KWARTALNIK PEDAGOGICZNY
2015 NUMER 4(238)

**Piotr Maciej Skorupiński**
*III Liceum Ogólnokształcące im. gen. Józefa Sowińskiego w Warszawie*

# Standardy trafności pomiaru

**Summary**

STANDARDS OF MEASUREMENT VALIDITY

According to the latest edition of the *Standards for Educational and Psychological Testing* (AERA et al. 2014), validity is still the most fundamental construct in educational measurement. The new edition, like the previous versions of the *Testing Standards*, contains standards of professional practice in validation. These standards of validity will be presented in the fourth section of this article. The first three sections explore validity standards included in the *Testing Standards* published before the 2014 edition, in the period of the shift from the trinitarian doctrine validity to the unitarian view of validity. The closing remarks consider the issue of the role of consequences in the most recent validity recommendations.

**Key words**: validity; validation; AERA, APA, NCME *Standards*.

red. Paulina Marchlik

Idea niezależnego i profesjonalnego monitorowania użytkowania testów ma już ponad 100-letnią historię. Jedną z form jej konkretyzacji stanowią próby ustanowienia standardów w dziedzinie stosowanego pomiaru edukacyjnego (Madaus 2001: 2).

Zarówno w pionierskich *Technicznych rekomendacjach dla testów osiągnięć*[1] (AERA i in. 1955), jak i w kolejnych pięciu edycjach standardów z lat 1966–2014, przygotowanych i popularyzowanych przez Amerykańskie Towarzystwo Badań Edukacyjnych (*American Educational Research Association* – AERA), Narodową Radę do spraw Pomiarów Edukacyjnych (*National Council on Measurement in Education*

[1] Cytaty z publikacji obcojęzycznych w tłumaczeniu autora.

– NCME) oraz Amerykańskie Towarzystwo Psychologiczne (*American Psychological Association* – APA), przed standardami rzetelności zarekomendowane zostały standardy trafności pomiaru. Rekomendacjom tym towarzyszyły wykładnie pojęcia trafności, w których wykorzystano dwa jej modele.

*Techniczne rekomendacje* współotwierały epokę dominacji trynitarnego modelu trafności, obejmującego trzy zasadnicze typy trafności: kryterialną, treściową oraz teoretyczną. Ostatnie trzy wydania *Standardów edukacyjnego i psychologicznego testowania* upowszechniają natomiast model unitarny, w którym typologię trynitarną zastąpiono zunifikowanym i poszerzonym pojęciem trafności, definiowanym jako sąd oceniający, w jakim stopniu dane empiryczne i konstrukcje teoretyczne upoważniają do inferencji i działań opartych na wynikach pomiaru edukacyjnego (Messick 1989).

W kolejnych częściach niniejszego artykułu te dwa modele trafności zostaną szerzej scharakteryzowane. Szczegółowo przedstawione będą także standardy trafności, rozumiane jako wzorce dobrej praktyki walidacyjnej, rekomendowane przez wspomniane wyżej towarzystwa w ciągu ostatnich sześciu dziesięcioleci. W finalnej części zasygnalizowana jest obiekcja odnosząca się do zasadniczej zmiany w pojęciu trafności, wprowadzonej w najnowszej edycji *Standardów* z 2014 roku.

## Pierwsze rekomendacje

*Techniczne rekomendacje dla testów osiągnięć* zostały przygotowane na podstawie *Technicznych rekomendacji dotyczących testów psychologicznych i technik diagnostycznych* (APA i in. 1954). Adresowane były zarówno do konstruktorów oraz wydawców, jak i do użytkowników testów osiągnięć – jako „standardy profesjonalizmu w praktyce" i „standardy aprobowanej praktyki" (AERA i in. 1955: 3, 5) – z myślą, iż poważna refleksja nad użytecznością danego testu oraz znaczeniem jego wyników wymaga informacji, które powinny zostać najpierw zgromadzone, a następnie udostępnione w materiałach towarzyszących testowi, czyli w tzw. podręczniku jego użytkownika.

Trzydziestosześciostronicowe *Techniczne rekomendacje dla testów osiągnięć* prezentują wspomniane standardy w sześciu działach poświęconych rozpowszechnianiu informacji o teście, interpretacji jego wyników, trafności, rzetelności, przeprowadzaniu testu, a także skalom i normom. Rozdział o trafności zawiera 42 standardy trafności oraz eksplikację trynitarnego jej modelu. W jej świetle,

celem zastosowania testu jest wyrobienie osądu, zaś stopień, w jakim test umożliwia osiągnięcie tego celu, jest przedmiotem informacji o trafności. Informacja ta – w formie dowodów walidacyjnych – może dotyczyć trafności treściowej, diagnostycznej, prognostycznej lub teoretycznej, w zależności od typu osądu, do którego zmierza testowanie. Jak to zostało ujęte: „Dla każdego typu osądu wymagany jest nieco inny typ dowodu walidacyjnego" (AERA i in. 1955: 15).

Standardy trafności również pogrupowano w kilku sekcjach poświęconych poszczególnym typom trafności oraz zagadnieniom ogólnym. W sekcji ogólnej zapisano, iż podręcznik, który należy dołączyć do każdego przeznaczonego do użytkowania testu, „powinien informować o trafności każdego typu inferencji, dla której dany test jest rekomendowany" (standard C 1), „powinien klarownie określać typ trafności, o którym mowa" (C 2) i „wskazywać, które [...] spośród interpretacji wyników testu [...] nie zostały poparte dowodami" walidacyjnymi (C 3).

W sekcji *Trafność treściowa* zawarto rekomendację, iż „jeśli rozwiązywanie testu będzie interpretowane jako próba czynności z pewnego uniwersum sytuacji, podręcznik powinien wyraźnie wskazywać, jakie uniwersum jest reprezentowane oraz jak adekwatny jest dobór próby" (C 4). W związku z powyższym podkreślono konieczność precyzyjnego zdefiniowania wspomnianego uniwersum treści (C 4.1) oraz opisu metody doboru zadań (C 4.2), a także wskazano, iż „jeśli eksperci proszeni są o osąd, czy zadania są odpowiednią próbą z danego uniwersum albo czy są prawidłowo punktowane, należy przedstawić liczbę ekspertów, fakty dotyczące ich odpowiedniego doświadczenia zawodowego i kwalifikacji, szczegóły procesu oceniania oraz dane dotyczące niezależnej zgodności między nimi w ocenie" (C 4.4).

W sekcjach poświęconych trafności diagnostycznej i prognostycznej zalecono zamieszczenie w podręczniku użytkownika testu stosownego opisu wykorzystanych miar kryterialnych, zawierającego także ocenę adekwatności użytej miary, wskazującej te aspekty mierzonej czynności, których nie odzwierciedla dana miara, oraz czynniki irrelewantne, które znalazły odbicie w wynikach (C 6). Opis taki powinien również informować o rzetelności miar kryterialnych (C 7) oraz wyczerpująco charakteryzować próby walidacyjne, pod wskazanymi względami porównywalne do grup poddawanych testowaniu (C 12). W tym zakresie rekomendowane jest przedstawienie „odpowiednich miar tendencji centralnej i wariancji wyników testu w próbie walidacyjnej" (C 12.1), oraz korelacji między wynikami testu i miar kryterialnych (C 6.1).

Rozdział *Trafność teoretyczna* zawiera 5 standardów poprzedzonych uwagami dotyczącymi roli tego typu trafności w pomiarze edukacyjnym. W świetle tych

uwag, „trafność teoretyczna jest szczególnie ważna w przypadkach, które zazwyczaj dominują w dziedzinie testów osiągnięć edukacyjnych, kiedy nieosiągalne jest odpowiednie niezależne kryterium. Walidacja testów osiągnięć poprzez zademonstrowanie ich mocy różnicowania pomiędzy wybranymi kryterialnymi grupami odniesienia jest w efekcie trafnością teoretyczną [...]. Dane wskazujące na moc różnicującą w koniunkcji z trafnością treści są wystarczające dla większości testów osiągnięć, zorganizowanych wokół materii przedmiotowej. Pozostałe rekomendacje [...] są znaczące dla nowszego typu testów osiągnięć, które usiłują mierzyć takie konstrukty, jak rozumienia, postawy, zainteresowania, zamiłowania i nawyki lub umiejętności uczenia się. Te [...] nie reprezentują identyfikowalnego uniwersum treści [...]" (AERA i in. 1955: 26). Podręcznik użytkownika testu mierzącego tego typu konstrukty „powinien przedstawiać wszelkie dostępne informacje, które pomogą użytkownikowi w określeniu, jakie właściwości [...] znajdują odbicie w wariancji wyników testu" (C 16), oraz „przedstawić zarys teorii, na której oparty jest test, i porządkować wszelkie dostępne dane walidacyjne, aby pokazać, w jaki sposób wspierają one teorię" (C 17).

## Trynitarne standardy trafności

W okresie dominacji trynitarnego modelu trafności opublikowane zostały – oprócz *Technicznych rekomendacji* – dwa zbiory standardów pomiaru, w których model ten propagowano.

Pierwszy z nich: *Standardy dla Testów Edukacyjnych i Psychologicznych i Podręczników Testowania* (APA i in. 1966), powielał strukturę *Technicznych rekomendacji*, przy czym w rozdziale *Trafność* w miejsce sekcji *Trafność prognostyczna* i *Trafność diagnostyczna* wprowadzono gruntownie przeredagowaną *Trafność kryterialną*. W *Trafności teoretycznej* powtórzono 4 rekomendacje z *Technicznych rekomendacji* APA, wskazujące konieczność przedstawienia korelacji między testem i innymi testami, „których interpretacja jest relatywnie klarowna", a także innymi akceptowanymi miarami tego samego konstruktu (standardy C 7.12 i C 7.13), konieczność zaprezentowania danych dotyczących efektu prędkości w wypadku testu przeprowadzanego w określonych granicach czasowych (C 7.25) oraz informowania o wynikach analizy czynnikowej, jeżeli w takiej wykorzystano dany test (C 7.22). Rekomendacje te stanowiły rozwinięcie jednego z dwóch podstawowych zaleceń APA z 1954 roku: o konieczności „przedstawienia wszelkich dostępnych informacji, które towarzyszyły będą użytkownikowi w rozpoznaniu tych

właściwości psychologicznych, które odbijają się w wariancji wyników" (C 18). W wydaniu z 1966 roku zalecenie to zmieniono, nadając mu postać: „Jeżeli autor proponuje interpretować test jako miarę pewnej zmiennej teoretycznej – zdolności, cechy czy postawy – proponowana interpretacja powinna zostać w pełni wyrażona", a także „odróżniona od interpretacji wynikających z innych teorii" (C 7). Ponadto uszczegółowiono je w dwóch nowych stwierdzeniach: o konieczności podsumowania badań związanych z hipotezami wywiedzionymi z teorii oraz przedstawienia danych dotyczących stopnia, w jakim inne konstrukty znajdują odbicie w wariancji wyników.

Wspomniane wyżej rekomendacje oraz 49 pozostałych standardów trafności powiązano z trynitarną koncepcją trafności. Świadczy o tym także wstęp wprowadzający pojęcie trafności, w którym powtórzono – z drobnymi korektami – sformułowania z *Technicznych rekomendacji* APA, na przykład stwierdzenie, iż „ocena trafności teoretycznej dokonywana jest w trakcie poszukiwania tych jakości, które mierzone są za pomocą testu, to znaczy ustalenia stopnia, w jakim określone pojęcia eksplikatywne albo konstrukty wyjaśniają wykonanie danego testu" (APA i in. 1966: 13). Jedyną znaczącą zmianą w tym wstępie jest termin „trafność kryterialna".

Trynitarna koncepcja trafności przyjęta została także w kolejnych *Standardach dla Testów Edukacyjnych i Psychologicznych* (APA i in. 1974). W rozdziale *Standardy dla raportów z badań rzetelności i trafności* zamieszczono 82 standardy trafności. W sekcji *Trafność treściowa* powtórzono 6 standardów z edycji poprzedniej, dotyczących doboru próby treści z uniwersum programowego, oraz dodano zalecenie sprawdzania treści testu pod kątem ewentualnej jej stronniczości (standard E 12.1.2) oraz sugestię dotyczącą dowodów trafności treściowej testów zawodowych (E 12.4). W *Trafności teoretycznej* powtórzono 10 z 12 rekomendacji z 1966 roku, utrzymując strukturę ich zapisu: zalecenie główne (pełnego wyrażenia proponowanej interpretacji wyników testu jako miary konstruktu), dwa zalecenia podrzędne (wskazania stopnia, w jakim proponowana interpretacja została zweryfikowana, i przedstawienia danych dotyczących stopnia, w jakim inne konstrukty znajdują odbicie w wariancji wyników) oraz kilka zaleceń szczegółowych. Dodano przy tym tylko jedną nową rekomendację wskazującą, iż „jeżeli różnice w strategiach rozwiązywania testu, które mogą wpływać na interpretację wyników, są związane z identyfikowalnymi właściwościami grup społecznych, ta informacja powinna zostać klarownie przedstawiona [...]" (E 13.2.25).

W omawianym rozdziale gruntownie zmieniono natomiast wstęp wprowadzający w problematykę trafności. Podjęto w nim próbę zdefiniowania terminu

„konstrukt" – jako „idei rozwiniętej lub «skonstruowanej» w wyniku pracy wyobraźni naukowej i poinformowanej [...], mającej na celu wyjaśnienie i uporządkowanie niektórych aspektów istniejącej wiedzy" (APA i in. 1974: 29) – oraz nowego opisu koncepcji trafności.

Trafność – w tym ujęciu – odnosi się do odpowiedniości wniosków wyprowadzonych z wyników testu lub innych form oceny. Wyróżnione typy trafności – kryterialna, treściowa i teoretyczna – dotyczą różnych typów wniosków, które sformułowane być mogą na podstawie wyników testowania. Te aspekty trafności są ze sobą powiązane tak, iż „wszelka wiedza dotycząca trafności jest relewantna w trakcie ewaluacji trafności teoretycznej" (APA i in. 1974: 26). Ocena testu w świetle określonego konstruktu wymaga zatem akumulacji wyników wszelkich dostępnych badań.

We wstępie zarysowano także przykładowe koncepcje badania trafności. Walidacja – rozumiana jako proces gromadzenia i oceny danych związanych z interpretacją wyników – rozpoczynać się zatem może od sformułowania hipotez dotyczących właściwości osób, które uzyskały wysokie i niskie wyniki. W jej trakcie możliwe jest chociażby wykorzystanie wyniku testu jako zmiennej zależnej bądź niezależnej w badaniu eksperymentalnym. Można również przeprowadzić weryfikację prognoz zachowania się osób o odmiennych wynikach w różnych sytuacjach czy analizę zadań w powiązaniu z zewnętrznym kryterium. Jak jednak podkreślono: „Dowody trafności teoretycznej nie są owocem pojedynczego badania; sądy dotyczące trafności teoretycznej opierają się raczej na zakumulowanych wynikach wielu badań" (APA i in. 1974: 30).

## Unitarne standardy trafności

Kolejna edycja *Standardów edukacyjnego i psychologicznego testowania* (AERA i in. 1985) różniła się zasadniczo od wersji z 1974 roku. Po pierwsze, była obszerniejsza. Po drugie, ustępy na temat trafności zostały ulokowane na początku części pierwszej: *Technicznych standardów konstrukcji i ewaluacji testu*. Po trzecie, zostały one zasadniczo przeredagowane.

Zmiany w podrozdziale *Trafność* wiązały się z fundamentalną zmianą w koncepcji trafności, która została wyrażona już w pierwszym jego akapicie. Ponieważ trafność – definiowana jako odpowiedniość, sensowność i użyteczność określonej inferencji wyprowadzonej z wyników testowania, a więc jako stopień, w jakim zgromadzone dowody empiryczne i teoretyczne uzasadniają daną infe-

rencję – jest „unitarną koncepcją" (AERA i in. 1985: 9), w dokumencie nie stosuje się już tradycyjnych terminów określających typy czy aspekty trafności. W ich miejsce wprowadza się nazwy trzech kategorii dowodów trafności, dotyczących konstruktu, treści testu i miar kryterialnych. Odchodzi się także od wyodrębniania sekcji zawierających standardy związane z typami trafności.

W *Trafności* zamieszczono 25 standardów trafności. Trzy spośród nich odnoszą się do problemów łączonych wcześniej z trafnością teoretyczną, a w tej wersji – z dowodami dotyczącymi konstruktu. Standard 1.8, czyli przeredagowany standard E 13 z poprzedniej edycji, stanowi, iż „jeżeli test proponowany jest jako miara konstruktu, konstrukt ów powinien być wyodrębniony spośród innych konstruktów; proponowana interpretacja wyników testu powinna być jednoznacznie określona; a związane z tym konstruktem dowody trafności powinny być przedstawione w celu uzasadnienia takich inferencji. W szczególności należy przedstawić dowody, które ukazują, że test nie zależy silnie od innych konstruktów". Dwie kolejne rekomendacje, nieobecne w *Standardach* z 1974 roku, precyzują, iż „powinno się przedstawić dowody wskazujące, że wynik znajduje się w bliższej relacji z wynikami pomiaru tego samego konstruktu innymi metodami niż z miarami innych konstruktów" (standard 1.9), oraz „demonstrujące, że wyniki testu są w bliższej relacji ze zmiennymi stanowiącymi przedmiot zainteresowania niż ze zmiennymi niewłączonymi do przyjętej teorii" (1.10).

Te trzy standardy poprzedzone są powtórzonymi za *Standardami* z 1974 roku ogólnymi zaleceniami. Zgodnie z nimi, „należy przedstawiać dowody trafności głównych typów inferencji, dla których rekomendowane jest użycie danego testu" (1.1), zaś w przypadku „jeżeli trafność danej interpretacji nie była przedmiotem badania, ten fakt powinno się wyraźnie wskazać" (1.2). Zaleca się także, aby stwierdzenia na temat trafności odnosiły się do określonych interpretacji czy decyzji (1.2).

Przywołanym powyżej standardom towarzyszą rekomendacje związane z pozostałymi dwoma typami dowodów trafności: treściowymi i kryterialnymi, a także charakterystyka źródeł tych dowodów, takich jak korelacje między zadaniami testu, korelacje między wynikami testu a wynikami zastosowania różnorodnych metod pomiaru danego i innych konstruktów, analiza procesów aktywnych w czasie pracy z zadaniem, a także sąd ekspercki, obejmujący relacje pomiędzy zadaniami a treściową dziedziną testowania. W tym ostatnim przypadku rekomenduje się – podobnie jak w poprzednich edycjach – udostępnienie użytkownikowi testu informacji na temat doświadczenia, kwalifikacji i przygotowania ekspertów oraz procedur wykorzystanych w trakcie uzgadniania sądów (1.7).

Unitarna zmiana w koncepcji trafności została pogłębiona w kolejnej edycji *Standardów edukacyjnego i psychologicznego testowania* (AERA i in. 1999). W tym wydaniu, podobnie jak w edycji poprzedniej, podkreśla się, że trafność jest unitarną koncepcją, i odchodzi od tradycyjnej terminologii, nazywającej typy trafności, w stronę terminologii określającej typy dowodów trafności.

Wśród wspomnianych dowodów znalazły się dowody oparte na badaniach treści i wewnętrznej struktury testu, procesów odpowiedzi na zadania, relacji pomiędzy wynikami zastosowania testu oraz miar tego samego i innych konstruktów, technicznej jakości systemu testowania (na przykład procedur konstrukcji i przeprowadzania testu), a także obserwowanych i przewidywanych konsekwencji testowania i ich źródeł. Do tych ostatnich zaliczono niedoreprezentowanie konstruktu (*construct underrepresetation*) w teście, definiowane jako stopień, w jakim test nie obejmuje ważnych aspektów konstruktu, oraz komponenty irrelewantne względem konstruktu (*construct-irrelevant components*), rozumiane jako te elementy składowe testu, które stymulują procesy niezwiązane z konstruktem.

Zasygnalizowane typy dowodów trafności składają się na argumentację uzasadniającą zamierzoną interpretację wyników, zaś stopień, w jakim przygotowane dowody wspierają tę interpretację, stanowi przedmiot osądu. Dziedziną walidacji jest zatem interpretacja wyników testu, nie zaś sam test.

Dwadzieścia cztery standardy trafności, dotyczące tak rozumianej walidacji, która rozpoczyna się od sformułowania proponowanej interpretacji wyników, obejmują zalecenia uzasadniania – za pomocą dowodów empirycznych i teoretycznych – każdej zamierzonej interpretacji wyników (standard 1.1), jednoznacznego sformułowania planowanej interpretacji wyników i opisu konstruktu stanowiącego przedmiot pomiaru (1.2), ostrzeżenia użytkowników testu w sytuacji braku dowodów uzasadniających interpretację wyników (1.3), a także zgromadzenia przez użytkownika testu danych niezbędnych do uzasadnienia takiego wykorzystania testu, które nie zostało poddane walidacji przez jego konstruktora (1.4). Oprócz tego włączono do zbioru standardów wymóg opisu właściwości próby walidacyjnej (1.5), procedur i kryteriów wyboru treści testu (1.6) oraz ekspertów – w przypadku walidacji z wykorzystaniem opinii eksperckich (1.7). Zalecono także przedstawienie odpowiednich dowodów w wypadku uzasadnienia odwołującego się do procesów psychologicznych i operacji kognitywnych, zachodzących u egzaminowanych (1.8), oraz opis warunków, w jakich zgromadzono dane wykorzystane w analizach statystycznych, jeżeli wyniki tego rodzaju analiz zawarte są w dowodach trafności (1.13). Ponadto zarekomendowano standard uzasadniania wyboru zmiennych (wraz z opisem konstruktów,

które one reprezentują) wykorzystanych w analizach porównawczych (1.14) oraz praktykę prezentowania danych o wewnętrznej strukturze testu w wypadku interpretacji opartej na wnioskowaniu z przesłanek dotyczących relacji między częściami testu (1.11).

W zbiorze standardów trafności znalazły się również rozstrzygnięcia w zakresie problematyki skutków zastosowania testu. Przyjęto, iż jeżeli oczekuje się, że zastosowanie lub interpretacja wyników testu przyniosą określone następstwa, podstawy antycypacji rezultatów użycia testu powinny być przedstawione (1.22 i 1.23). Zalecono także podjęcie próby badania niezamierzonych konsekwencji użycia testu w celu rozstrzygnięcia, czy nie wynikają one z niedoreprezentowania konstruktu w teście lub obecności komponentów irrelewantnych względem konstruktu, lub czułości testu na inne niż przedmiot pomiaru właściwości (1.24).

W komentarzu poprzedzającym standardy trafności podkreślono, że za walidację – rozumianą jako rozwijanie naukowych argumentów dotyczących trafności zamierzonej interpretacji wyników – odpowiedzialni są wspólnie konstruktor i użytkownik testu. Ten pierwszy odpowiada za zebranie dowodów trafności proponowanego odczytania wyników pomiaru, ten drugi – za ocenę tych dowodów z uwzględnieniem tych okoliczności, w których projektuje on zastosowanie testu, a także za zgromadzenie dodatkowych dowodów, jeżeli użytkowanie testu różni się od określonego przez jego konstruktora.

## Translokacja trafności konsekwencyjnej

Unitarny model trafności został zastosowany także w najnowszych – niemalże 230-stronicowych – *Standardach edukacyjnego i psychologicznego testowania* (AERA i in. 2014). Edycja ta stanowi zrewidowaną wersję *Standardów* z 1999 roku.

W *Trafności*, w pierwszym jej rodziale, do tekstu z poprzedniej edycji wprowadzono bądź poprawki, bądź dodatkowe ustępy. Zmieniono także układ sekcji *Standardy trafności*, zawierającej tym razem 26 standardów trafności, pogrupowanych w trzech tzw. klastrach. Klastry te poprzedzono rekomendacją nadrzędną, wyrażającą fundamentalną zasadę konstruktorską – zasadę konstruowania i udostępniania językowej interpretacji wyników testu oraz jej uzasadnienia – w formule: „Klarowna artykulacja każdej zamierzonej interpretacji wyniku testu dla określonego zastosowania powinna być przedstawiona i odpowiednie dowody trafności na poparcie każdej zamierzonej interpretacji powinny być udostępnione" (standard 1.0).

W klastrze *Ustalanie zamierzonych zastosowań i interpretacji* umieszczono siedem standardów trafności, przeniesionych – ze zmianami – z edycji poprzedniej (1.1, 1.2, 1.3, 1.4, 1.9, 1.22 i 1.23). Obejmują one między innymi standard „klarownego przedstawiania, jak wyniki testu mają być interpretowane i w konsekwencji stosowane", a w tym także przejrzystego opisu mierzonych konstruktów (1.1), standard prezentowania uzasadnienia każdej zamierzonej interpretacji wyników, obejmującego dowody empiryczne i teoretyczne (1.2), oraz standard informowania użytkowników testu o ewentualnym braku walidacji określonej interpretacji jego wyników lub o jej niezgodności z dostępnymi dowodami (1.3). Wspomniane wyżej zmiany leksykalne uściślają, iż przedmiotem walidacji jest interpretacja wyników testu w konkretnej sytuacji jego użytkowania.

W klastrze *Kwestie dotyczące prób i metod wykorzystanych w walidacji* powtórzono trzy standardy z edycji wcześniejszej (1.4, 1.7 i 1. 13) właściwie bez zmian. Rekomendacje te dotyczą sposobu opisu próby walidacyjnej (1.8), procedur związanych z udziałem sędziów kompetentnych w badaniach trafności (1.9) oraz metod gromadzenia danych wykorzystanych w analizach statystycznych (1.10).

Klaster *Specyficzne formy dowodów trafności* został podzielony na 6 części – zawierających 15 standardów, w tym 14 przeniesionych z wydania poprzedniego – zatytułowanych: *Dowody ukierunkowane na treść* (standard 1.11, czyli 1.6 z 1999 roku), *Dowody ukierunkowane na procesy kognitywne* (1.12/1.8), *Dowody ukierunkowane na wewnętrzną strukturę* (1.13/1.11, 1.14/1.12 i 1.15/1.10), *Dowody ukierunkowane na relacje z konceptualnie powiązanymi konstruktami* (1.16/1.14), *Dowody ukierunkowane na relacje z kryteriami* (1.17/1.16, 1.18/1.15, 1.19/1.17, 1.20, 1.21/1.18, 1.22/1.20, 1.23/1.21 i 1.24/1.19) oraz *Dowody oparte na konsekwencjach testów* (1.25/1.24). Jedyny nowy standard w tym klastrze stanowi zalecenie raportowania stopnia niepewności związanej z miarami wielkości efektu – w sytuacji wykorzystywania ich we wnioskowaniu (1.20). Nieliczne zmiany w pozostałych standardach to albo korekty leksykalne, albo kilkuwyrazowe uzupełnienia.

Najobszerniejsze zmiany w komentarzach dołączonych do każdego standardu wprowadzone zostały w ustępie poświęconym standardowi 1.25, który określa działania badawcze, jakie należy podjąć w sytuacji wystąpienia niezamierzonych konsekwencji zastosowania testu, w słowach: „Gdy niezamierzone konsekwencje stanowią rezultat zastosowania testu, należy podjąć próbę zbadania, czy konsekwencje te wynikają z czułości testu na cechy inne niż te, które zamierzano ocenić, czy też z defektu testu w zakresie pełnego reprezentowania zamierzonego konstruktu" (1.25). W ustępie dodanym do trzyzdaniowego komentarza z 1999 roku, poświęconego kwestii relacji między konsekwencjami użycia testu a zachodzącym w nim niedoreprezentowaniem konstruktu i funkcjonującymi w jego

ramach komponentami irrelewantnymi względem konstruktu, stwierdzono, że „odkrycie niezamierzonych konsekwencji może prowadzić do namysłu nad odpowiedniością zastosowanego konstruktu". Dopowiedziano także, iż „odpowiedzialność za upewnienie się, że niezamierzone konsekwencje są przedmiotem ewaluacji, leży po stronie tych, którzy podejmują decyzję w sprawie użycia danego testu" (AERA i in. 2014: 31).

Również w sekcji poprzedzającej *Standardy trafności*, w której zreferowano problematykę źródeł dowodów trafności interpretacji wyników testu w danym jego zastosowaniu, najobszerniejsze zmiany wprowadzono w dziale dotyczącym konsekwencji testowania, który w edycji z 1999 roku nosił tytuł *Dowody oparte na konsekwencjach testowania*, zaś w wydaniu ostatnim – *Dowody trafności i konsekwencje testowania*. W dziale tym – obok ustępów z poprzedniej edycji, w których podjęto próbę modyfikacji unitarnego modelu Messicka w jego wymiarze konsekwencyjnym – uwzględniono bardziej zdecydowane rozgraniczenie między problematyką trafności interpretacji wyników testu w danym jego zastosowaniu a problematyką konsekwencji praktyk pomiarowych. I o ile we wcześniejszej edycji zwracano już uwagę na „rozróżnienie pomiędzy dowodami, które są bezpośrednio relewantne względem trafności, i dowodami, które mogą wspierać decyzje dotyczące polityki społecznej, ale które znajdują się poza sferą trafności", i przestrzegano, iż „chociaż informacje o konsekwencjach testowania mogą mieć wpływ na decyzje dotyczące stosowania testów, to takie konsekwencje nie umniejszają, same w sobie, trafności zamierzonych interpretacji wyników testu" (AERA i in. 1999: 16, AERA i in. 2014: 20), to dopiero w wydaniu ostatnim jednoznacznie stwierdzono, że konstruktor testu odpowiedzialny jest jedynie za przeprowadzenie walidacji rozumianej jako zebranie dowodów trafności proponowanej przez niego interpretacji wyników otrzymanych w zamierzonym zastosowaniu testu i że odpowiedzialność za zgromadzenie dowodów dotyczących twierdzeń wskazujących na pozytywne lub negatywne konsekwencje stosowania testu spoczywa na stronie, która takie twierdzenia formułuje.

W dziale *Dowody trafności i konsekwencje testowania* przywołano także przypadek ścisłego związku pomiędzy trafnością i konsekwencjami. Podobnie jak w edycji z 1999 roku, stwierdzono bowiem, iż „dowody dotyczące konsekwencji są relewantne względem trafności wtedy, kiedy mogą być powiązane ze źródłem braku trafności, takim jak niedoreprezentowanie konstruktu czy komponenty irrelewantne względem konstruktu" (AERA i in. 2014: 21). Przypadek ten ujęty został także w ramach standardu 25, stanowiącego, iż w sytuacji wystąpienia niezamierzonych konsekwencji użycia testu należy podjąć próbę zbadania, czy nie wypływają one z przywołanych wyżej źródeł braku trafności.

W innych działach sekcji poprzedzającej *Standardy trafności*, zawierających tekst z poprzedniej edycji wraz z drobnymi korektami oraz obszernymi uzupełnieniami, zasygnalizowaną powyżej kwestię relacji między trafnością i konsekwencjami poruszono w ustępach dodanych, na przykład w drugim akapicie tej sekcji, a więc już na pierwszej stronie rozdziału poświęconego trafności. Stwierdzono w nim, iż – po pierwsze – warunkiem *sine qua non* uzasadnionego zastosowania testu jest posiadanie dowodu trafności interpretacji jego wyników uzyskanych w podobnych okolicznościach jego zastosowania oraz że – po drugie – dysponowanie takim dowodem stanowi niewystarczający warunek jego użycia. Jak to ujęto w przywołanym ustępie: „W przypadku, gdy istnieją wystarczające dowody trafności, decyzja, czy przeprowadzić właśnie dany test, wymaga zazwyczaj dodatkowego namysłu. Obejmuje on kwestie kosztów i korzyści, ujęte w różnych subdyscyplinach jako analiza użyteczności lub refleksja nad negatywnymi konsekwencjami zastosowania testu, oraz wszelkie negatywne konsekwencje skonfrontowane z pozytywnymi konsekwencjami zastosowania testu" (AERA i in. 2014: 11). Takie ujęcie problemu konsekwencji użycia testu, iż oto należy je rozważyć, dysponując już dostarczonymi przez konstruktora testu dowodami trafności, stanowi poważną próbę zawężenia unitarnego pojęcia trafności i wyłączenia z jego zakresu tzw. trafności konsekwencyjnej (Niemierko 2009: 148).

Pozostałe ustępy dodane do tekstu zamieszczonego we wspomnianych wyżej działach nie zawierają tak zasadniczych zamian pojęciowych. Znajdują się w nich dodatkowe przykłady ilustrujące kontaminację konstruktu (wariancję irrelewantną względem konstruktu) oraz uwagi dotyczące dostosowania treści testu do programu kształcenia czy eksperckiego osądu jako metody oceny, czy jakość i ilość dowodów trafności są wystarczające. Podkreśla się tam także, iż w celu określonego zastosowania testu nie zachodzi potrzeba gromadzenia wszystkich typów dowodów trafności. W dziale końcowym, poświęconym integrowaniu dowodów trafności, zawarta jest konstatacja, iż mimo że „wnioskowanie walidacyjne podobne jest do wnioskowania naukowego", gdyż „proces walidacji nigdy się nie kończy", to „w pewnym punkcie dowody walidacyjne pozwalają na sumatywny osąd zamierzonej interpretacji, że jest dobrze poparta i możliwa do obronienia", przy czym „ilość i charakter dowodów wymaganych do wsparcia prowizorycznego osądu trafności często różnią się pomiędzy dziedzinami" (AERA i in. 2014: 22).

Pomimo zasygnalizowanych powyżej zmian w koncepcji trafności samą jej definicję powtórzono – z drobną korektą leksykalną – za *Standardami* z 1999 roku. Do definicji tej – w brzmieniu: „trafność odnosi się do stopnia, w jakim

dowody empiryczne i teoria wspierają interpretacje wyników testu dla zaproponowanych zastosowań testu" – dodano jednakże komentarz uściślający, iż „twierdzenia o trafności powinny odnosić się do konkretnych interpretacji związanych ze ściśle określonymi zastosowaniami". W tym kontekście zdecydowanie podkreślono także, że „używanie frazy «trafność testu» [...] jest niepoprawne" (AERA i in. 2014: 11).

## Obiekcja

Standardom trafności, sześciokrotnie uzgadnianym i publikowanym przez NCME, AERA i APA[2], towarzyszyła konceptualizacja pojęcia trafności – od modelu trynitarnego, popularyzowanego w trzech pierwszych wydaniach standardów pomiarowych, do modelu unitarnego, upowszechnianego w ostatnich trzech edycjach. W wersji najnowszej model unitarny poddany został istotnej modyfikacji, która polega na translokacji odpowiedzialności za badanie trafności konsekwencyjnej i której skutkiem może okazać się zwrot pragmatyczny w teoriotwórczej refleksji nad trafnością.

Podstawą unitarnego modelu trafności jest bowiem pojęcie trafności konstruktu reprezentującego – jako to ujął Samuel Messick – „nasze najlepsze, chociaż niedoskonałe i omylne, wysiłki uchwycenia esencji cech, które posiadają walor realności w zachowaniach niezależnych od naszych prób ich scharakteryzowania" (Messick 1998: 35). Czy zatem wspomniana translokacja stanowi jedynie próbę scedowania odpowiedzialności za niedoskonałą i omylną interpretację wyników z konstruktora na użytkownika testu, czy raczej wyraża zmianę poważniejszą – epistemologiczną i etyczną zarazem? Czy przejawem tej właśnie zmiany nie jest także argumentacyjny model trafności Michaela T. Kane'a (Skorupiński 2013), będący wyrazem – jak to określił sam Kane (2013: 64) – „bardziej pragmatycznego podejścia", w ramach którego trafność stanowi funkcję nie tylko aktywności badawczej, lecz także perswazji?

## Bibliografia

AERA, APA, NCME. 1985. *Standards for educational and psychological testing*, AERA, Washington.
AERA, APA, NCME. 1999. *Standards for educational and psychological testing*, AERA, Washington.

[2] Dwie edycje standardów zostały w całości przełożone na język polski i opublikowane pod redakcją Jerzego Brzezińskiego (Polskie Towarzystwo Psychologiczne 1985; AERA i in. 2007).

AERA, APA, NCME. 2007. *Standardy dla testów stosowanych w psychologii i pedagogice*, tłum. E. Hornowska, GWP, Gdańsk [pierwodruk: AERA i in. 1999].

AERA, APA, NCME. 2014. *Standards for educational and psychological testing*, AERA, Washington.

AERA, NCMUE. 1955. *Technical recommendations for achievement tests*, NEA, Washington.

APA, AERA, NCMUE. 1954. *Technical recommendations for psychological tests and diagnostic techniques*, „Psychological Bulletin", t. 51, nr 2 (suplement), s. 1–38 (201–238).

APA, AERA, NCME. 1966. *Standards for educational and psychological tests and manuals*, APA, Washington.

APA, AERA, NCME. 1974. *Standards for educational and psychological tests*, APA, Washington.

Kane M.T. 2013. *Validating the Interpretations and Uses of Test Scores*, „Journal of Educational Measurement", t. 50, nr 1, s. 1–73.

Madaus G.F. 2001. *A Brief History of Attempts To Monitor Testing*, „NBETPP Statements", t. 2, nr 2, Chestnut Hill, MA, dostępny na: http://files.eric.ed.gov/fulltext/ED456145.pdf [otwarty 14 marca 2015].

Messick S. 1989. *Validity*, [w:] *Educational measurement*, red. R.L. Linn, New York, s. 13–103.

Messick S. 1998. *Test validity: a matter of consequence*, „Social Indicators Research", t. 45, nr 1–3, s. 35–44.

Niemierko B. 2009. *Diagnostyka edukacyjna. Podręcznik akademicki*, Wydawnictwo Naukowe PWN, Warszawa.

Polskie Towarzystwo Psychologiczne. 1985. *Standardy dla testów stosowanych w psychologii i pedagogice*, tłum. E. Hornowska, PTP, Warszawa [pierwodruk: APA i in. 1974].

Skorupiński P.M. 2013. *Modele trafności pomiaru*, [w:] *Ścieżki rozwoju edukacyjnego młodzieży – szkoły pogimnazjalne. Trafność wskaźników edukacyjnej wartości dodanej dla szkół maturalnych*, red. M. Karwowski, IFiS PAN, Warszawa, s. 13–25.