jest niekwestionowana, jednakże określenie, czym jest trafność, nadal wzbudza kontrowersje.

MODEL TRAFNOŚCI OPARTY NA KRYTERIUM

Około 1915 roku, choć nie istniała jeszcze formalna definicja trafności, zaczął się kształtować model, który Kane (2001) określa jako *model trafności oparty na kryterium*. Z pewnością wpływ na to miało opublikowanie przez Pearsona w roku 1896 pracy dotyczącej współczynnika korelacji. Był to ekscytujący nowy wskaźnik statystyczny, który można było zastosować bezpośrednio, rozwiązując kwestię tego, jak dobrze wyniki poddanych ocenie osób wiążą się z innymi przejawami testowanej własności (Sireci, 2009). Pogląd, że test jest trafny dla wszystkiego, z czym koreluje (Guilford, 1946), utrzymywał się do połowy XX wieku. Choć większość badań nad trafnością prowadzonych w latach 30. i 40. XX wieku koncentrowała się na ustalaniu trafności prognostycznej, pojawił się też inny jej rodzaj – trafność diagnostyczna. Pojęcie trafności diagnostycznej związane jest ze stopniem, w jakim wyniki testowe umożliwiają ocenę kryterium (pewnej zmiennej pozatestowej, której dotyczy wnioskowanie na podstawie wyników testu), a oba pomiary: pomiar testowy i pomiar kryterium przeprowadzane są w zbliżonym bądź w tym samym czasie, a więc ocena dotyczy stanu aktualnego. W przypadku trafności prognostycznej pomiar kryterium dokonywany jest w czasie późniejszym niż pomiar testowy, zatem ocena odnosi się do możliwości przewidywania stanu przyszłego. Testy były wykorzystywane przede wszystkim do celów selekcji i klasyfikacji. Nie dziwi zatem, że te dwie formy trafności odzwierciedlały cel, w jakim testy były powszechnie w tym czasie przeprowadzane. Celem konstrukcji testu było przewidywanie pewnych zewnętrznych zachowań (Lissitz, Samuelsen, 2007). Do ewolucji poglądów na temat trafności niewątpliwie przyczyniło się również opracowanie przez Spearmana metody analizy czynnikowej. Zwolennikiem wykorzystania tej metody do szacowania trafności był Guilford (1946), postulujący dwa rodzaje trafności: praktyczną – odnoszącą się do korelacji między wynikami testu i miarami kryterium – oraz czynnikową.

> Trafność czynnikową testu charakteryzują jego ładunki na znaczeniowo zinterpretowanych wspólnych, istotnych czynnikach. To jest ten rodzaj trafności, o który rzeczywiście chodzi, gdy pada pytanie: Czy test mierzy to, do mierzenia czego został opracowany? Stosowniejszym pytaniem winno być: Co mierzy test? Przewiduję, że przyjdzie czas, gdy od każdego autora będzie się oczekiwać podania informacji na temat kompozycji czynnikowej jego testu (Guilford, 1946, s. 428 – tłum. J. R.).

MODEL TRAFNOŚCI OPARTY NA TREŚCI

Kłopoty z modelem trafności opartym na kryterium dotyczą konieczności odwołania się do dobrze zdefiniowanego i trafnego kryterium, które nie zawsze jest łatwo osiągalne. Pojawienie się testów osiągnięć, dla których nie było lepszego kryterium niż sam test, przyczyniło się do powstania kolejnego modelu trafności – modelu opartego na treści (Kane, 2008). Trafność treściowa jest wbudowana w test od samego początku przez wybór odpowiednich pozycji (Anastasi, Urbina, 1999). Pojęcie trafności treściowej dotyczy oceny stopnia, w jakim treść pozycji testowych reprezentuje dziedzinę będącą przedmiotem pomiaru. Spotkało się ono z krytyką (por. Kane, 2006a; Sireci, 1998), która wskazywała na subiektywizm ocen wydawanych na podstawie analizowania treści pozycji testowych, a także na możliwość obciążenia tych ocen silną tendencją do potwierdzania, jako że analiza treści pozycji zwykle przeprowadzana jest w trakcie tworzenia narzędzia bądź wkrótce po jego opracowaniu przez osoby zaangażowane w ten proces. Ponadto – co podkreślał Messick (1989) – ocena trafności treściowej, jako niezwiązana z wynikami testu, nie może stanowić uzasadnienia dla interpretacji wyników testu. Guion (1978) wręcz odrzucał pojęcie trafności treściowej, przestrzegając, że rozważności przy konstruowaniu testu czy wyborze próbki z uniwersum pozycji nie powinno się mylić z trafnością treściową. Warto zauważyć, że problem użyteczności szacowania trafności treściowej nadal budzi kontrowersje wśród badaczy, o czym świadczy dyskusja, jaka odbyła się na łamach pisma *Industrial and Organizational Psychology*, zapoczątkowana artykułem Murphy'ego (2009) zatytułowanym prowokacyjnie: *Content validation is useful for many things, but validity isn't one of them*.

MODEL TRAFNOŚCI TEORETYCZNEJ

> Trafność teoretyczna badana jest zwykle, gdy testujący nie dysponuje żadną dokładnie określoną miarą kryterium cechy, którą jest zainteresowany, i musi wykorzystać miary pośrednie. W takim przypadku właśnie cecha czy własność leżąca u podstaw testu ma zasadnicze znaczenie, a nie zachowania, które angażuje wykonanie testu czy wyniki otrzymane w zakresie kryteriów. (Cronbach, Meehl, 1955, s. 282 – tłum. J. R.)

Cytat ten, zamieszczony w przełomowym artykule autorstwa Cronbacha i Meehla (1955), zatytułowanym *Construct validity in psychological test*, pochodzi z dokumentu *Technical recommendations for psychological tests and diagnostic techniques*. W dokumencie tym opublikowano wnioski uzgodnione przez Komitet Testów Psychologicznych Amerykańskiego Towarzystwa Psychologicznego, który w latach 1950–1954 starał się określić, jakie właściwości testu winny być przedmiotem analizy, zanim zostanie on opublikowany. Cronbach i Meehl (1955), wyjaśniając pojęcie trafności teoretycznej, stwierdzają: „Z określeniem trafności teoretycznej mamy do czynienia, ilekroć test ma być interpretowany jako miara pewnego atrybutu lub własności, która nie jest «zdefiniowana operacyjnie». Problem, przed którym stoi badacz, można sformułować następująco: Jakie konstrukty wyjaśniają wariancję poziomu wykonania testu?" (s. 282 – tłum. J. R.).

W niedługim czasie Campbell i Fiske (1959) opracowali metodę badania trafności teoretycznej za pomocą macierzy wielu cech-wielu metod. Metoda ta dotyczyła przede wszystkim adekwatności testów jako miar konstruktu, a nie adekwatności konstruktów określanej przez potwierdzenie przewidywanych teoretycznie związków z miarami innych konstruktów. „Sądzimy, że przed przystąpieniem do testowania związków między daną cechą

a innymi cechami należy nabrać zaufania do miary tej cechy. To zaufanie mogą podtrzymać dowody trafności zbieżnej i różnicowej" (Campbell, Fiske, 1959, s. 100 – tłum. J. R.).

Osadzając znaczenie konstruktu w sieci nomologicznej, Cronbach i Meehl (1955) wskazują na konieczność sformułowania teorii danej cechy: „Dowiedzenie się więcej" o konstrukcie teoretycznym jest ich zdaniem kwestią opracowania sieci nomologicznej, w której konstrukt występuje, bądź zwiększenia precyzji definiowania jego składników. Wskazują tym samym na konieczność zdefiniowania konstruktu przed przystąpieniem do ustalania trafności testu. Stwierdzając natomiast, że badanie trafności teoretycznej testu nie różni się zasadniczo od ogólnych naukowych procedur tworzenia i potwierdzania teorii, autorzy akcentują, że metodologia ustalania trafności nie może sprowadzać się do podania wartości jednego współczynnika (Cronbach, Meehl, 1955).

Trafność ujmowana z perspektywy teorii przestaje być trafnością związaną z samym testem, a staje się trafnością dotyczącą inferencji na temat znaczenia wyników testu czy też trafnością dotyczącą nadawanej im przez badacza interpretacji. „Nie ustala się trafności testu, lecz trafność interpretacji danych uzyskanych za pomocą specyficznej procedury" (Cronbach, 1971, s. 447 – tłum. J. R.), a ustalanie trafności odnosi się do wszystkich interpretacji wyników testowych. Różne zastosowania danego testu wiążą się z innymi interpretacjami jego wyników, a każda z tych interpretacji wymaga uzasadnienia. „Ponieważ każdą interpretację charakteryzuje właściwy dla niej stopień trafności, nigdy nie da się wyprowadzić prostego wniosku, że dany test «jest trafny»" (Cronbach, 1971, s. 447 – tłum. J. R.).

Trafność teoretyczna stanowiła kolejny typ trafności, jednakże Cronbach i Meehl (1995) wyrażali przekonanie, że określenie, które konstrukty psychologiczne wyjaśniają wykonanie testu, jest pożądane dla niemal każdego testu. Podkreślali też, iż z trafnością teoretyczną wiąże się wiele typów dowodów, łącznie z: trafnością treściową, korelacjami między pozycjami, korelacjami między testami, korelacjami między testem i «kryterium», badaniami stabilności w czasie oraz stabilności w różnych warunkach eksperymentalnych (Cronbach, Meehl, 1995). Tym samym pojęcie trafności teoretycznej stało się najlepszym kandydatem do objęcia nim wszystkich dotychczasowych rodzajów trafności.

W KIERUNKU UJEDNOLICENIA POJĘCIA TRAFNOŚCI

To, że pojęcie trafności teoretycznej stanowi doskonałą podstawę do ujednolicenia wszystkich rodzajów trafności, zostało rozpoznane przez badaczy wcześnie. Już w 1957 roku Loevinger twierdziła, że trafność teoretyczna stanowi całościowe ujęcie trafności z naukowego punktu widzenia. Loevinger (1957) wyróżniła trzy aspekty trafności teoretycznej powiązane z trzema stadiami procesu konstrukcji testu. Aspekt esencjalny związany z tworzeniem puli pozycji testowych dotyczy zakresu, w jakim treść pozycji można wyjaśnić w terminach mierzonej cechy i w kontekście teorii psychologicznej. Aspekt strukturalny wiąże się z analizą struktury wewnętrznej puli pozycji i ich selekcją w celu opracowania procedury obliczania wyników, odnosząc się do odpowiedniości czy też – używając terminu Loevinger – wierności modelu strukturalnego w stosunku do charakterystyk strukturalnych pozatestowych manifestacji cechy i samej struktury utworzonej przez pozycje testowe. Aspekt zewnętrzny obejmuje analizę związków wyników testowych z miarami zachowań pozatestowych i wynikami innych testów.

Trzy dekady po rozważaniach Loevinger podobny pogląd wyraził sam Cronbach (1988): „licząca 30 lat idea trzech typów trafności, rozłącznych, choć może sobie równych, to idea, której czas minął. Większość teoretyków trafności uznała, że trafność treściowa i kryterialna nie są niczym więcej niż żyłami w kablu argumentacji na rzecz trafności" (s. 4 – tłum. J. R.). Postulując ujmowanie procesu walidacji jako konstruowania argumentacji

na rzecz trafności, Cronbach podkreślał, że argumentacja taka kierowana jest do różnorodnego i potencjalnie krytycznego audytorium i dlatego „musi łączyć pojęcia, dowody empiryczne, konsekwencje społeczne i indywidualne oraz wartości" (s. 4 – tłum. J. R.).

Wymiar konsekwencji społecznych związanych z interpretacją wyników testowych i ich wykorzystaniem został wprowadzony przez Messicka (1989; 1995) jako jeden z aspektów trafności teoretycznej stanowiącej podstawę ujednoliconego modelu ujmowania trafności. Na pozostałe aspekty trafności teoretycznej składają się wskazane już przez Loevinger (1957): esencjalny, strukturalny i zewnętrzny oraz aspekt treściowy, obejmujący dowody na rzecz treściowej ważności, reprezentatywności i jakości technicznej, a także aspekt ogólności, odnoszący się do stopnia generalizowalności wyników testowych i ich interpretacji. Konsekwencje mogą być pozytywne bądź negatywne, a zgromadzenie dowodów na rzecz obu tych rodzajów jest równie istotne. Szczególnie ważne – zwłaszcza dla konsekwencji potencjalnie niekorzystnych – jest ustalenie, czy nie wynikają one z niedoreprezentowania konstruktu lub wariancji niezwiązanej z konstruktem. W pierwszym przypadku test nie obejmuje ważnych wymiarów konstruktu, tym samym interpretacja wyników zostaje zawężona, w drugim zaś na wyniki testowe wpływają czynniki niezwiązane z konstruktem, a powiązane z innymi konstruktami, samym testem lub warunkami badania. W takim ujęciu pojęcie trafności dotyczy nie tylko interpretacji wyników testowych, lecz także pożądanych i niepożądanych konsekwencji testowania: „ustalanie trafności jest empiryczną oceną znaczenia i konsekwencji pomiaru. Termin *ocena empiryczna* został użyty dla podkreślenia, że proces walidacji ma zarówno naukowy, jak i retoryczny charakter, wymaga zatem i dowodów, i argumentów" (Messick, 1995, s. 747 – tłum. J. R.).

Jednolite ujęcie trafności przez Messicka (1989) stało się podstawą dla współczesnego jej pojmowania: „trafność to całościowy osąd stopnia, w jakim dowody empiryczne i racjonalne uzasadnienia natury teoretycznej potwierdzają adekwatność i poprawność interpretacji i działań podejmowanych na podstawie wyników testu lub innych sposobów oceny". Messick (1995) dobitnie podkreśla, że:

> (…) trafność nie jest własnością testu lub oceny jako takiej, a stanowi ona znaczenie wyników testowych. Wyniki te są funkcją nie tylko pozycji testowych lub warunków bodźcowych, ale również odpowiadających na nie osób badanych i kontekstu przeprowadzania oceny. Tym, co powinno być trafne, jest znaczenie lub interpretacja wyniku, podobnie jak każda implikacja dla działań, jaką to znaczenie za sobą pociąga (Cronbach, 1971). Problem zasięgu, w jakim znaczenie wyniku i implikacje działań obejmują osoby oraz układy warunków lub konteksty, stanowi zawsze aktualne i odwieczne pytanie empiryczne. Trafność jest własnością ewoluującą, a proces jej ustalania ma charakter ciągły przede wszystkim z tego właśnie powodu. (s. 741 – tłum. J. R).

Główne zmiany, jakie zaszły w rozumieniu pojęcia trafności, można podsumować w następujący sposób:

1. Rozumienie trafności zmieniło się od praktycznego, empirycznego sposobu jej ujmowania, a tym samym wprowadzania różnych rodzajów trafności, do osadzenia pojęcia trafności w teorii, dzięki czemu stało się możliwe jednolite jej ujęcie. Podstaw do zintegrowania różnych rodzajów trafności dostarczyło pojęcie trafności teoretycznej. Współcześnie trafność traktowana jest jako pojęcie jednolite (trafność teoretyczna), mające kilka aspektów. „Dane pochodzące z poszczególnych źródeł mogą potwierdzać różne aspekty trafności, ale nie reprezentują różnych rodzajów trafności. Trafność jest pojęciem spójnym" (AERA, APA, NCME, 2007, s. 35).

2. Trafność nie jest własnością testu. Jest trafnością uzasadnienia proponowanej interpretacji i wykorzystania wyników testowych dostarczanego z punktu widzenia teorii leżącej u podstaw testu oraz zgromadzonych dowodów empirycznych. Tym samym twierdzenia

na temat trafności nie mają charakteru dychotomicznego (trafny vs nietrafny), ale ujmowane są jako kontinuum. Trafność to „stopień, w jakim wszystkie kumulujące się dane potwierdzają zamierzoną interpretację wyników testowych" (AERA, APA, NCME, 2007, s. 35). Zmianę tę doskonale ujmuje stwierdzenie Messicka (1980): „trafność jest oceną dowodów, osądem, a nie samoistnym bytem" (s. 1020).

3. Trafność powinna być ustalona dla każdego sposobu interpretacji i wykorzystania wyników testowych.

Twierdzenia dotyczące trafności powinny odnosić się do konkretnych sposobów interpretacji czy zastosowania wyników testowych. Posługiwanie się sformułowaniem: *trafność wyników testowych* nie jest poprawne. Żaden test nie jest trafny w wypadku wszystkich celów i w każdej sytuacji. Wszystkie rekomendowane zastosowania czy interpretacje testu wymagają potwierdzenia trafności (AERA, APA, NCME, 2007, s. 45).

4. Ustalanie trafności wymaga uwzględnienia danych pochodzących z różnych źródeł. „Obejmuje to dane wynikające z analizy procesu tworzenia testu, dane dotyczące adekwatnej rzetelności testu, właściwej procedury badania testem i procedury obliczania wyników testowych, procedury skalowania i wyrównywania wyników oraz analizy testu z punktu widzenia potencjalnej stronniczości testu (…)" (AERA, APA, NCME, 2007, s. 44). Dane te służą jako podstawa zbudowania wiarygodnej naukowo argumentacji na rzecz trafności przyjmowanej interpretacji wyników.

5. Ustalanie trafności nie jest jednorazowym przedsięwzięciem, stanowi ciągły proces. Wymaga ono starannego zaplanowania i zrealizowania programu badawczego, a nie przeprowadzenia pojedynczego badania. „Proces walidacji obejmuje ciągłe zbieranie danych w celu dostarczenia mocnych podstaw naukowych proponowanej interpretacji wyników testowych" (AERA, APA, NCME, 2007, s. 31).

USTALANIE TRAFNOŚCI JAKO PROCES INTEGRACJI DANYCH POCHODZĄCYCH Z RÓŻNYCH ŹRÓDEŁ

W *Standardach dla testów stosowanych w psychologii i pedagogice* (AERA, APA, NCME, 2007) wymieniono pięć podstawowych źródeł danych dotyczących trafności.

1. Zgromadzenie danych opartych na treści testu (włączając poruszane tematy, sformułowania słowne, format pozycji testowych, zadań lub pytań tworzących test oraz procedury badania i obliczania wyników) umożliwia ocenę stopnia, w jakim treść testu reprezentuje dziedzinę będącą przedmiotem pomiaru. Dane te mogą pochodzić z analiz logicznych czy empirycznych dotyczących stopnia reprezentatywności treści testu oraz stopnia, w jakim owa treść wiąże się z proponowaną interpretacją wyników. Ważnym źródłem tych danych może być przeprowadzona przez ekspertów ocena stopnia, w jakim pozycje czy części testu odpowiadają definicji mierzonego konstruktu. Oceny ekspertów mogą być wykorzystane do określenia stopnia adekwatności, ważności i precyzji sformułowania pozycji testowych oraz zakresu, w jakim niedoreprezentowanie konstruktu lub włączenie niepowiązanych z nim treści może dostarczać nieuczciwej przewagi jakiejś grupie badanych osób.

2. Informacji o stopniu zgodności między konstruktem a zachowaniami czy odpowiedziami pojawiającymi się w procesie rozwiązywania testu mogą dostarczyć dane oparte na analizie procesu udzielania odpowiedzi. Dane te można uzyskać, śledząc proces konstruowania odpowiedzi czy przeprowadzając wywiad z osobami badanymi i analizując odpowiedzi indywidualne na pytania dotyczące wykorzystywanych strategii, zachowań czy zasad odpowiadania na konkretne pozycje testu. Może ich też dostarczyć analiza procesu rozwiązywania testu uwzględniająca podobieństwa i różnice w sposobach odpowiadania, przeprowadzona w różnych grupach. Analiza taka przyczynia się do określenia zakresu,

w jakim umiejętności mające niewielkie znaczenie lub nieistotne z punktu widzenia konstruktu wpływają na proces odpowiadania. Proces walidacji może obejmować także badanie sposobów gromadzenia, rejestrowania i interpretowania danych przez ekspertów, obserwatorów i osoby przeprowadzające wywiady. W takim przypadku istotna jest analiza poprawności tych procesów i ustalenie zgodności przyjmowanych kryteriów bądź wydawanych ocen z planowaną interpretacją wyników testowych.

3. Dane wynikające z analizy struktury wewnętrznej testu pozwalają na ustalenie stopnia, w jakim relacje między pozycjami czy składowymi testu odpowiadają zdefiniowanemu konstruktowi. Źródło takich danych może stanowić analiza związków między pozycjami testu oraz badanie jego struktury czynnikowej (w szczególności za pomocą konfirmacyjnej analizy czynnikowej). Analiza pozycji testowych może ujawnić te z nich, które funkcjonują odmiennie w różnych (np. ze względu na płeć lub warunki środowiskowe czy kulturowe) grupach badanych, nieróżniących się jednak ze względu na mierzoną wielkość. Zróżnicowane funkcjonowanie pozycji (*differential item functioning* – DIF) związane jest z wystąpieniem nieoczekiwanej różnicy między grupami osób badanych, które przypuszczalnie powinny być porównywalne co do atrybutu mierzonego przez pozycję i test, w którym ona występuje (Dorans, Holland, 1993). Taka pozycja może mierzyć coś innego niż pozostała część testu (np. składniki niezwiązane z konstruktem) lub też może mierzyć to samo, ale z różnym stopniem precyzji w różnych grupach. Doskonały przegląd zarówno podejść do problemu zróżnicowanego funkcjonowania pozycji, jak i metod służących identyfikowaniu tego zjawiska zawiera artykuł Zumbo (2007).

4. Kolejne źródło danych walidacyjnych stanowią dane oparte na analizie związków z innymi zmiennymi. Interpretacja takich danych umożliwia określenie rodzaju i zakresu zależności między wynikami testu a innymi zmiennymi zewnętrznymi oraz ustalenie stopnia, w jakim te zależności są spójne z konstruktem leżącym u podstaw proponowanej interpretacji wyników. W skład tych danych wchodzą dane walidacyjne zbieżne, dotyczące związków (silnych) między wynikami testu a innymi miarami odzwierciedlającymi podobne konstrukty oraz dane walidacyjne o charakterze rozbieżnym, odnoszące się do związków (słabych bądź ich braku) z miarami reprezentującymi inne konstrukty. Dane te mogą pochodzić z analizy związków między testem a kryterium, mierzonym w zbliżonym bądź tym samym czasie (dane diagnostyczne) lub też mierzonym w czasie późniejszym niż pomiar testowy (dane prognostyczne). Badania nad efektywnością decyzji selekcyjnych, klasyfikacyjnych czy lokacyjnych także stanowią źródło takich danych. Mogą ich dostarczyć badania różnic międzygrupowych, w których sprawdzana jest, sformułowana na podstawie teorii mierzonej wielkości, hipoteza dotycząca wystąpienia oczekiwanych różnic w wynikach testu między badanymi grupami lub badania eksperymentalne, których celem jest wykazanie wystąpienia zmiany w wynikach testu, przewidywanej w następstwie wprowadzonego oddziaływania eksperymentalnego.

Rezultaty badań wykorzystujących technikę metaanalizy mogą posłużyć do uogólnienia danych dotyczących trafności. Zanim pojawiła się przełomowa praca Schmidta i Huntera (1977) wprowadzająca metaanalizę do badań nad generalizacją trafności, sądzono, że bardzo duże zróżnicowanie współczynników trafności testów uzdolnień stosowanych do selekcji personelu, dla których kryterium stanowił poziom wykonania pracy, związane jest ze specyfiką sytuacji badania lub specyfiką danej pracy. Przekonanie o *sytuacyjnej specyficzności* (Schmidt, Hunter, 1998) powodowało konieczność każdorazowego ustalania trafności w nowej sytuacji wykorzystania testów, poważnie ograniczając ich użyteczność. Okazało się, że większość tych różnic jest artefaktem wynikającym z błędu próby, zazwyczaj niewielkich rozmiarów (od 40 do 70 osób). To, na ile dane walidacyjne uzyskane w konkretnej sytuacji badawczej mogą być uogólnione na nowe sytuacje, stanowi bardzo ważny problem, zwłaszcza w obszarze edukacji i pracy. Warto zaznaczyć, że nie zawsze jest możliwe zgromadzenie

danych z przeprowadzonych już badań walidacyjnych adekwatnych do sytuacji, na którą mają być uogólniane. Wtedy przeprowadzenie badań lokalnych jest nieodzowne.

5. Ostatni rodzaj danych to dane oparte na konsekwencjach testowania. Jeżeli zamierzony kierunek interpretacji wyników testu wiąże się z odniesieniem określonych korzyści związanych z jego stosowaniem (np. ulepszenie programów nauczania, efektywności procesu rekrutacji pracowników, skuteczności programów terapeutycznych), to w badaniach walidacyjnych należy wykazać, że dane korzyści uda się uzyskać. Dane dotyczące konsekwencji testowania, zarówno pożądanych, jak i niepożądanych, winny być odróżniane od danych, które mogą wpłynąć na decyzje dotyczące polityki społecznej, jednakże z trafnością niezwiązane. Jest to szczególnie ważne, gdy badanie różnych grup osób ma określone konsekwencje społeczne. I tak np. gdy w przypadku zastosowania testu jako narzędzia selekcyjnego na gruncie zawodowym czy edukacyjnym otrzymana różnica wyników między grupami stanowi odzwierciedlenie zróżnicowanego rozkładu umiejętności, które test ma mierzyć, to nie świadczy ona o braku trafności wnioskowania. Może jednak wynikać z wrażliwości testu na niektóre cechy osób badanych nienależące do zakresu konstruktu bądź z jego niedoreprezentowania, wtedy bezpośrednio wiąże się z trafnością. Jak podkreślono w *Standardach dla testów…* (AERA, APA, NCME, 2007):

> Chociaż informacje na temat konsekwencji testowania mogą wpłynąć na decyzje o zastosowaniu testu, to same konsekwencje jako takie nie mają wpływu na trafność zamierzonej interpretacji wyników. Sądy dotyczące trafności lub jej braku, formułowane w świetle konsekwencji testowania, powinny zależeć raczej od racjonalnego wyjaśnienia ich źródeł. (s. 42)

Żaden rodzaj danych nie jest preferowany na tle innych, a zadaniem badacza jest synteza informacji pochodzących z różnych źródeł i przedstawienie ich w postaci spójnej argumentacji na rzecz zamierzonej interpretacji wyników testu i sposobów jego wykorzystania. Problem polega na tym, w jaki sposób to zrobić.

PROBLEMY ZE WSPÓŁCZESNĄ KONCEPCJĄ TRAFNOŚCI

Najbardziej kontrowersyjnym z wymienionych wyżej źródeł danych walidacyjnych jest ostatnie z nich. Dyskusje nad zasadnością włączania danych opartych na konsekwencjach testowania w proces walidacyjny prowadzone są od dawna. Już w 1997 roku poświęcono tej problematyce cały numer czasopisma *Educational Measurement: Issues and Practice* (t. 16, nr 2), podobnie stało się w roku kolejnym (t. 17, nr 2). Od tego czasu, podobnie jak od czasu opublikowania w roku 1999 przełożonego na język polski (w roku 2007) wydania *Standardów dla testów…*, upłynęło ponad 20 lat, warto zatem przyjrzeć się roli, jaką w badaniach walidacyjnych odgrywa ten aspekt trafności.

Cizek, Rosenberg i Koons (2008) dokonali przeglądu informacji na temat aspektów trafności relacjonowanych w odniesieniu do 283 testów psychologicznych i pedagogicznych opublikowanych w *Mental Measurements Yearbook* (Spies, Plake, 2005). Okazało się, że informacje dotyczące trafności rzadko były podawane w terminach zgodnych z jej współczesnym ujęciem. Podczas gdy dane dotyczące trafności teoretycznej (58% wszystkich testów), diagnostycznej (50,9%) i treściowej (48,4%) były zamieszczane stosunkowo często, to informacje na temat aspektu trafności związanego z konsekwencjami testowania pojawiły się jedynie w przypadku dwóch testów (0,7%). W kolejnym badaniu autorzy (Cizek, Bowen, Church, 2010) objęli analizą opublikowane w latach 1999–2008 numery ośmiu wiodących czasopism (takich jak *Educational and Psychological Measurement* czy *Practical Assessment, Research & Evaluation*). Spośród 2408 artykułów problematyki trafności dotyczyło 1007

(41,8%), i aspekt trafności związany z konsekwencjami nie został wymieniony w żadnym z nich. Taki sam rezultat uzyskano, analizując treść wystąpień prezentowanych na corocznych, odbywających się w latach 2006–2008, spotkaniach trzech organizacji sponsorujących wydawanie *Standardów…*: American Educational Research Association, American Psychological Association i National Council on Measurement in Education. Autorzy konkludują (Cizek, Rosenberg, Koons, 2008), że aspekt trafności dotyczący konsekwencji testowania nie jest przez badaczy uwzględniany, ponieważ aspektu tego po prostu nie da się w logiczny sposób połączyć z pozostałymi. Tym samym stanowi on usterkę we współczesnej teorii trafności (Cizek i in., 2008, s. 410).

Ustalanie trafności proponowanej interpretacji wyników testu i uzasadnianie sposobów jego wykorzystania obejmujące związane z tym konsekwencje stanowią zadania, które należy od siebie oddzielić, choć niewątpliwie zgromadzenie przekonujących dowodów na rzecz trafności interpretacji wyników testu stanowi warunek konieczny, ale niewystarczający wykorzystania testu. Nie oznacza to umniejszania wagi konsekwencji testowania – wręcz przeciwnie. W związku z tym, że aspekt ten jest niemal całkowicie ignorowany w praktyce walidacyjnej, pojawia się potrzeba opracowania wskazań dotyczących gromadzenia i ewaluacji danych uzasadniających różne sposoby wykorzystania testów (Cizek, 2012, 2016, 2020). Ponadto autorzy podkreślają, że w dokonanym przez nich przeglądzie nie udało im się znaleźć ani jednego przypadku, w którym dane pochodzące z różnych źródeł zostałyby zintegrowane w spójną argumentację na rzecz (lub przeciw) trafności.

Problem z integracją dowodów na rzecz trafności w spójną argumentację sprowadza się do odpowiedzi na pytanie, czym jest argument na rzecz trafności i w jaki sposób powinien być konstruowany. Z problemem tym badacz musi się zmierzyć niezależnie od tego, czy integruje wszystkie rodzaje dowodów czy też tylko cztery (pierwsze) z nich, zważywszy na odmienny status logiczny danych opartych na konsekwencjach testowania. Użyteczne może się okazać rozwijane przez Kane'a (1992, 2004, 2006b, 2013, 2016) podejście do walidacji oparte na argumentacji (*argument-based validation*). Kane w swoich pracach odwołuje się do logiki nieformalnej i struktury argumentu wprowadzonej przez Toulmina (1958). Na strukturę argumentu składają się trzy podstawowe elementy: wynikająca z jakiejś przesłanki (*datum*) teza (*claim*) i uzasadniająca łączącą je relację gwarancja (*warrant*). Teza nie musi być wyrażona z całkowitą pewnością, może być poprzedzona jakimś wyrażeniem modalnym, np. *prawdopodobnie*, pełniącym funkcję modyfikatora (*qualifier*). Gwarancja, do której się odwołujemy, może wymagać wskazania, w jakich przypadkach stanowiących odstępstwa od reguły tezy nie da się utrzymać, czyli wymagać de facto wprowadzenia ograniczenia lub obalenia (*rebuttal*). Sama gwarancja może też wymagać uzasadnienia, określanego jako wsparcie dla niej (*backing*). Argumentacją posługujemy się w celu poparcia stwierdzeń, co do których nie mamy pewności, i czynimy to, odwołując się do stwierdzeń, względem których pewność tę posiadamy. Oceniając poprawność argumentacji, rozważamy akceptowalność wszystkich wykorzystanych przesłanek oraz to, czy stanowią one dostateczne uzasadnienie dla proponowanej tezy. W przypadku oceny poprawności argumentacji budowanej zgodnie z modelem Toulmina kluczową rolę odgrywa ocena poprawności stwierdzeń występujących w funkcji gwarancji, a tym samym – także twierdzeń funkcjonujących jako jej poparcie. W tym sensie zgodnie z poglądem Toulmina (1958) poprawność argumentacji jest zrelatywizowana do dziedziny, której dotyczy.

Proponowane przez Kane'a podejście obejmuje dwa etapy i odnosi się do dwóch rodzajów argumentów. Najpierw konstruowany jest argument dotyczący interpretacji (*interpretive argument*), obejmujący każdą interpretację i wykorzystanie wyników testu przedstawione w formie wnioskowania prowadzącego od wyników testu do nadawanej im interpretacji / sposobu ich wykorzystania oraz uzasadniającej je gwarancji. I tak np. gdy za pomocą równania regresji na podstawie wyników testu przewidujemy średnią ocen, to wyniki testu są przesłanką dla tezy, którą jest przewidywana średnia ocen, gwarancję natomiast stanowi

równanie regresji. Na drugim etapie budowany jest argument na rzecz trafności (*validity argument*), dzięki któremu możliwe jest wykazanie trafności proponowanych interpretacji i sposobów wykorzystania wyników testu przez dostarczenie stosownego wsparcia dla gwarancji, na których opierają się wnioskowania. Odwołując się do powyższego przykładu: równanie regresji pełniące funkcję gwarancji wymaga odpowiedniego wsparcia. Takiego wsparcia dla korzystania z niego w celu przewidywania średniej ocen dostarczają przeprowadzone badania empiryczne, w których równanie zostało otrzymane. W przypadku wykazania, że któreś z wnioskowań jest nieuprawnione, argument na rzecz trafności należy odrzucić.

Jak podkreśla Kane (2009):

> …konstruując argument dotyczący interpretacji, mamy do czynienia z istotnym 'coś za coś'. Wykazanie trafności prostej interpretacji wymaga stosunkowo skromnej inwestycji w zgromadzenie danych na rzecz trafności. Interpretacja ambitniejsza, związana z bardziej rozbudowaną strukturą wnioskowań i wspierających je założeń, wymaga więcej wsparcia, by osiągnąć porównywalny poziom trafności, ale ambitniejsza interpretacja może być bardziej użyteczna, jeśli zostanie uzasadniona. Jeżeli nie twierdzimy, że test będzie przewidywał przyszłe zachowanie, to nie jesteśmy zobowiązani do zgromadzenia danych odnośnie do trafności prognostycznej, jednakże powstały argument na rzecz trafności nie będzie uzasadniał takich przewidywań. (s. 48–49 – tłum. J. R.)

Chapelle, Enright i Jamieson (2010) przedstawiają strategię konstruowania argumentu na rzecz trafności zrewidowanej wersji TOEFL® (Test of English as a Foreign Language). Autorki jako jedne z pierwszych wykorzystują podejście Kane'a do syntezy danych walidacyjnych. Chapelle i in. (2010) oceniają różnice między podejściem przedstawionym w *Standardach dla testów…* (AERA, APA, NCME, 2007) i tym proponowanym przez Kane'a odnośnie do sposobu formułowania zamierzonej interpretacji wyników testu, prezentacji przeprowadzonych badań walidacyjnych, syntetyzowania wyników tych badań w argument na rzecz trafności i prób jego podważenia. W ich opinii opierająca się na argumentacji strategia walidacyjna stanowi cenne uzupełnienie rekomendacji przedstawionych w *Standardach dla testów…* (AERA, APA, NCME, 2007), dostarczając zarówno ogólną strukturę dla przeprowadzenia syntezy danych pochodzących z różnych źródeł, jak i precyzyjne wskazówki co do sposobu, w jaki należy to zrobić. Warto zaznaczyć, że wydane ostatnio *Standards for educational and psychological testing* (AERA, APA, NCME, 2014) zalecają podejście do walidacji oparte na argumentacji, obejmujące jasne sformułowanie zamierzonej interpretacji i wykorzystania wyników testów, w tym zidentyfikowanie podstawowych założeń i wniosków (argument na rzecz interpretacji / wykorzystania wyników testowych) i zgromadzenie dowodów na ich poparcie lub obalenie.

Badacze od dawna wyrażali obawy, że podnieśli pojęcie trafności teoretycznej na tak wysoki poziom, że wydaje się ono nieosiągalnym celem (Fremer, 2000). Coraz częściej pojawiają się badania, których rezultaty wskazują na rozdźwięk między teorią i praktyką walidacyjną, co stanowi poważny problem zwłaszcza w przypadku testów mających wielu interesariuszy, a ich autorzy konkludują, że teorię trafności bardzo trudno przełożyć na działania praktyczne wobec braku wskazówek odnośnie do syntetyzowania danych dotyczących trafności i ich interpretacji (por. np. Wolming, Wikström, 2010). „Teoria trafności stopniowo doszła do traktowania każdej ważnej kwestii wiążącej się z testem jako istotnej dla pojęcia trafności i dąży do integracji wszystkich tych kwestii pod jednym nagłówkiem. Jednakże postępując w taki sposób, nie służy ani teoretycznie zorientowanym psychologom, ani stosującym testy praktykom" (Borsboom, Mellenbergh, van Heerden, 2004, s. 1061 – tłum. J. R.). Borsboom (2006) stwierdza, że trafność teoretyczna działa jak czarna dziura, z której nic nie może uciec: gdy dana kwestia zostanie określona jako problem trafności teoretycznej, jego trudność uważa się za nadludzką, a rozwiązanie – za przekraczające

możliwości zwykłego śmiertelnika. Trudność to być może nie nadludzka, ale jednak bardzo poważna, co potwierdza aktualna praktyka. Dowody na rzecz trafności zazwyczaj mają charakter fragmentaryczny, a badacze nadal bardzo często, prezentując dane na temat trafności, odwołują się do klasycznych terminów trafności teoretycznej, treściowej, diagnostycznej i prognostycznej.

Kolejną konsekwencją tego stanu rzeczy są propozycje odmiennych sposobów ujmowania trafności, odrzucające jednolitą teorię trafności; zgodnie z nimi trafność jest własnością testu, a nie własnością interpretacji jego wyników. Lissitz i Samuelsen (2007) postulują, że współczesna teoria trafności, skoncentrowana na trafności teoretycznej, ma bardzo niewiele do zaoferowania badaniom testowym prowadzonym w obszarze edukacji. W proponowanym przez nich ujęciu trafności centralną rolę odgrywa trafność treściowa, którą łącznie z rzetelnością testu określają mianem trafności wewnętrznej. Tym samym ustalenie definicji mierzonej przez test wielkości w procesie jego konstrukcji oraz stabilność wyników mają zasadnicze znaczenie dla trafności testu. Wewnętrzne charakterystyki testu nie zależą od czynników zewnętrznych, np. związków wyników testu z innymi miarami. Aspekty zewnętrzne w stosunku do testu, związane z oceną użyteczności wykorzystania testu w danym celu, wsparciem ze strony teorii oraz konsekwencjami testowania, pomimo że bardzo ważne, nie powinny być włączane do definicji trafności.

> Te odrębne charakterystyki – które aktualnie określa się jako ustalanie trafności kryterialnej i teoretycznej – są bardzo ważne, jednakże użytkownik tych technik powinien zdawać sobie sprawę, że dostarczają one odpowiedzi na fundamentalnie różne pytania. Te pytania nie powinny odciągać uwagi badacza od skoncentrowania się na trafności wewnętrznej. (Lissitz, Samuelsen, 2007, s. 446 – tłum. J. R.)

Borsboom i współpracownicy (Borsboom i in., 2004; Borsboom, Cramer, Kievit, Scholten, Franic, 2009) krytykują pojęcie trafności teoretycznej, określając je mianem doktryny, i wskazują, że jednolita teoria trafności nie jest potrzebna psychologii, ponieważ nie ma czego ujednolicać. Integracja wszystkich związanych z testem działań (*validation*) nie jest trafnością (*validity*). Trafność jest własnością przysługującą testowi, a zadaniem badacza jest sprawdzenie, czy test taką własność posiada (czy test mierzy to, co powinien mierzyć). „Trafność jest pojęciem takim jak prawda: reprezentuje idealną lub pożądaną sytuację" (Borsboom i in., 2004, s. 1063). Zgodnie z proponowaną przez autorów koncepcją trafności test jest trafną miarą danego atrybutu, gdy ten atrybut istnieje i jego zróżnicowanie stanowi przyczynę zróżnicowania wyników testu, co oznacza, że relacja między wynikami testu i atrybutami ma charakter przyczynowy, a nie korelacyjny. Tym samym wnioskowanie o trafności wyłącznie na podstawie macierzy wielu cech-wielu metod nie jest możliwe, jako że pożądana konfiguracja współczynników korelacji nie ustanawia trafności. Tym, co stanowi o trafności, jest istnienie atrybutu i wykazanie jego przyczynowego wpływu na wyniki testu. „Problem trafności testu odnosi się do kwestii, czy atrybut psychologiczny (np. 'inteligencja ogólna') istnieje i odpowiada atrybutowi, który test rzeczywiście mierzy (przyjmując, że jakiś mierzy). W tym przypadku terminy 'inteligencja ogólna' i 'atrybut, który test mierzy' odnoszą się do tej samej struktury. Walidacja testu to po prostu przeprowadzenie badania w celu dowiedzenia się, czy jest to prawda, czy nie" (Borsboom i in., 2009, s. 153 – tłum. J. R.). Ujmowana w ten sposób trafność staje się własnością jakościową. Określanie „stopnia trafności" traci sens, a na pytanie o to, czy dany test jest trafny, można udzielić prostej odpowiedzi: tak lub nie.

PODSUMOWANIE

Prowadzona od ponad stu lat dyskusja nad konceptualizacją trafności jako pojęcia psychometrycznego trwa. Zainteresowany Czytelnik może zapoznać się z aktualnie podejmowanymi problemami, sięgając choćby do poświęconego pojęciu trafności wydania czasopisma *Assessment in Education: Principles, Policy & Practice* (2016, t. 23, nr 2) czy też do wydania *Educational Assessment* (2020, t. 25, nr 1), dedykowanego zagadnieniom praktyki walidacyjnej.

Czy tocząca się dyskusja doprowadziła do jakichś konkluzywnych ustaleń? Odpowiedź na to pytanie musi być – niestety! – przecząca. Kontrowersje nadal dotyczą kluczowej kwestii, związanej z tym, do czego się trafność odnosi: czy jest ona własnością narzędzia, czy związana jest z interpretacją wyników testu czy też z konkretnymi jego zastosowaniami? Pod szyldem jednolitej teorii trafności każdy z ważnych aspektów dotyczących testu i procesu testowania został objęty terminem *trafność*, w efekcie czego sam termin stał się wyjątkowo złożony znaczeniowo i – rzec by można – nieco uciążliwy. Nie dziwi zatem, że niektórzy badacze postulują, że nadszedł czas, by z niego zrezygnować: „Wziąwszy pod uwagę powszechność nieprecyzyjnego i niejednoznacznego użycia oraz fakt, że nawet specjaliści używają tego słowa w całkiem różny sposób, trudno zrozumieć, jak cokolwiek koncepcyjnie fundamentalnego mogłoby zostać utracone, gdyby słowo to przeszło na emeryturę" (Newton, Shaw, 2016, s. 190 – tłum. J. R.). Termin *trafność* można by zastąpić bardziej ogólnym terminem *jakość* (*quality*), który – obejmując znaczeniowo wszystkie odniesienia terminu *trafność* – wymagałby również od użytkowników sprecyzowania jego znaczenia w konkretnym kontekście użycia (Newton, Shaw, 2013). Z różnicami w ujmowaniu trafności związane są kontrowersje dotyczące właściwego sposobu jej ustalania.

Wyniki badania przeprowadzonego przez Camargo, Herrerę i Traynor (2018) za pomocą metody delfickiej (którym to badaniem objęto siedmiu ekspertów zajmujących się problematyką trafności) wydają się potwierdzać różnice w odniesieniu do tych dwóch podstawowych kwestii: czemu przysługuje trafność i jaki jest właściwy sposób jej ustalania? Biorący udział w badaniu eksperci uznali definicję trafności zamieszczoną w ostatnim wydaniu *Standardów dla testów…* za niejasną i interpretacyjnie wieloznaczną, zgadzając się, że określenie, do czego odnosi się trafność, jest utrudnione przez brak jasnej definicji tego terminu. Zgodzili się także co do potrzeby sformułowania jasnych wskazań odnośnie do przeprowadzania procesu walidacji, nie osiągając jednak zgody co do roli, jaką konsekwencje testowania miałyby w nim odgrywać, ani też w odniesieniu do tego, ile danych i dane jakiego typu są konieczne lub wystarczające do ustalenia trafności. Eksperci nie osiągnęli również zgody w kwestii zastąpienia terminu *trafność* innym terminem, takim jak np. *jakość*, czy w sprawie definiowania walidacji w sposób obejmujący wszystko, co wiąże się z naukową oceną jakości pomiaru. Pozwala to sądzić, że termin *trafność* pozostanie w słowniku nauki, a dyskusja nad konceptualizacją trafności będzie toczyć się nadal.

BIBLIOGRAFIA

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

American Educational Research Association, American Psychological Association, National Council on Measurement in Education (2014). *Standards for educational and psychological testing* [wersja Adobe Digital Edition]. Pobrane z: https://www.testingstandards.net/uploads/7/6/6/4/76643089/standards_2014edition.pdf

Anastasi, A., Urbina, S. (1999). *Testy psychologiczne*. Warszawa: Pracownia Testów Psychologicznych PTP.

Angoff, W. H. (1988). Validity: an evolving concept. W: H. Wainer, H. Braun (red.), *Test validity* (s. 19–32). Hillsdale, NJ: Lawrence Erlbaum.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrica*, *71*, 425–440. https://doi.org/10.1007/s11336-006-1447-6

Borsboom, D., Mellenbergh, G. J., van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*, 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061

Borsboom, D., Cramer, A. O. J., Kievit, R. A., Scholten, A. Z., Franic, S. (2009). The end of construct validity. W: R. Lissitz (red.), *The concept of validity. Revisions, new directions, and applications* (s. 135–170). Charlotte, NC: Information Age Publishing.

Campbell, D. T., Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81–105. https://doi.org/10.1037/h0046016

Camargo, S. L., Herrera, A. N., Traynor, A. (2018). Looking for a consensus in the discussion about the concept of validity: a Delphi study. *Methodology*, *14*, 146–155. https://doi.org/10.1027/1614-2241/a000157

Chapelle, C. A., Enright, M. K., Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice*, *29*(1), 3–13. https://doi.org/10.1111/j.1745-3992.2009.00165.x

Cizek, G. J. (2012). Defining and distinguishing validity: interpretations of score meaning and justifications of test use. *Psychological Methods*, *17*, 31–43. https://doi.org/10.1037/a0026975

Cizek, G. J. (2016). Validating test score meaning and defending test score use: different aims, different methods. *Assessment in Education: Principles, Policy & Practice*, *23*, 212–225. https://doi.org/10.1080/0969594X.2015.1063479

Cizek, G. J. (2020). *Validity: an integrated approach to test score meaning and use*. Nowy Jork, NY: Routledge.

Cizek, G. J., Rosenberg, S., Koons, H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement*, *68*, 397–412. https://doi.org/10.1177/0013164410379323

Cizek, G. J., Bowen, D., Church, K. (2010). Sources of validity evidence for educational and psychological tests: a follow-up study. *Educational and Psychological Measurement*, *70*, 732–743. https://doi.org/10.1177/0013164407310130

Cronbach, L. J. (1971). Test validation. W: R. L. Thorndike (red.), *Educational measurement* (wyd. 2, s. 443–507). Waszyngton, DC: American Council on Education.

Cronbach, L. J. (1988). Five perspectives on validation argument. W: H. Wainer, H. Braun (red.), *Test validity* (s. 3–17). Hillsdale, NJ: Lawrence Erlbaum.

Cronbach, L. J., Meehl, P. E. (1955). Construct validity in psychological test. *Psychological Bulletin*, *52*, 281–302. https://doi.org/10.1037/h0040957

Dorans, N. J., Holland, P. W. (1993). DIF detection and description: Mantel-Haenzel and standardization. W: P. W. Holland, H. Wainer (red.), *Differential item functioning* (s. 35–66). Hillsdale, NJ: Lawrence Erlbaum.

Fremer, J. (2000). Promoting high standards and the „problem" with construct validation. *NCME Newsletter*, *8*(3), 1.

Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement*, *6*, 427–438. https://doi.org/10.1177/001316444600600401

Guion, R. M. (1978). „Content validity" in moderation. *Personnel Psychology*, *31*, 205–213. https://doi.org/10.1111/j.1744-6570.1978.tb00440.x

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*, 527–535. https://doi.org/10.1037/0033-2909.112.3.527

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, *38*, 319–342. https://doi.org/10.1111/j.1745-3984.2001.tb01130.x

Kane, M. T. (2004). Certification testing as a illustration of argument-based validation. *Measurement*, *2*, 135–170. https://doi.org/10.1207/s15366359mea0203_1

Kane, M. T. (2006a). Content-related validity evidence in test development. W: S. M. Downing, T. M. Haladyna (red.), *Handbook of test development* (s. 115–154). Mahwah, NJ: Lawrence Erlbaum.

Kane, M. T. (2006b). Validation. W: R. L. Brennan (red.), *Educational measurement* (wyd. 4, s. 17–64). Westport, CT: Greenwood Publishing.

Kane, M. T. (2008). Terminology, emphasis, and utility in validation. *Journal of Educational Measurement*, *37*, 76–82. https://doi.org/10.3102/0013189X08315390

Kane, M. T. (2009). Validating the interpretations and uses of test scores. W: R. Lissitz (red.), *The concept of validity. Revisions, new directions, and applications* (s. 39–64). Charlotte, NC: Information Age Publishing.

Kane, M. T. (2013). Validating the interpretations and uses of test scores'. *Journal of Educational Measurement*, *50*, 1–73. https://doi.org/10.1111/jedm.12000

Kane, M. T. (2016). Explicating validity. *Assessment in Education: Principles, Policy & Practice*, *23*, 198–211. https://doi.org/10.1080/0969594X.2015.1060192

Lissitz, R. W., Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, *36*, 437–448. https://doi.org/10.3102/0013189X07311286

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, Monograph Supplement*, *3*, 635–694. https://doi.org/10.2466/pr0.1957.3.3.635

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, *35*, 1012–1027. https://doi.org/10.1037/0003-066X.35.11.1012

Messick, S. (1989). Validity. W: R. L. Linn (red.), *Educational measurement* (s. 13–103). Nowy Jork, NY: American Counsil on Education and Macmillan.

Messick, S. (1995). Validity of psychological assessment. Validation of inferences of persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*, 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Murphy, K. R. (2009). Content validation is useful for many things, but validity isn't one of them. *Industrial and Organizational Psychology*, *2*, 453–464. https://doi.org/10.1111/j.1754-9434.2009.01173.x

Newton, P. E., Shaw, S. D. (2013). Standards for talking and thinking about validity. *Psychological Methods*, *18*, 301–319. https://doi.org/10.1037/a0032969

Newton, P. E., Shaw, S. D. (2016). Disagreement over the best way to use the word 'validity' and options for reaching consensus. *Assessment in Education: Principles, Policy & Practice*, *23*, 178–197. https://doi.org/10.1080/0969594X.2015.1037241

Schmidt, F. L., Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology*, *62*, 529–540. https://doi.org/10.1037/0021-9010.62.5.529

Schmidt, F. L., Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, *124*, 262–274. https://doi.org/10.1037/0033-2909.124.2.262

Sireci, S. (1998). The construct of content validity. *Social Indicators Research*, *45*, 83–117. https://doi.org/10.1023/A:1006985528729

Sireci, S. (2009). Packing and unpacking sources of validity evidence: History repeats itself again. W: R. Lissitz (red.), *The concept of validity. Revisions, new directions, and applications* (s. 19–37). Charlotte, NC: Information Age Publishing.

Spies, R. A., Plake, B. S. (red.). (2005). *The sixteenth mental measurements yearbook*. Lincoln, NE: Buros Institute of Mental Measurements.

Toulmin, S. (1958). *The uses of argument*. Cambridge: Cambridge University Press.

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, *4*, 223–233. https://doi.org/10.1080/15434300701375832

Wolming, S., Wikström, C. (2010). The concept of validity in theory and practice. *Assessment in Education: Principles, Policy & Practice*, *17*, 117–132. https://doi.org/10.1080/09695941003693856