

Testy wiedzy – tworzenie i stosowanie w weryfikacji efektów uczenia się

**Wskazówki dla instytucji certyfikujących
w Zintegrowanym Systemie Kwalifikacji**



Testy wiedzy – tworzenie i stosowanie w weryfikacji efektów uczenia się

**Wskazówki dla instytucji certyfikujących
w Zintegrowanym Systemie Kwalifikacji**

Warszawa 2020

Autorzy

dr Maksymilian Bielecki, Uniwersytet SWPS
dr Magdalena Kaczmarek, Uniwersytet SWPS
Ludmiła Rycielska

Konsultacja merytoryczna

Barbara Fijałkowska
Agnieszka Grabowska
Karolina Malinowska

Korekta językowa

Iwona Stachowicz

Opracowanie graficzne

Wojciech Maciejczyk

© Copyright by: Instytut Badań Edukacyjnych, Warszawa 2020

ISBN 978-83-66612-64-8

Wydawca

Instytut Badań Edukacyjnych
ul. Górczewska 8
01-180 Warszawa
tel. +48 22 241 71 00; www.ibe.edu.pl

Wzór cytowania: Bielecki, M., Kaczmarek, M. i Rycielska, L. (2020). *Testy wiedzy – tworzenie i stosowanie w weryfikacji efektów uczenia się. Wskazówki dla instytucji certyfikujących w Zintegrowanym Systemie Kwalifikacji*. Warszawa: Instytut Badań Edukacyjnych.

Publikacja współfinansowana przez Unię Europejską ze środków Europejskiego Funduszu Społecznego w ramach projektu „Wspieranie realizacji II etapu wdrażania Zintegrowanego Systemu Kwalifikacji na poziomie administracji centralnej oraz instytucji nadających kwalifikacje i zapewniające jakość nadawania kwalifikacji”.

Spis treści

Słowniczek	4
Wstęp	5
Wprowadzenie	6
1. Test wiedzy jako jedna z metod walidacji	7
2. Teoria i praktyka stosowania testów	9
3. Etyka stosowania testów	13
4. Projektowanie testu wiedzy	19
5. Kryteria dobroci psychometrycznej testu	22
6. Konstruowanie testów wiedzy	31
7. Rodzaje zadań/odpowiedzi stosowanych w testach wiedzy	34
8. Sposób przeprowadzania testów	45
9. Czas i miejsce stosowania testów wiedzy	51
10. Ocenianie odpowiedzi na zadania i pytania testowe	56
11. Podsumowanie: wskazówki dla instytucji certyfikujących dotyczące tworzenia i stosowania testów wiedzy w weryfikacji efektów uczenia się	66
Dodatkowe źródła informacji	70
Bibliografia	72
Autorzy	75

Słowniczek¹

asesor – osoba weryfikująca efekty uczenia się osoby przystępującej do walidacji.

certyfikowanie – proces, w wyniku którego osoba ubiegająca się o nadanie określonej kwalifikacji rynkowej, po uzyskaniu pozytywnego wyniku walidacji, otrzymuje od uprawnionego podmiotu certyfikującego dokument potwierdzający jej nadanie.

efekty uczenia się – wiedza, umiejętności oraz kompetencje społeczne nabyte w procesie uczenia się; efekty uczenia się wymagane dla danej kwalifikacji powinny być opisane w sposób zawierający: a) syntetyczną charakterystykę efektów uczenia się, b) wyodrębnione zestawy efektów uczenia się, c) poszczególne efekty uczenia się w zestawach oraz kryteria weryfikacji osiągnięcia poszczególnych efektów uczenia się.

instytucja certyfikująca (IC) – podmiot, który uzyskał uprawnienia do przeprowadzania walidacji i certyfikowania.

metody walidacji – różne metody sprawdzania osiągniętych efektów uczenia się podczas walidacji, np.: obserwacja pracy w warunkach realnych, test, analiza dowodów i deklaracji.

podmiot zewnętrznego zapewniania jakości (PZZJ) – instytucja wpisana na listę podmiotów zewnętrznego zapewniania jakości, której minister właściwy powierzył funkcję zewnętrznego zapewniania jakości wobec instytucji certyfikującej dla danej kwalifikacji.

wewnętrzny system zapewniania jakości (WSZJ) – sformułowany na potrzeby instytucji certyfikującej zbiór zasad postępowania, procedur, metod oraz rozwiązań organizacyjnych, służący zapewnianiu poprawności walidacji i certyfikowaniu oraz doskonaleniu walidacji i certyfikowania.

test wiedzy (inaczej test teoretyczny) – metoda walidacji najczęściej stosowana na etapie weryfikacji, polegająca na zadawaniu kandydatowi pytań lub wykonywaniu przez niego określonych zadań w zakresie wiedzy i umiejętności oraz opierająca się na analizie rezultatów tych działań.

walidacja – sprawdzenie, czy osoba ubiegająca się o nadanie określonej kwalifikacji, niezależnie od sposobu uczenia się tej osoby, osiągnęła wyodrębnioną część lub całość efektów uczenia się wymaganych dla tej kwalifikacji.

weryfikacja – dotyczy efektów uczenia się i oprócz identyfikowania oraz dokumentowania jest jednym z etapów walidacji.

wewnętrzny system zapewniania jakości (WSZJ) – sformułowany na potrzeby instytucji certyfikującej zbiór zasad postępowania, procedur, metod oraz rozwiązań organizacyjnych służący zapewnianiu poprawności walidacji i certyfikowaniu oraz doskonaleniu walidacji i certyfikowania.

Zintegrowany System Kwalifikacji (ZSK) – wyodrębniona część Krajowego Systemu Kwalifikacji, w której obowiązują określone w ustawie standardy opisywania kwalifikacji oraz przypisywania poziomu Polskiej Ramy Kwalifikacji do kwalifikacji, zasady włączania kwalifikacji do Zintegrowanego Systemu Kwalifikacji, a także zasady i standardy certyfikowania oraz zapewniania jakości nadawania kwalifikacji.

Zintegrowany Rejestr Kwalifikacji (ZRK) – rejestr publiczny prowadzony w systemie teleinformatycznym, ewidencjonujący kwalifikacje włączone do Zintegrowanego Systemu Kwalifikacji.

¹ Definicje spójne z przepisami ustawy o ZSK, *Słownikiem Zintegrowanego Systemu Kwalifikacji* (2017) oraz innymi publikacjami dotyczącymi ZSK opracowanymi przez Instytut Badań Edukacyjnych.

Wstęp

W publikacji tej przedstawiono podstawowe zasady projektowania oraz stosowania pisemnych testów wiedzy do weryfikacji efektów uczenia się wymaganych dla danej kwalifikacji rynkowej. Opisano najważniejsze standardy zarówno tworzenia, jak i późniejszego stosowania testów, a także wybrane problemy etyczne związane z testowaniem. Omówiono rodzaje testów ze wskazaniem ich najważniejszych wad i zalet, zasady konstruowania liczbowych wskaźników na podstawie wyniku testu oraz sposoby dokonywania oceny, czy dany efekt uczenia się został lub nie pozytywnie zweryfikowany.

Podczas spotkań roboczych i seminariów² podmioty opisujące kwalifikacje i przygotowujące się do pełnienia funkcji instytucji certyfikujących poruszały kwestie związane z tworzeniem i stosowaniem testów. Wybór zakresu treści opracowania jest konsekwencją tych wątpliwości.

Publikację przygotowano z myślą o osobach projektujących walidację, tworzących testy wiedzy na potrzeby weryfikacji oraz asesorach i osobach odpowiadających za przebieg walidacji w uprawnionych do tego instytucjach certyfikujących (IC), a także specjalistach opisujących kwalifikacje rynkowe.

² Spotkania robocze z podmiotami opisującymi kwalifikacje rynkowe i przygotowującymi się do pełnienia funkcji instytucji certyfikujących oraz seminaria otwarte pt. „Projektowanie walidacji w ZSK” były prowadzone przez zespoły „Wsparcia procesu opisywania kwalifikacji” i „Przygotowania ogółu podmiotów mogących pełnić funkcję instytucji certyfikującej w ramach projektów współfinansowanych przez Unię Europejską ze środków Europejskiego Funduszu Społecznego «Wspieranie realizacji I etapu wdrażania Zintegrowanego Systemu Kwalifikacji na poziomie administracji centralnej oraz instytucji nadających kwalifikacje i zapewniających jakość nadawania kwalifikacji» i «Wspieranie realizacji II etapu wdrażania Zintegrowanego Systemu Kwalifikacji na poziomie administracji centralnej oraz instytucji nadających kwalifikacje i zapewniających jakość nadawania kwalifikacji»”.

Wprowadzenie³

Zapewnianie jakości nadawanych kwalifikacji powinno być przedmiotem uwagi każdego podmiotu uczestniczącego w działaniach związanych z ustanawianiem i nadawaniem kwalifikacji w ZSK (Dybaś, 2015). Walidacja i certyfikowanie prowadzone przez instytucję certyfikującą są objęte obowiązkowo systemem zapewniania jakości (art. 50 pkt 1 ustawy o ZSK). Wewnętrzny system zapewniania jakości, opracowany i stosowany przez IC, służy m.in. do ciągłego monitorowania oraz bieżącej oceny trafności i rzetelności stosowanych metod walidacji. Z tego względu metody i narzędzia powinny być odpowiednio modyfikowane i doskonalone przez IC, tak aby walidacja przy ich zastosowaniu gwarantowała rzetelną i trafną weryfikację wszystkich efektów uczenia się określonych dla danej kwalifikacji.

Podmioty składające wniosek o nadanie uprawnień do certyfikowania danej kwalifikacji rynkowej muszą zapewnić kompetentną kadrę oraz zagwarantować zasoby organizacyjne do przeprowadzania dokumentowania i identyfikowania oraz weryfikacji za pomocą metod dopuszczalnych w opisie kwalifikacji. W przypadku kwalifikacji rynkowych włączonych do ZSK jedną z często stosowanych metod weryfikacji, która jest wskazywana jako możliwa do wykorzystania przez poszczególne IC, jest test teoretyczny (inaczej: test wiedzy). Dlatego warto pogłębiać i upowszechniać wiedzę na temat projektowania i stosowania testów teoretycznych.

Może się wydawać, że test teoretyczny jest łatwym, oszczędnym i szybkim sposobem na przeprowadzenie weryfikacji efektów uczenia się, a jego przygotowanie nie wymaga specjalistycznej wiedzy i umiejętności z zakresu projektowania i układania testów służących do pomiaru wiedzy. Niska świadomość metodologiczna i psychometryczna osób projektujących i stosujących testy wiedzy w IC niesie ze sobą ryzyko nieprawidłowej weryfikacji efektów uczenia. To z kolei stwarza możliwość wydawania certyfikatów osobom, które nie powinny ich otrzymać, lub niewydanie ich tym, którzy byli gotowi do potwierdzenia swojej wiedzy, umiejętności czy kompetencji społecznych, ale nie mieli takiej możliwości, ponieważ test nie był rzetelny i trafny. Ostatecznie może to wpłynąć negatywnie na jakość nadawanej kwalifikacji przez IC. Dlatego osoby tworzące testy wiedzy i asesory oraz osoby odpowiedzialne za organizację walidacji w IC, wzięwszy pod uwagę dbałość o jakość nadawanych kwalifikacji, powinny stale podnosić swoje kompetencje i poszerzać wiedzę z zakresu metodologii, psychometrii i pomiaru edukacyjnego. Niniejsze opracowanie nie tylko pozwoli poznać zainteresowanym najważniejsze zasady tworzenia i późniejszego stosowania testów, lecz także uświadomi kwestie etyczne związane z testowaniem.

³ W rozdziałach 2–11 wykorzystano fragmenty niepublikowanej ekspertyzy Maksymiliana Bieleckiego i Magdaleny Kaczmarek *Konstruowanie i stosowanie testów wiedzy w celu weryfikacji efektów uczenia się*, zamówionej przez Instytut Badań Edukacyjnych w 2019 roku.

1. Test wiedzy jako jedna z metod walidacji

Do potwierdzenia efektów uczenia się nabytych poza systemem edukacji formalnej (Council of the European Union, 2012) można wykorzystać różne metody identyfikowania, dokumentowania i weryfikacji efektów uczenia się w celu potwierdzenia efektów uczenia się/ by potwierdzić efektów uczenia się (Gmaj i in., 2016). Wśród metod stosowanych w walidacji można wymienić: obserwację w warunkach rzeczywistych lub symulowanych, wywiad, analizę dowodów i deklaracji czy test teoretyczny, inaczej zwany testem wiedzy (Katalog Metod Walidacji, <http://walidacja.ibe.edu.pl/metody/pl/>, Zintegrowany Rejestr Kwalifikacji, <http://www.rejestr.kwalifikacje.gov.pl>). Test jest jedną z najczęściej dopuszczanych metod weryfikacji efektów uczenia się wymaganych dla kwalifikacji rynkowych (por. ZRK, <http://www.rejestr.kwalifikacje.gov.pl>). Jest to też najpopularniejsza metoda walidacji w 35 innych krajach europejskich (Cedefop, Komisja Europejska i ICF, 2017).

Instytut Badań Edukacyjnych opracował Katalog Metod Walidacji, <http://walidacja.ibe.edu.pl/metody/pl/>, w którym wyróżniono m.in. testy pisemne i ustne. Ich celem jest określenie, czy dany poziom umiejętności został osiągnięty przez kandydata, oraz porównanie wyników uzyskanych przez różnych kandydatów.

Formy sprawdzania wiedzy za pomocą testów można podzielić ze względu na:

- obszar sprawdzanej wiedzy lub umiejętności (testy poznawcze i pozapoznawcze);
- rodzaj sprawdzanej wiedzy (wiedza proceduralna typu „jak” – sprawdzana w testach wykonaniowych i wiedza deklaratywna typu „co” – sprawdzana w testach wiedzy);
- sposób udzielania odpowiedzi (testy pisemne i ustne) (Katalog Metod Walidacji).

W weryfikacji efektów uczenia się wszystkie te formy testów mogą mieć zastosowanie, jeśli tak zostanie określone w obwieszczeniu ministra właściwego, informującym o włączeniu danej kwalifikacji rynkowej lub uregulowanej do ZSK, w polu opisu kwalifikacji „wymagania dotyczące walidacji oraz podmiotów przeprowadzających walidację”. W opracowaniu skupiono się na pisemnych testach wiedzy⁴.

1.1. Testy wiedzy w opisie kwalifikacji rynkowej

Wymagania dotyczące walidacji i podmiotów ją przeprowadzających są określone w opisie każdej kwalifikacji rynkowej (Ziewiec-Skokowska i in., 2016). Zapis tych wymagań jest punktem odniesienia do planowania i przeprowadzania walidacji w podmiotach, które uzyskają uprawnienia do certyfikowania.

Wymagania odnoszące się do walidacji w opisie kwalifikacji rynkowej dotyczą:

⁴ Ilekczoć w publikacji pojawia się pojęcie „test”, odnosi się ono do metody testu teoretycznego, inaczej zwanego testem wiedzy (por. Katalog Metod Walidacji).

- metod stosowanych w walidacji – służących weryfikacji efektów uczenia się wymaganych dla kwalifikacji, lecz także (o ile to potrzebne) identyfikowaniu i dokumentowaniu efektów uczenia się;
- zasobów kadrowych – wymagań kompetencyjnych w stosunku do osób przeprowadzających walidację;
- sposobu prowadzenia walidacji oraz warunków organizacyjnych i materialnych, niezbędnych do prawidłowego prowadzenia walidacji (Ziewiec-Skokowska i in., 2016, s. 32).

Wymagania dotyczące weryfikacji odnoszą się do sprawdzenia, czy wszystkie efekty uczenia, a tym samym kryteria weryfikacji podane w opisie kwalifikacji, zostały osiągnięte przez osobę przystępującą do walidacji. Test wiedzy jest często wskazywany w opisach kwalifikacji rynkowych jako metoda weryfikacji, a sposób jej przeprowadzania zwykle nie jest precyzyjnie określony. Rzadko w opisie kwalifikacji znajdują się zapisy dotyczące wymagań wobec osób opracowujących metody i narzędzia walidacji. Zazwyczaj te informacje znajdują się w szczegółowym opisie walidacji, tzw. scenariuszu walidacji tworzonemu przez IC (Brzozowska i in., 2020, rozdz. 2, <https://kwalifikacje.edu.pl/rekomendacje-w-zakresie-funkcjonowania-instytucji-certyfikujacych-w-zintegrowanym-systemie-kwalifikacji/>).

1.2. Test wiedzy w projektowaniu walidacji

Podczas projektowania walidacji warto zaprosić do współpracy osoby, które będą zaangażowane w jej przeprowadzenie na różnych etapach, w tym m.in. mające aktualną wiedzę i doświadczenie w projektowaniu narzędzi walidacji, np. pytań i zadań testowych, oraz praktyków z doświadczeniem w wykonywaniu zadań, do których realizacji powinna być przygotowana osoba przystępująca do walidacji, a także osoby, które już przeprowadzały walidację czy egzaminy, np. egzaminatorzy, asesorzy i doradcy walidacyjni (Brzozowska i in., 2020).

Scenariusze walidacji, tworzone w IC na podstawie wymagań określonych w opisie kwalifikacji rynkowej, zawierają m.in. szczegółowy opis weryfikacji, a konkretnie metod i narzędzi walidacji stosowanych przez IC, np. test wiedzy on-line. Dodatkowo powinno się przygotować materiały i procedury dla asesorów i osób przeprowadzających weryfikację, np. arkusze testowe z instrukcjami oraz klucze oceny czy protokoły z wynikiem weryfikacji dla asesorów. Zadaniem IC jest również określenie, jaki wynik weryfikacji, np. wynik testu, powinien stanowić próg uznania przez asesora lub komisję weryfikacyjną, że zostały osiągnięte wymagane efekty uczenia się przez osobę ubiegającą się o daną kwalifikację (Brzozowska i in., 2020). Ponieważ scenariusz walidacji stanowi jeden z ważniejszych elementów wewnętrznego systemu zapewniania jakości (Brzozowska i in., 2020; Dybaś, 2015), to zaplanowanie i realizacja opisanych zadań IC bezpośrednio wpływają na jakość weryfikacji efektów uczenia się, a co za tym idzie jakość nadawanej kwalifikacji.

2. Teoria i praktyka stosowania testów

Dziedziną, której głównym przedmiotem zainteresowania są teoria i praktyka konstrukcji testów, jest psychometria. Ma ona charakter interdyscyplinarny i łączy wiedzę z obszarów psychologii, statystyki i matematyki. Jednym z jej istotnych osiągnięć jest precyzyjne określenie wyróżników pozwalających na ocenę jakości testów. Jako że każdy test jest rodzajem narzędzia pomiarowego własności osoby rozwiązującej test, jego jakość – podobnie jak w przypadku każdego innego narzędzia – można ocenić, sprawdzając, czy właściwie pełni on swoją funkcję.

Wyniki popularnych testów edukacyjnych służą ocenie postępów lub osiągnięć indywidualnych uczniów, ale również całych krajów czy systemów edukacji (np. egzamin ósmoklasisty w Polsce czy międzynarodowe badanie *PISA*) (*Standards for Educational and Psychological Testing*, 2014). Specjaliści zajmujący się pomiarem edukacyjnym (ang. *educational measurement*) wykorzystują wiedzę psychometryczną do tworzenia i stosowania testów edukacyjnych.

Testy stosuje się też jako podstawę do nadawania licencji lub certyfikowania. Wówczas wynik testu stanowi dowód, że osoba ubiegająca się o licencję czy certyfikat ma wiedzę i umiejętności niezbędne do wykonywania jakichś czynności zawodowych, np. certyfikaty językowe, certyfikaty zawodowe w księgowości. Do celów zawodowych służą zwykle testy wiedzy pisemne i ustne (Brennan, 2006).

W przypadku testów wiedzy stosowanych w nadawaniu kwalifikacji ich podstawową funkcją jest jak najlepsze oszacowanie poziomu wiedzy, umiejętności i kompetencji społecznych kandydatów. Skutkiem poprawnie przeprowadzonego pomiaru powinno być jak najbardziej precyzyjne określenie, czy kandydat potwierdził efekty uczenia się określone w opisie danej kwalifikacji. Dzięki wykorzystaniu narzędzi psychometrycznych i statystycznych można to zmierzyć, a także wyrazić obiektywnymi, liczbowymi wskaźnikami (np. Brzeziński, 2019).

2.1. Testy poznawcze i pozapoznawcze

Testy poznawcze (ang. *cognitive tests*) są wykorzystywane do sprawdzania wiedzy i umiejętności związanych z poznawczymi aspektami funkcjonowania, tzw. umiejętności twarde, np.: wiedza przedmiotowa z matematyki, umiejętności czytania i pisania. Testy pozapoznawcze (ang. *non-cognitive tests*) służą do oszacowania tzw. umiejętności miękkich, czyli wszystkich umiejętności, cech, postaw, które mogą wpływać na osiągnięcie sukcesu szkolnego i zawodowego. Umiejętności pozapoznawcze zwykle są traktowane jako związane z uczeniem się formalnym, m.in. postrzegana własna skuteczność działania, motywacja do nauki, samokontrola, zaangażowanie w naukę, umiejętności radzenia sobie ze stresem i porażkami (Brunello i Schlotter, 2011; Majkut i in., 2016).

W przypadku testów poznawczych pomiar zależy na przykład od takich czynników zakłócających jak zgadywanie. Oczywiście pomiar ten – jako że stanowi pewną

próbkę wiedzy zebraną w danym momencie i określonej sytuacji – jest obarczony pewnym błędem losowym, co oznacza, że jego rzetelność nie jest najwyższa.

W przypadku testów pozapoznawczych również można wskazać wymienione źródła zakłóceń, a ponadto z reguły określić ryzyko zaistnienia dysymulacji, czyli tendencji do przedstawiania się w lepszym świetle. Ludzie powszechnie jej ulegają i nawet, gdy z wykonaniem testu nie łączą się decyzje dotyczące uznania lub nieuznania danego rodzaju umiejętności, wolą prezentować swoje zachowania jako lepsze (bardziej akceptowane społecznie), niż są one w istocie, i świadomie starają się wywrzeć pozytywne wrażenie na egzaminatorach. W sytuacji, w której od wyników testu zależy coś ważnego, tendencja ta zdecydowanie się nasila. Oznacza to, że testy, w których pyta się o typowe lub hipotetyczne zachowania, obarczone są ryzykiem, że mierzą bardziej życzeniowy poziom wiedzy i umiejętności niż ten realnie występujący.

2.2. Testy wykonaniowe, testy ustne i pisemne

Testy można podzielić na takie, które mierzą poziom wykonania danej czynności (testy wykonaniowe), oraz w których badany jest poziom tzw. wiedzy deklaratywnej. Ten pierwszy można określić mianem wiedzy „jak zrobić”, drugi zaś „co zrobić” (Marzano, 2001). Należy wyraźnie podkreślić, że mimo istnienia pewnego transferu uczenia się między tymi rodzajami wiedzy, to na pewno nie są one tożsame i poziom tych typów wiedzy w danym obszarze w skrajnych przypadkach może być bardzo rozbieżny.

Z reguły osoba, która nabywa wiedzę przez działanie, dysponuje dobrymi umiejętnościami (wiedzą „jak zrobić”), ale nie zawsze jest świadoma tej wiedzy i potrafi o niej opowiedzieć; ta zaś, która uczy się w sposób teoretyczny, zdobywa wiedzę „co zrobić”, ale niekoniecznie potrafi ją zastosować w praktyce. Testując dany poziom wiedzy, nie można więc pochopnie wnioskować o drugim rodzaju wiedzy. Testy wykonaniowe (wiedza „jak zrobić”) w najbardziej trafny sposób pozwalają ocenić umiejętność wykonania danej czynności. Na ich podstawie trudno jest jednak wnioskować, w jakim stopniu wiedza ta jest szeroka i uniwersalna. Przykładowo dana osoba doskonale opanowała określoną czynność, ale jeśli nie rozumie jej istoty, w nieco zmienionych warunkach nie będzie potrafiła adaptować tej wiedzy. Jeśli kandydat na kierowcę nauczy się przejeżdżać trasę z punktu A do punktu B, to nie znając przepisów ruchu drogowego, na nieco innej trasie (z B do C) może popełniać istotne błędy. Problem ten można częściowo wyeliminować, gdy zastosuje się odpowiednio dużą i wszechstronną próbkę wiedzy wykonaniowej. Ocena tej wiedzy w testach wykonaniowych wymaga również standaryzacji i określenia oczekiwanego poziomu wykonania (standard czy kryterium, z którym porównywane jest działanie i efekt działania osoby egzaminowanej). Zaleca się, aby elementem przygotowania takiego testu był nie tylko opis zadania, które ma przy asesorze zrealizować dana osoba, lecz także arkusz oceny, którym posługują się asesory. Podobnie jak w przypadku testów poznawczych, wypełnianych przez samego testowanego, arkusz oceny dla asesora powinien mieć formę szeregu konkretnych pytań na temat przebiegu działania osoby podchodzącej do walidacji

oraz skutków tych działań. Pytania powinny odnosić się do zjawisk dających się zaobserwować i obiektywnie ocenić. Częścią arkusza obserwacji jest też system punktacji, czyli klucz, na podstawie którego są przydzielane punkty lub jest formułowana jakościowa ocena opisowa. Sposób oceniania powinien być możliwie najbardziej:

- zobiektywizowany (wolny od subiektywnych ocen egzaminatora czy poziomu sympatii wobec kandydata),
- powtarzalny (a więc rzetelny) oraz
- trafny (odnoszący się do efektów uczenia się zawartych w opisie kwalifikacji).

Z kolei wiedza deklaratywna (typu „co zrobić”) pozwala dobrze określić poziom teoretycznej wiedzy kandydata. Pytając o nią, można określić, czy dana osoba dobrze rozumiała dane zjawisko, zna jego uwarunkowania, potrafi określić jego różne typy. Ten rodzaj pomiaru jest szczególnie przydatny, gdy od kandydata wymaga się wiedzy pamięciowej, nieintuicyjnej w samym działaniu, np. znajomości przepisów prawa. Na tej podstawie trudno jednak przewidzieć, czy potrafi on także dane czynności wykonać, a nie tylko je rozumieć lub o nich opowiedzieć.

Obie metody, pomiar testami wykonaniowymi oraz testami wiedzy, się uzupełniają. Pomiar wykonania określa rzeczywisty przebieg realizacji danych zadań, a test teoretyczny pozwala ustalić, w jakim stopniu wiedza ta jest uświadomiona. Warto dodać, że test wykonaniowy może obejmować prawdziwe czynności lub ich symulację. Na przykład od protokolanta oczekuje się sprawności zapisu przebiegu rozmowy. Zamiast jednak prosić o zapisanie rozmowy, osobie biorącej udział w walidacji można dyktować tekst, sprawdzając, jak szybko potrafi go zapisać. Zakłada się wtedy, że szybkość pisania ma związek z jakością protokolowania. Taki rodzaj sprawdzania umiejętności jest szczególnie przydatny, gdy dane zadanie trudno jest przeprowadzić w warunkach rzeczywistych. Wadą symulowania sytuacji, w której ma się ujawnić określona umiejętność, jest to, że obie sytuacje, prawdziwa i zaaranżowana, mogą nie być tożsame. Na przykład można sprawdzić, jak szybko i sprawnie kandydat notuje przebieg rozmowy, trudno jednak zweryfikować, jak będzie wyglądać protokolowanie ważnych i trudnych rozmów (np. w biznesie czy polityce), gdy dochodzi również czynnik stresu spowodowany wymagającą sytuacją.

Testy pisemne i ustne mają zalety i wady. Zaletą testów ustnych jest większa łatwość udzielenia odpowiedzi – z reguły prościej jest o czymś opowiedzieć niż napisać. Z odpowiedziami ustnymi zazwyczaj wiąże się jednak większy stres. Ich wadą jest także problem z archiwizacją przebiegu weryfikacji, a także ogólnie trudniejsza ocena tej odpowiedzi. Szczególnie w przypadku dłuższych odpowiedzi na tzw. pytania otwarte członkowie komisji weryfikacyjnej mogą pominąć niektóre wątki z odpowiedzi, różnie je rozumieć i inaczej ocenić ich jakość. Dlatego w przypadku takiej metody weryfikacji szczególnie ważne jest, aby wcześniej opracować klucz do oceny odpowiedzi – założyć, jakie elementy w wypowiedzi muszą się pojawić i jak będą oceniane. Istotne jest także wcześniejsze przeszkolenie asesorów. Oceniając wypowiedź ustną, warto posilkować się dosłownymi cytatami z wypowiedzi

kandydata i unikać własnych interpretacji (nie należy rozważać, co kandydat miał na myśli, co chciał powiedzieć, ale nie powiedział itd.).

W testach pisemnych również istotne są standaryzacja warunków, uprzednie skonstruowanie klucza do oceny odpowiedzi i inne procedury, które wpływają na jakość pomiaru wiedzy. W tym wypadku łatwiej jednak poddać etap oceny archiwizacji, pogłębionej refleksji czy krytycznej ocenie przez inny zespół asesorów. Szczególną pisemną metodą weryfikacji jest test pisemny z pytaniami zamkniętymi. To najbardziej oszczędna czasowo metoda weryfikacji. Nie jest jednak wolna od wad (np. opiera się na rozpoznawaniu dobrych odpowiedzi, a nie na aktywnym przywoływaniu ich z pamięci), a proces konstrukcji takiego testu wymaga dużego nakładu pracy, a więc czasu i środków na etapie jego przygotowywania.

3. Etyka stosowania testów

Kwestie etyczne stają się wyjątkowo istotne, gdy działania są podejmowane w obszarze, w którym określone postępowanie ma istotne i trwałe konsekwencje dla innych ludzi. Decyzja, podjęta na podstawie pomiaru testowego, aby uznać czyjś zakres wiedzy za wystarczający, np. w celu nadania kwalifikacji lakiernika samochodowego, może wpływać nie tylko na osobę uczestniczącą w walidacji, ale również – potencjalnie – również na bezpieczeństwo klientów korzystających z jego usług w przyszłości. Podobnie błędna decyzja o niepotwierdzeniu efektów uczenia się osobie przystępującej do testu będzie się wiązać z wieloma negatywnymi konsekwencjami ekonomicznymi, psychologicznymi i społecznymi nie tylko dla uczestnika walidacji, lecz także często dla jego bliskich. Ze względu na kompleksowość zagadnienia przyjęto dwa porządkujące założenia. Po pierwsze, poszczególne problemy etyczne omówiono w porządku chronologicznym zgodnym z „cyklem życia” testu – od jego tworzenia, przez realizację pomiaru i komunikowanie jego wyników, aż po procedury zapewniania jakości. Po drugie, prezentowane dalej rekomendacje nie mają charakteru wyczerpującego, ale przede wszystkim odnoszą się do tych aspektów walidacji, które w sposób specyficzny są związane ze stosowaniem testów wiedzy. Oczywiście jest, że cały proces powinien być przede wszystkim zgodny z ogólnymi zaleceniami dotyczącymi projektowania walidacji (Brzozowska i in., 2020), a także szczegółowymi zapisami sformułowanymi w opisie konkretnej kwalifikacji (Ziewiec-Skokowska i in., 2016)⁵.

Warto też zaznaczyć, że poniższe skrótowe uwagi w większości przypadków jedynie sygnalizują istnienie pewnych problemów, a także dobrych praktyk czy standardów, które pozwalają zapobiec występowaniu tych trudności.

3.1. Konstruowanie testu

Podstawowe etyczne wyzwania podczas tworzenia testu są związane z odpowiedzialnością za jakość tworzonego narzędzia, a także refleksją nad zasadnością samej decyzji o zastosowaniu tej właśnie formy weryfikacji efektów uczenia się jako najwłaściwszej w kontekście danej kwalifikacji. Instytucje opracowujące procedury walidacji muszą zdecydować, czy posiadane przez nie zasoby kompetencyjne są wystarczające do podjęcia decyzji co do optymalnej formy weryfikacji umiejętności kandydatów oraz stworzenia narzędzia walidacji odpowiedniej jakości. W konstruowanie testu, a także podejmowanie decyzji co do wyboru tej właśnie metody weryfikowania efektów uczenia się powinni być zaangażowani zarówno eksperci dziedzinowi, rozumiejący specyfikę danej kwalifikacji i dysponujący odpowiednią wiedzą specjalistyczną, jak i osoby mające umiejętności wymagane do stworzenia i oceny jakości powstającego narzędzia.

⁵ Osobom zainteresowanym poszerzeniem wiedzy związanej z problematyką stosowania testów, nie tylko w kontekście procesów walidacji, polecamy zapoznanie się z rekomendacjami i standardami omawianymi w publikacjach z zakresu psychometrii (np. Hornowska, 2001), w tym przede wszystkim opisanych w *Standardach dla testów stosowanych w psychologii i pedagogice* (AERA; APA; NCME, 2014).

Takie kompetencje mogą mieć m.in. psychometrycy, psychologowie, socjologowie czy pedagodzy o odpowiednim wykształceniu kierunkowym oraz posiadający praktyczne doświadczenie związane z konstruowaniem testów i zadań sprawdzających wiedzę i umiejętności. W przypadku wielu instytucji może to oznaczać konieczność sięgnięcia po pomoc ekspertów zewnętrznych. Jeżeli w czasie planowania tego działania pojawiają się wątpliwości, czy wsparcie eksperckie jest potrzebne, warto się zastanowić, czy w zespole zajmującym się tworzeniem testu obecne są osoby posiadające:

- doświadczenie w tworzeniu testów wiedzy;
- wiedzę ekspercką i doświadczenie w obszarze branży, do której odnosi się kwalifikacja;
- wiedzę z obszaru psychometrii i pomiaru edukacyjnego, w tym kompetencje statystyczne, pozwalające na ocenę podstawowych własności w zakresie trafności i rzetelności testu;
- wiedzę i umiejętności niezbędne do zaprojektowania poprawnego metodologicznie narzędzia walidacji oraz zapewniania jego jakości.

Jeżeli odpowiedź na któreś z wymienionych powyżej pytań jest negatywna, istotne wydaje się rozważenie uzupełnienia zespołu eksperckiego o dodatkowe kompetentne osoby.

Innym istotnym aspektem konstrukcji testu jest ustalenie, czy dobór osób zaangażowanych w ten proces nie naraża ich na wystąpienie konfliktu interesów. Przykładem takiej sytuacji może być oddelegowanie do tworzenia testu wiedzy pracowników, którzy w przyszłości będą się zajmować szkoleniem osób podchodzących do walidacji w tej samej IC. Nakładanie się ról i brak należytej dbałości o oddzielenie procesów szkolenia i walidacji (zwłaszcza w ramach tej samej instytucji) zawsze będzie narażał pracowników na istotne ryzyko etyczne, a także może prowadzić do problemów z uznaniem jakości nadawanych kwalifikacji. Szczegółowa znajomość treści testów może prowadzić do takiego kształtowania procesu nauczania i uczenia się, które będą zapewniać wysoką zdawalność, a nie zagwarantują rzeczywistego przyswojenia wymaganego zakresu wiedzy (problem ten można traktować jako skrajny przykład tzw. *washback effect*, omawianego w dalszej części opracowania). W kontekście projektowania testu warto zastanowić się:

- Czy (i w jakim stopniu) osoby zaangażowane w tworzenie testu będą zajmować się kształceniem osób przystępujących do walidacji?
- Czy i w jakim stopniu pytania testowe będą upubliczniane?
- Czy przebieg walidacji gwarantuje równe traktowanie wszystkich przystępujących do walidacji – w szczególności osób zdobywających kompetencje poza instytucją tworzącą narzędzia walidacji lub poza systemem oświaty i szkolnictwa wyższego?

3.2. Stosowanie testu

Stosując test, również można natrafić na sytuacje, w których ważne są kwestie etyczne. Przede wszystkim istotny jest sposób prowadzenia pomiaru testowego.

Zgodnie ze *Standards for Educational and Psychological Testing* (2014) osoby testowane mają w szczególności prawo do:

- Uzyskania pełnej informacji o celu testowania, formie i jego przebiegu.
- Uzyskania pełnej informacji o tym, jak zostaną wykorzystane zebrane wyniki, komu będą udostępnione i w jakim celu.
- Wyboru metody testowania, o ile twórca testu przewidział metody alternatywne (testy równoległe) i nie ma innych przeciwwskazań do tego, aby osoba egzaminowana mogła decydować o preferowanej formie testowania.
- Ochrony swoich danych osobowych, które podaje na potrzeby testowania.
- Ochrony poufności uzyskanego wyniku.
- Wyjaśnienia wszelkich niezrozumiałych dla niej pojęć, które pojawiają się w opisie wyniku jej testu.
- Otrzymania informacji w krótkim czasie od zakończenia weryfikacji, a jeśli czas ten miałby się wydłużyć, np. z powodu analizowania procesu weryfikacji pod kątem nieprawidłowości przez IC, okoliczności te powinny zostać przedstawione osobie testowanej.
- Do uczciwego traktowania, skorzystania z procedury odwoławczej oraz rzetelnego procesu zgodnie z konkretnymi okolicznościami w przypadku pojawienia się sporu dotyczącego wyników testowania pomiędzy kandydatem a komisją walidacyjną*.

Na osobie rozwiązującej test spoczywają też pewne obowiązki, o których powinna zostać poinformowana przez IC już na etapie przystępowania do walidacji. Obowiązki te to m.in.:

- staranność i szczerść w odpowiedziach (np. w testach pozapoznawczych, w których pytania dotyczą postaw, preferencji lub zachowania w hipotetycznych sytuacjach, mogących być pomocą przy doradztwie);
- traktowanie wszelkich materiałów związanych z sytuacją testowania w ramach walidacji jako poufnych.

* Ramki bez podania źródła zostały opracowane przez autorów.

Warto dodać, że treść testu jest objęta prawem autorskim. Dysponuje nim twórca testu lub IC, która go zatrudnia, zleca skonstruowanie danego testu lub wykupuje prawo do jego używania. Nie można więc bez uzyskania odpowiedniej zgody (np. wykupienia licencji, odkupienia praw autorskich) korzystać z gotowych testów lub ich fragmentów.

3.3. Wpływ wykorzystywania testów wiedzy na edukację oraz na sposób przygotowania się do walidacji

Dalej omówiono dwa zjawiska, które mają istotne znaczenie dla jakości pomiaru testowego, a jednocześnie właściwe poradzenie sobie z nimi wymaga opracowania odpowiednich środków zapobiegawczych już podczas projektowania testu. Pierwszym z nich jest tzw. *efekt testowania*. Punktem wyjścia do zrozumienia tego zjawiska może być następująca refleksja: korzystanie z dowolnego testu ma aspekt diagnostyczny (pozwala zmierzyć poziom wiedzy), ale jednocześnie jest też wydarzeniem, które w istotny sposób może zmieniać osobę podlegającą ocenie. Innymi słowy, uczestnictwo w teście można traktować jako specyficzne doświadczenie edukacyjne. Czego w takiej sytuacji może się nauczyć osoba testowana i jakie są tego niepożądane konsekwencje? Po pierwsze, uczestnictwo w teście jest okazją, aby lepiej zrozumieć sposób rozwiązywania danego rodzaju zadań. Osoby testowane mogą wypracować strategię udzielania odpowiedzi, lepiej rozumieją logikę konstrukcji pytań, optymalizują sposób zarządzania swoim czasem i wreszcie – oswajają się nieco z emocjami i sytuacją bycia testowanym. Po drugie, w czasie rozwiązywania testu można się przekonać, jakich treści dotyczą pytania, co w przyszłości pozwala zoptymalizować proces nauki i skoncentrować się na odpowiednich zagadnieniach. Dodatkowo należy też liczyć się z tym, że uczestnicy procesu walidacji będą wymieniać się informacjami na temat treści testów. Jeżeli zbiór stosowanych zadań będzie niewielki, a zadania proste do pamięciowego opanowania, to bardzo szybko zostanie on zrekonstruowany i upubliczniony.

Wymienione zjawiska prowadzą do podobnych niepożądanych konsekwencji i stanowią istotne zagrożenie dla jakości walidacji. W przypadku ponownego podejścia do testu uzyskiwane wyniki mogą być znacząco lepsze, ale zmiana ta nie będzie wynikać z wyższego poziomu wiedzy, ale ze znajomości testu (lub sytuacji testowania). Oznacza to zatem spadek trafności pomiaru i ryzyko podjęcia błędnej decyzji walidacyjnej przez komisję.

Problem „zużywania się” testów ma nie tylko wymiar indywidualny. Wprowadzenie ich jako metody weryfikacji wiedzy, zwłaszcza w dłuższym czasie, w istotny sposób może modyfikować także sam sposób organizowania kształcenia. Zjawisko modyfikacji praktyk dydaktycznych i programów nauczania pod wpływem stosowania testów znane jest pod angielską nazwą *washback effect* (Shohamy, 1992, 1993; Shohamy i in., 1996). Skrajnym przykładem tego efektu będzie sytuacja, w której osoba przygotowująca się do walidacji zamiast starać się zrozumieć przyswajane zagadnienia i uczyć się, jak stosować zdobywaną wiedzę, skupia się wyłącznie na pamięciowym opanowywaniu definicji czy wręcz pytań testowych. Pojawia się tu

uzasadniona wątpliwość, czy uzyskany wynik w czasie weryfikacji będzie odzwierciedlał rzeczywisty poziom wiedzy osoby testowanej, a nie jedynie skuteczność w radzeniu sobie z danym rodzajem testu.

W sytuacji idealnej uczenie się mające na celu zdobycie wiedzy, umiejętności oraz kompetencji społecznych, potrzebnych w danej branży, i przygotowywanie się do walidacji powinny być procesami jak najbardziej zbliżonymi. Duże rozbieżności, wynikające np. z pojawienia się *washback effect*, świadczą o niepoprawnie przygotowanych procedurach i metodach weryfikacji.

Sugestie i dobre praktyki pozwalające zredukować ryzyko związane z efektem testowania

- Zasady przeprowadzania testu i przykładowe pytania powinny być udostępniane wszystkim osobom zainteresowanym walidacją. Pozwala to wyrównać ich szanse i zmniejsza zakłócenia wynikające z lepszej znajomości procedury przez osoby testowane ponownie.
- W czasie układania testu warto pamiętać o tym, aby był on opatrzony nie tylko klarowną instrukcją, ale również – zwłaszcza w przypadku bardziej skomplikowanej konstrukcji zadań – odpowiednimi przykładami. Pozwala to zredukować efekty wynikające zarówno z wprawy zdobytej w powtórnym przystępowaniu do walidacji, jak i poprzez większe doświadczenia z testami w ogóle, np. w wyniku uczestnictwa w edukacji formalnej.
- Często stosowaną praktyką jest rozpoczynanie testów od pytań stosunkowo mniej trudnych, co redukuje zakłócenia pomiaru wynikające ze zróżnicowanej podatności kandydatów na stres związany z przystąpieniem do walidacji.
- Testy albo powinny mieć wersje równoległe (wersje testu o takich samych parametrach psychometrycznych, które mogą być użyte alternatywnie, a pomiar nimi będzie równie dobry), albo też wykorzystywać odpowiednio szerokie banki pytań, z których, w konkretnym procesie walidacji, losowana jest jedynie niewielka ich część. W istotny sposób zapobiega to możliwości przygotowania się kandydatów z ograniczonego zakresu treści, co do których wiadomo, że „będzie na teście”.
- Kosztownym, choć bardzo skutecznym zabiegiem pozwalającym poradzić sobie z problemem „zużywania się” testu jest przygotowanie tak dużego banku pytań (obejmującego zakres efektów uczenia się dla danej kwalifikacji rynkowej), że możliwe jest jego pełne upublicznienie. Przykładem takiego rozwiązania są testy stosowane w egzaminach teoretycznych na prawo jazdy.

- Minimalizacja *washback effect* jest problemem bardziej złożonym. Częściowo można ograniczać go przez właściwą konstrukcję testu, który powinien wymagać nie tylko mechanicznego przyswajania informacji, ale również zastosowania wiedzy w różnych kontekstach. Zadania tego typu, jeżeli są wystarczająco urozmaicone, mogą być dość skuteczną ochroną.
- Istotne jest też właściwe skomponowanie metod walidacji, tak aby test wiedzy był tylko jedną ze stosowanych metod. Testy takie uzupełniać powinny inne narzędzia i metody walidacji (np. obserwacja, próbki pracy, testy wykonaniowe czy analizy przypadków), które w mniejszym stopniu są obarczone ryzykiem wystąpienia *washback effect*.

4. Projektowanie testu wiedzy

4.1. Klasyczna teoria testów

Jakość testu najogólniej można zdefiniować jako siłę związku zachodzącego między wynikiem uzyskanym w procesie testowania a rzeczywistym poziomem wiedzy. Aby ująć tę relację precyzyjniej, można się odwołać do terminologii wprowadzonej przez tzw. klasyczną teorię testu (KTT) (por. np. Hornowska, 2001; Livingstone, 2004).

WYNIK TESTU to wynik obserwowany = wynik prawdziwy + błąd pomiaru

Zgodnie z założeniami klasycznej teorii testu wynik, który został uzyskany w teście (tzn. wynik obserwowany), jest sumą dwóch składowych: wyniku prawdziwego, czyli idealnie odzwierciedlającego poziom mierzonej wiedzy, oraz błędu pomiaru, który wynik ten zniekształca. Zgodnie z KTT wynik obserwowany, czyli uzyskany przez kandydata w czasie weryfikacji metodą testu, ma źródła w dwóch podstawowych procesach. Z jednej strony jest wyznaczany rzeczywistym poziomem mierzonej wiedzy. Przykładowo, jeżeli test ma mierzyć poziom znajomości przepisów prawa, to osoba przygotowana i mająca odpowiednią wiedzę powinna poradzić sobie z nim znacznie lepiej niż ktoś, kto nie ma wiedzy prawniczej. Z drugiej strony wpływ na wynik mają wszystkie czynniki, które zakłócają pomiar i powodują, że wynik obserwowany różni się od „prawdziwego”. Są to czynniki, które KTT określa łącznie jako „błąd pomiaru”. Na przykład niski lub wysoki poziom zmęczenia osoby wypełniającej test może podnosić lub obniżać jej wyniki. Podobnie może działać szereg czynników sytuacyjnych, np. poziom hałasu w czasie przeprowadzania testu. W obu tych sytuacjach zmianom w wynikach obserwowanych nie towarzyszą żadne różnice w faktycznych zdolnościach osób przystępujących do testu. Czynniki tego rodzaju nazywa się błędem losowym, jako że występują one w sposób zupełnie przypadkowy i nie wiążą się z charakterystyką osoby uczestniczącej w teście.

Inne błędy nie mają charakteru losowego, ale systematyczny. Są to czynniki, które zniekształcają uzyskiwane wyniki w określonym kierunku. Przykładem tego rodzaju sytuacji może być np. pominięcie w teście pytań dotyczących określonej grupy zagadnień. W przypadku weryfikacji efektów uczenia się będzie to pominięcie jakichś efektów uczenia się, które powinny być zweryfikowane testem. Takie rozwiązanie może skutkować np. wydaniem certyfikatu osobom, które nie osiągnęły efektów uczenia się w pominiętym w walidacji obszarze. Podobnie formułowanie pytań z zastosowaniem wyrafinowanego słownictwa, niemającego z testowaną wiedzą bezpośredniego związku, może powodować systematyczne obniżenie wyników osób o niższym poziomie wykształcenia (mimo że w danym teście kompetencje językowe nie powinny podlegać ocenie). Istotnym zadaniem osób konstruujących oraz stosujących później test jest ograniczenie wpływu poszczególnych

kategorii błędu na uzyskiwane w pomiarze wyniki, w tym bezwzględnie – dążenie do usunięcia lub zminimalizowania systematycznych źródeł błędów.

W dalszej części opracowania nieco dokładniej skoncentrowano się na poszczególnych charakterystykach testów, które determinują ich wysoką jakość. Często są łącznie określane mianem *kryteriów dobroci psychometrycznej*. Odwołując się do klasycznej teorii testu, można uznać, że spełnienie wszystkich tych kryteriów ma jeden zasadniczy cel: jak największe zmniejszenie błędu pomiaru, a zatem zapewnienie, że wynik uzyskany w procesie weryfikacji będzie potwierdzał wiedzę, umiejętności i kompetencje społeczne danej osoby, opisane w efektach uczenia się w danej kwalifikacji.

4.2. Teoria odpowiedzi na pozycje testowe

Klasyczna teoria testów nie jest jedyną ramą teoretyczną, którą dysponują współcześnie psychometrycy (Holland i Hoskens, 2002; Pokropek i in., 2015). Jedną z najbardziej wpływowych (i nowoczesnych) podejść w psychometrii jest teoria odpowiedzi na pozycje testowe (zazwyczaj określana angielskim akronimem IRT, *Item Response Theory*). IRT opiera się na zupełnie odmiennej filozofii w porównaniu z KTT. Kluczową różnicą jest to, że w czasie tworzenia testu w centrum uwagi jest nie tyle osoba podlegająca ocenie, ile same zadania testowe. Proces konstrukcji testu w ujęciu IRT polega na bardzo precyzyjnym określeniu własności szerokiej puli zadań, z których przy testowaniu jednej osoby zazwyczaj wykorzystuje się niewielką część. IRT, w porównaniu z podejściem klasycznym, ma wiele istotnych zalet. Najważniejszą z nich jest możliwość dostosowania zestawu zadań do poziomu mierzonej cechy osoby podchodzącej do walidacji (np. poziomu wiedzy), czyli konstruowanie tzw. testów adaptatywnych (ang. *adaptive testing*) (Pokropek, 2015). W konsekwencji IRT umożliwia osiągnięcie większej precyzji pomiaru i mierzenie zdolności w znacznie szerszym zakresie niż podczas wykorzystania KTT. Wykorzystanie IRT często pozwala też znacząco zredukować czas pomiaru wymagany do osiągnięcia danego poziomu precyzji pomiaru. Odwołując się do metafory, można powiedzieć, że IRT pozwala na dokładne „przyglądanie się” temu fragmentowi skali pomiarowej, w którym prawdopodobnie znajduje się mierzona osoba. Dzięki temu np. można uniknąć zbędnego zadawania bardzo prostych pytań osobie, która już wykazała się zaawansowaną wiedzą, lub bardzo trudnych pytań kandydatowi z deficytami poznawczymi (Stone i Davey, 2011). Klasyczna teoria testu nie daje takiej elastyczności.

Patrząc z perspektywy pragmatycznej, należy podkreślić, że bardziej wyrafinowany charakter modeli statystycznych, na których opiera się IRT, ma swoją cenę. Przede wszystkim ocena właściwości zadań wykorzystywanych w tego rodzaju testach wymaga zgromadzenia ogromnych ilości danych, co w procesie tworzenia testu przekłada się na znaczące koszty czasowe i finansowe. Nie bez znaczenia jest też konieczność angażowania w proces konstrukcji testu i opracowania wyników ekspertów mających specjalistyczne kompetencje statystyczne i psychometryczne. Podsumowując, wykorzystanie IRT poleca się w sytuacjach, gdy jest możliwość zainwestowania znacznych środków w konstrukcję testu

i jednocześnie można oczekiwać wysokiej stopy zwrotu z tej inwestycji (np. skrócenie czasu realizacji testów może przynieść istotne oszczędności finansowe), lub planuje się testowanie dużych grup kandydatów (zakłada się ograniczenie „zużywania się” testu, wynikającego z powtarzania się pytań i ryzyka ich upublicznienia) (więcej szczegółowych informacji na temat tego podejścia można znaleźć w źródłach na końcu opracowania).

5. Kryteria dobroci psychometrycznej testu

5.1. Rzetelność

Pojęcie rzetelności, rozumiane w kontekście tworzenia testów, jest stosunkowo zbliżone do definicji, jaką ma ono w języku potocznym. „Rzetelny” może znaczyć, w zależności od kontekstu, m.in. solidny, stabilny, powtarzalny czy niesprawiający nieprzyjemnych niespodzianek. Dokładnie tego samego należy oczekiwać od wyników testów. Uzyskiwane pomiary powinny cechować się jak największą precyzją. Wyniki rzetelnego testu nie powinny się zmieniać, jeżeli nie zmienia się osoba testowana. Jednocześnie nawet niewielka modyfikacja w zakresie mierzonej wiedzy w teście o wysokiej rzetelności powinna mieć swoje odzwierciedlenie w uzyskiwanych w nim wynikach.

Rzetelność jest fundamentem wszystkich charakterystyk omawianych w dalszej części tego rozdziału. Jeżeli realizowany pomiar jest mało precyzyjny, trudno mówić np. o jego obiektywizmie czy trafności. Odwołując się do metafory, można powiedzieć, że korzystanie z testu o niskiej rzetelności jest jak słuchanie czyjejś wypowiedzi z bardzo słabego nagrania. Jest dużo szumu, niewiele można zrozumieć i na pewno nie da się ocenić, w jakim stopniu trafne czy obiektywne są informacje przekazywane przez opowiadającego. Rzetelność należy do tych charakterystyk testów, które można wyrazić także liczbowo. Najczęściej stosowaną miarą jest tu współczynnik *alfa Cronbacha*, nazywany od nazwiska twórcy. Przyjmuje on zawsze wartości z przedziału między 0 i 1, gdzie 1 oznacza maksymalną precyzję pomiaru, stan idealny i – w pomiarze psychologicznym czy pedagogicznym – właściwie nigdy niewystępujący. Rzetelność równa 0 stanowi drugą skrajność, czyli opisuje taki test, który w rzeczywistości niczego nie mierzy – na uzyskiwany wynik składają się wyłącznie czynniki losowe, niemające ze sobą żadnego systematycznego związku (Hornowska, 2011). Testy wykorzystywane w praktyce do diagnozy indywidualnej (a z taką mamy do czynienia podczas weryfikacji) powinny charakteryzować się wartościami rzetelności przekraczającymi poziom 0,8 (więcej szczegółów na temat różnych metod ilościowego szacowania rzetelności patrz źródła na końcu opracowania).

Z perspektywy praktycznej kluczowe jest pytanie, jak można zadbać o wysoką rzetelność tworzonych przez nas testów wiedzy. Co o niej decyduje? Istotna jest liczba wykorzystywanych w teście pytań czy też zadań. Co do reguły, dłuższe narzędzia cechują się wyższą rzetelnością pomiaru. W szczególności ma to znaczenie w przypadku stosowania pytań w formacie zamkniętym, czyli takich, gdzie wybrana odpowiedź (czy też ich zestaw) to jedyne informacje, które uzyskuje się od osoby testowanej. Ogólnie, im mniej informacji dostarcza osoba udzielająca odpowiedzi w pojedynczym zadaniu, tym więcej tego rodzaju zadań należy uwzględnić w teście. Przykładowo, jeżeli użyje się najprostszego formatu zadań wymagających tylko wyboru odpowiedzi pomiędzy opcjami „tak” lub „nie”, osiągnięcie satysfakcjonującej rzetelności bardzo często będzie obligowało do zastosowania nawet 50 czy 60 pytań w teście. W przypadku pytań otwartych, zwłaszcza jeżeli są

wykorzystywane podczas testów ustnych, już kilka rozbudowanych odpowiedzi może pozwolić na precyzyjne oszacowanie poziomu wiedzy kandydata.

Poza kwestią długości walidacji pozytywny wpływ na rzetelność pomiaru mają też wszystkie te działania, które pozwalają zredukować wpływ czynników losowych na uzyskiwany wynik przez osobę testowaną. Chodzi tu o cechy stosowanego testu (np. klarowność i jednoznaczność instrukcji), okoliczności przeprowadzania testu (np. zachowanie ciszy) czy zachowania osób zaangażowanych w jego realizację (więcej szczegółów na ten temat zawierają kolejne podrozdziały, zwłaszcza te dotyczące standaryzacji i obiektywizmu pomiaru).

5.2. Trafność

Trafność testu gwarantuje, że za pomocą danego narzędzia można zmierzyć tę cechę czy własność, która nas interesuje. Rzetelność to informacja o tym, jak dokładnie potrafimy coś zmierzyć, a trafność mówi nam o tym, czy mierzymy to, co powinniśmy. Trafność jest też najbardziej złożonym spośród wszystkich omawianych w tym rozdziale kryteriów oceny testów. Wziąwszy pod uwagę kontekst zastosowania testów wiedzy, tj. walidację w ramach ZSK, zakres informacji zostanie ograniczony do omówienia dwóch aspektów trafności, które mają kluczowe znaczenie dla konstrukcji i stosowania testów wiedzy na etapie weryfikacji, a mianowicie trafności treściowej oraz diagnostycznej.

5.2.1. Trafność treściowa

Trafność treściowa określa, w jakim stopniu zakres wiedzy mierzonej za pomocą danego testu jest reprezentatywny dla całości określonego zagadnienia (np. jej zakresu wyznaczonego zestawami efektów uczenia się w opisie kwalifikacji). Znaczenie tej definicji najlepiej można zrozumieć, korzystając z przykładu: jeżeli ma zostać poddana ocenie czyjaś wiedza, np. na temat historii Albanii, to nie należy pytać o zdarzenia dotyczące innego kraju, niemającego z Albanią nic wspólnego. Pytania takie cechowałyby się bardzo niską trafnością treściową. Podobnie jeżeli cały test ma zawierać 20 pytań, to dobrze byłoby zadbać o to, aby poszczególne z nich dotyczyły różnorodnych aspektów polskiej historii (politycznej, kultury, gospodarczej itd.) oraz różnych okresów historycznych. Trafne treściowo testy po pierwsze powinny zawierać wyłącznie pytania powiązane w istotny sposób z określonym zakresem wiedzy. Po drugie, powinno się zagwarantować, że treść pytań odzwierciedla różne obszary czy aspekty danej dziedziny. Właściwie żaden test nie ma charakteru wyczerpującego, nie obejmuje całości wiedzy, którą powinna mieć osoba testowana. Dlatego tak istotne jest znaczenie właściwego jej doboru – tak aby test był reprezentatywnym odzwierciedleniem tego zakresu. W sytuacji przygotowywania testów wiedzy wykorzystywanych w walidacji kluczowym zadaniem jest więc staranne wyznaczenie, co dokładnie powinna wiedzieć osoba ubiegająca się o uzyskanie danej kwalifikacji. Głównym źródłem tej wiedzy jest poprawnie przygotowany opis kwalifikacji, a konkretnie – efekty uczenia się oraz wymagania dla walidacji. Efekty uczenia się, opisane w danej kwalifikacji, które nie będą sprawdzane testem wiedzy, oraz opis wymagań dla walidacji determinują,

w jakich kontekstach kandydat powinien umieć z wiedzy skorzystać, co w istotny sposób będzie się przekładać na typ problemów czy zadań, które warto uwzględnić w teście.

Dopiero po dokonaniu tego rodzaju analizy warto przystąpić do przygotowywania pytań i zadań. Na tym etapie prac często zapadają decyzje dotyczące proporcji, w których poszczególne zagadnienia będą reprezentowane w finalnej wersji testu. Naturalnie wiedza, która ma znaczenie kluczowe i jest wykorzystywana intensywniej, powinna być weryfikowana większą liczbą pytań niż zagadnienia marginalne.

Opracowanie trafnych treściowo pozycji nierzadko wymaga zastosowania procedury dwuetapowej. Najpierw poszczególne zadania są układane przez ekspertów dziedzinowych przy ewentualnym wsparciu osoby mającej odpowiednią wiedzę i doświadczenie w konstrukcji testów. Podczas drugiego etapu przygotowana pula pytań podlega weryfikacji przez (przynajmniej od dwóch do pięciu) niezależnych oceniających (tzw. sędziów kompetentnych lub asesorów), którzy oceniają, w jakim stopniu przygotowane pytania rzeczywiście można traktować jako trafną reprezentację efektów uczenia się. Do pilotażowej wersji testów zazwyczaj są włączane tylko takie pozycje testowe, które nie budzą istotnych wątpliwości czy rozbieżnych opinii niezależnych ekspertów.

5.2.2. Trafność diagnostyczna

Testowanie nigdy nie jest celem samym w sobie. Jeżeli testy spełniają swoją funkcję, pozwalają zidentyfikować osoby mające daną wiedzę, określone umiejętności czy kompetencje społeczne. Tę właśnie cechę określa się *trafnością diagnostyczną narzędzia*.

W przypadku walidacji cel diagnostyczny jest jasny. W świecie idealnym proces ten pozytywnie przechodzą wszystkie (i tylko te) osoby, które reprezentują wymagany poziom wiedzy, umiejętności czy kompetencji społecznych.

Jak można zagwarantować, że stworzony przez IC test będzie spełniał tak zdefiniowane kryterium trafności? Po pierwsze, należy pamiętać, że żadne narzędzie nie będzie diagnostycznie trafne, jeżeli nie zapewnia odpowiedniej precyzji pomiaru (czyli rzetelności). Po drugie, trafność diagnostyczna jest też uwarunkowana trafnością treściową, a także (choć pośrednio) poprawnością i precyzją opisu efektów uczenia się w danej kwalifikacji.

Po trzecie, ograniczając już refleksję tylko do weryfikacji wiedzy, kluczowe znaczenie ma trafna identyfikacja niezależnych obszarów czy tematów, które powinny być potwierdzone w ramach walidacji. Jej specyfiką jest konieczność osiągnięcia w wystarczającym stopniu wszystkich wymaganych efektów uczenia się. Oznacza to, że bardzo często nawet sama weryfikacja wiedzy będzie wymagać niezależnej oceny kilku obszarów. Przykładowo osoba posiadająca kwalifikację z zakresu gotowania pewnie musi dysponować zarówno wiedzą dotyczącą przygotowania potraw i ich receptur, jak i informacjami dotyczącymi zasad BHP pracy w kuchni. Nie ma tu możliwości kompensacji braków w jednym obszarze przez dodatkową wiedzę w drugim. Trafność pomiaru będzie uwarunkowana także doбором rodzajów pytań czy zadań wykorzystywanych w testach wiedzy (przykłady dobrych

i złych praktyk w przygotowywaniu pytań opisano w dalszej części opracowania). Wreszcie *trafność diagnostyczna* w istotnym stopniu jest uwarunkowana sposobem oceniania udzielanych odpowiedzi i kryteriami decyzyjnymi.

Ile punktów powinien zdobyć kandydat w każdym z ocenianych obszarów, aby uznać, że jego wiedza jest wystarczająca? (Zasady ustalania progów punktowych opisano dokładniej w dalszej części opracowania). Trafność diagnostyczna jest też funkcją sposobu korzystania z testu. Tracimy ją, gdy test nie jest przeprowadzany zgodnie z instrukcją, materiały testowe są upubliczniane, stosowane pytania tracą aktualność, a także (co istotne zwłaszcza w przypadku form testowania wymagających aktywnego udziału asesora), gdy nie są przestrzegane standardy dotyczące szkoleń i wymagań wobec osób przeprowadzających walidację. Trafna identyfikacja tego rodzaju zagrożeń i zapobieganie ryzyku ich wystąpienia przez IC stanowią uzasadnienie dla tworzenia i stosowania odpowiednich zapisów służących wewnętrznemu zapewnianiu jakości.

5.3. Obiektywizm

Obiektywizm, jako wymaganie stawiane przed dobrymi narzędziami walidacji, można interpretować jako dosłowne rozszerzenie znaczenia tego słowa funkcjonującego w codziennej polszczyźnie. Obiektywizm jest przeciwieństwem subiektywizmu, czyli zależności ocen od preferencji, cech czy osobistych przekonań. W kontekście testowania oznacza, że uzyskiwane wyniki, a przede wszystkim opierające się na nich decyzje asesorów, nie są w żaden sposób uwarunkowane ani właściwościami osób uczestniczących w procesie weryfikacji (np. w roli asesora lub członka komisji walidacyjnej), ani też tymi cechami kandydata, które nie mają merytorycznego znaczenia z perspektywy walidacji, np. płeć lub wyznanie. W tym kontekście warto pamiętać, że jednym z założeń ZSK jest zapewnienie równych szans wszystkim osobom przystępującym do walidacji, niezależnie od tego, jaką drogą zdobyły wiedzę, kompetencje społeczne czy umiejętności. W szczególności w walidacji tak samo powinni być traktowani uczestnicy mający wykształcenie formalne, jak i ci, którzy wiedzę czy umiejętności zdobyli w innych okolicznościach. A zatem, konstruując zadania, należy unikać korzystania z przykładów czy też języka, które mogłyby dyskryminować jedną z wymienionych grup. Specjalistyczna terminologia czy żargon zawodowy, wykorzystywany głównie przez praktyków, powinny być ograniczone wyłącznie do tych sytuacji, gdy taka wiedza jest istotną składową danego efektu uczenia się.

Opracowując test wiedzy, należy też pamiętać, że nie wszystkie osoby przystępujące do walidacji dysponują takim samym kapitałem kulturowym. Może to dotyczyć nie tylko samej znajomości polskiej historii czy obyczajów, ale również kwestii znacznie bardziej fundamentalnej, tj. stopnia opanowania języka. Ze względu na zachodzące w Polsce zmiany społeczno-demograficzne w nadchodzących latach należy się spodziewać, że polski nie będzie ojczystym językiem dla coraz większego odsetka osób przystępujących do walidacji. Jednocześnie, dla tej właśnie grupy osób, możliwość uzyskania certyfikatów potwierdzających nabycie kwalifikacji

często będzie stanowić bardzo atrakcyjne rozwiązanie poprawiające pozycję na rynku pracy.

Oznacza to, podobnie jak w przypadku specjalistycznej terminologii, że konstruując zadania testowe, należy rozważyć, czy (i w jakim stopniu) są one osadzone w specyficznym polskim kontekście czy polskich doświadczeniach, które mogą być nieznane np. osobom wychowanym w innym kręgu kulturowym, ale też osobom z określonej grupy społecznej czy innego pokolenia. Wymagania wobec kandydatów (także te językowe) muszą być adekwatne do opisu kwalifikacji. Skuteczne funkcjonowanie w wielu rolach zawodowych wymaga płynnego komunikowania się, nie oznacza to jednak, że w procesie walidacji dodatkowo premiowane ma być posługiwanie się literacką polszczyzną. Praktyki „wymuszające” na osobie testowanej znajomość treści wykraczających poza wymagania wyrażone wprost w opisie kwalifikacji często stanowią ukrytą formę dyskryminacji. W zależności od profilu tych „dodatkowych” wymagań grupami dyskryminowanymi mogą być np. osoby starsze, pochodzące z innych kultur, o niższym statusie socjoekonomicznym czy osoby z niepełnosprawnością. Problem ten jest określany w literaturze specjalistycznej jako *stronniczość testu*. Narzędzia obciążone tego rodzaju błędem po pierwsze będą mniej trafne (pomiar nie odzwierciedla tego, co powinien, tylko jest uzależniony od wpływu dodatkowych czynników), a po drugie ich wykorzystywanie będzie prowadzić do podejmowania krzywdzących i błędnych decyzji walidacyjnych. W tym kontekście nie powinno dziwić, że dbałość o ten wymiar jakości walidacji jest niezwykle istotna. Oczywiście problematyka obiektywizmu pomiaru testowego nie ogranicza się do konstrukcji zadań testowych.

Jak jeszcze można zadbać o to, aby tworzone narzędzia były obiektywne? W dużym stopniu będzie to zależało od rodzaju testu. Ogólnie ryzyko zagrożeń obiektywizmu jest zawsze nieco większe, gdy dochodzi do relacji między osobą podchodzącą do walidacji a asesorem. Im mniejszy ma ona zakres, tym mniej jest okazji do wystąpienia zjawisk mogących wpłynąć negatywnie na obiektywizm w czasie weryfikacji. Dlatego też testy mające postać pisemną zmniejszają tego rodzaju ryzyko. W jeszcze mniejszym stopniu problem ten występuje w kontekście testów w formie elektronicznej, gdy prezentowanie instrukcji, zbieranie danych czy pomiar czasu wykonania są zautomatyzowane. Warto pamiętać, że brak obiektywizmu może się ujawniać w sposób bardzo subtelny (np. jako niestaranne prezentowanie instrukcji słownych), może też ograniczać się do sygnałów niewerbalnych, np. manifestacji zniecierpliwienia czy znudzenia przez asesora przeprowadzającego test ustny. Aby zapobiegać tego typu problemom, konieczne jest jak najbardziej precyzyjne formułowanie i przestrzeganie instrukcji dla osób przeprowadzających test. W tym przypadku obiektywizm łączy się ściśle ze – omawianą w kolejnym punkcie – standaryzacją narzędzi walidacji.

Naruszenia zasady obiektywizmu mogą też występować podczas analizy uzyskanych wyników w teście. Tu kluczowe znaczenie mają zarówno kompetencje i jakość szkolenia asesorów oraz innych osób przeprowadzających test, jak i jakość klucza używanego w kodowaniu odpowiedzi i rozwiązań zadań testowych. Wymagania w tym względzie w dużym stopniu będą zależały od zastosowanej formy zadań. Poprawne obliczenie wyników w teście wykorzystującym jedynie pytania lub zadania zamknięte (ze ściśle określonymi odpowiedziami do wyboru) jest czynnością stwarzającą minimalne ryzyko.

W przypadku pytań lub zadań otwartych pole do interpretacji (a więc także zmniejszania obiektywizmu) wyników jest znacznie większe. Tu też z pomocą powinny przychodzić, przygotowane przez autorów testu, ściśle określone zasady interpretacji w postaci kryteriów oceny odpowiedzi (tzw. klucza). Warto też pamiętać o psychologicznych procesach mogących zakłócać obiektywizm formułowanych ocen i systematycznie przyczyniać się do ich zniekształcenia. Takie sytuacje częściej mogą występować w przypadku procedur, w których dochodzi do bezpośredniej interakcji pomiędzy osobą podchodzącą do walidacji a asesorami. Między innymi wówczas może ujawnić się tzw. *efekt aureoli*, gdy pozytywna ocena osobowości uczestnika procesu walidacji (np. kompetencje interpersonalne egzaminowanego, podobieństwo do asesora, rodzaj niepełnosprawności, status społeczny itd.) może – nawet w sposób nieuświadomiony – wpływać na sposób oceniania odpowiedzi. Ponownie ograniczenie ryzyka występowania tego rodzaju możemy uzyskać poprzez staranne określenie instrukcji, procedur, kryteriów oceny oraz dbałość o ich przestrzeganie. W przypadku oceny wielu kandydatów i pisemnej metody weryfikacji przydatna może być też anonimizacja odpowiedzi na etapie ich punktowania (np. przez przypisanie poszczególnym osobom piszącym test losowo generowanych kodów) lub też wykorzystywanie testów w formie elektronicznej, które są sprawdzane automatycznie przez program komputerowy.

5.4. Standaryzacja

Standaryzację testu najkrócej można zdefiniować jako zagwarantowanie stałości sposobu jego realizacji. Wymaganie to przekłada się na wiele różnych działań, które łącznie zakładają, że wszyscy uczestniczący w procesie walidacji będą traktowani w podobny sposób. Zagwarantowanie standaryzacji jest więc ważne zarówno dla zapewnienia rzetelności pomiaru (minimalizuje ryzyko występowania różnych losowych i systematycznych zakłóceń), jak i jego obiektywizmu (jeżeli weryfikacja jest prowadzona w podobny sposób, zmniejsza się potencjalne zniekształcenia wynikające z charakterystyk osób uczestniczących w procesie walidacji). Dzięki temu standaryzacja testu zwiększa trafność podejmowanych na jego podstawie decyzji diagnostycznych. Opisanie, wdrożenie i utrzymanie odpowiednich standardów testowania jest wspólną odpowiedzialnością autorów testu oraz instytucji certyfikujących i walidujących. Po stronie autorów spoczywa obowiązek na tyle precyzyjnego i zarazem realistycznego zdefiniowania warunków przeprowadzenia weryfikacji za pomocą testu, aby było możliwe odtworzenie ich w zbliżonej formie podczas realnej walidacji. W wielu aspektach są to czynniki oczywiste, takie jak np. zapewnienie odpowiednich warunków przestrzennych, akustycznych (niski poziom hałasu, nieobecność dystraktorów) czy przygotowanie stanowisk spełniających standardy ergonomiczne (np. w zakresie dobrego oświetlenia stanowisk pracy w czasie testu pisemnego). Należy też podkreślić standard zapewnienia możliwości realizacji testów osobom z różnymi kategoriami niepełnosprawności (np. sensorycznej lub motorycznej), co powinno znaleźć odzwierciedlenie zarówno w opisie narzędzia, jak i w przygotowaniu instytucji przeprowadzających walidację (por. np. wskazówki CKE, 2020).

Standaryzacja powinna też obejmować dalsze etapy opracowywania wyników, zwłaszcza proces przekształcania odpowiedzi na wskaźniki ilościowe oraz podejmowania na ich podstawie decyzji przez asesorów i komisję walidacyjną.

Niewątpliwie istotną składową standaryzacji jest także dbałość o dobór i szkolenie asesorów oraz stały nadzór nad jakością ich pracy. Jako że w grę wchodzi tu zmienne związane z tzw. czynnikiem ludzkim (np. sumienność asesorów, błędy), a często także pokusy IC natury ekonomicznej (np. redukcji kosztów szkoleń czy rezygnacji z angażowania trzyosobowych zespołów asesorskich), bardzo istotne jest też opracowanie i wdrożenie procesów monitoringu i ewaluacji stosowanych metod walidacji przez IC oraz pracy asesorów.

5.5. Normalizacja

Jedną z głównych zalet wynikających ze stosowania testów jest możliwość uzyskania po ich przeprowadzeniu liczbowych wskaźników określających poziom wiedzy (czy nasilenia jakiejś cechy mierzonej danym narzędziem). Jak jednak powinniśmy takie wskaźniki interpretować? Czy jeżeli wiadomo, że osoba podchodząca do bliżej nieznanego nam testu uzyskała 20 poprawnych odpowiedzi na 24 możliwe, to czy dostarcza nam to jakichś informacji na temat poziomu jej wiedzy? Intuicja (słusznie) każe w to wątpić. Przykład ten wyraźnie wskazuje na podstawową własność testów – znaczenie uzyskiwanych w nich wskaźników ujawnia się wyłącznie wówczas, gdy dysponuje się dodatkowym kontekstem umożliwiającym ich interpretację. W zależności od celu i kontekstu stosowania danego narzędzia różne mogą być źródła „tła”, aby nadać znaczenie uzyskanym wskaźnikom.

Kompetencje, wiedza i umiejętności osoby przystępującej do walidacji mają być ocenione względem określonego kryterium (efektów uczenia się), stąd też nazwa tego rodzaju narzędzi określanych jako *testy kryterialne* (ang. *criterion-referenced test*). Korzystając z nich, weryfikuje się, czy kandydat w sposób wystarczający opanował określone opisanie kwalifikacji wiedzy czy umiejętności i kompetencje społeczne. W obszarze zainteresowania ani nie jest to, czy dany wynik jest typowy (np. dla pozostałych kandydatów), ani czy przekracza on bardzo znacząco zakładany w opisie kwalifikacji poziom wymagań (walidacja nie przewiduje bowiem żadnych „wyróżnień”). Należy podkreślić, że testy kryterialne nie wymagają opracowania norm, ponieważ bardzo często właściwy poziom ich realizacji można określić względem obiektywnych standardów dotyczących np. poziomu wykonania określonych czynności czy opanowania pewnego zakresu wiedzy zgodnie z opisem zawartym w danej kwalifikacji.

W określaniu kryteriów satysfakcjonującego wykonania zadań lub odpowiedzi na pytania testowe istotną rolę odgrywają eksperci, którzy w czasie tworzenia testu dokonują „tłumaczenia” (operacjonalizacji) opisu kwalifikacji na język pomiaru.

Omówione kryteria *dobroci psychometrycznej* nie wyczerpują oczywiście całości tego zagadnienia, bardzo złożonego i badanego od wielu lat przez specjalistów.

W podręcznikach psychometrii i pomiaru dydaktycznego można znaleźć jeszcze szereg innych jej aspektów, które nie mają bezpośredniego zastosowania w sytuacji tworzenia testów wiedzy wykorzystywanych w weryfikacji efektów uczenia się.

5.6. Adaptacja testów

Konstruowanie testu jest kosztowne i czasochłonne. To złożony proces. Samo zapoznanie się z listą kryteriów *dobroci psychometrycznej* może być doświadczeniem nieco deprymującym, bo wyraźnie sugeruje, że stworzenie dobrego narzędzia do weryfikacji efektów uczenia się jest zadaniem trudnym. Pojawia się wobec tego pytanie, czy w procesie walidacji możliwe jest wykorzystanie istniejących już narzędzi, które z sukcesami są stosowane np. za granicą. Takie rozwiązanie ma wiele zalet, gdyż z pozoru jedyną konieczną czynnością jest dokonanie tłumaczenia materiałów testowych i tych przeznaczonych dla asesorów. Niestety w rzeczywistości proces adaptacji jest wieloetapowy i złożony. Samo słowo adaptacja (a nie „tłumaczenie”) sugeruje już, że nie jest to proste przełożenie treści zestawu pytań na inny język.

Przed podjęciem wysiłków związanych z realizacją adaptacji testu przede wszystkim należy się upewnić, że jest to proces możliwy do przeprowadzenia od strony prawnej. Pierwszym krokiem powinno być tu uzyskanie zgody na adaptację od instytucji rozporządzającej prawami autorskimi do danego testu. Jako że narzędzia diagnostyczne podlegają ochronie prawnej, korzystanie z adaptacji bez tego rodzaju zgody będzie nielegalne.

Kolejną decyzją, która powinna poprzedzać podjęcie prac nad adaptacją, jest rozważenie, czy przeprowadzenie adaptacji jest przedsięwzięciem racjonalnym z punktu widzenia kosztów czasowych, finansowych itd. Adaptacja jest działaniem opłacalnym wówczas, gdy wysiłek włożony w stworzenie polskojęzycznej wersji narzędzia będzie mniejszy niż stworzenie analogicznego narzędzia od podstaw. Należy też podkreślić, że adaptacja testu wymaga zaangażowania specjalistów mających szerokie kompetencje – z jednej strony muszą bardzo dobrze znać zarówno język oryginalnego testu, jak i język polski, z drugiej zaś dysponować wiedzą przedmiotową, pozwalającą na podejmowanie świadomych decyzji w przypadku konieczności modyfikacji treści zadań. Sam proces adaptacji ma charakter złożony i składa się z wielu etapów. W rozbudowanej wersji obejmuje m.in. takie działania, jak:

- opracowanie kompletu materiałów testowych (instrukcji, treści zadań, instrukcji dla asesorów itd.);
- przeprowadzenie badań pilotażowych pozwalających sprawdzić, czy przetłumaczone materiały testowe są zrozumiałe dla kandydatów;
- opracowanie tłumaczenia zwrotnego, w którym test jest ponownie tłumaczony z języka polskiego na język oryginału, a następnie sprawdzenie, czy

tak przetłumaczony test daje wyniki odpowiednio zbliżone do oryginalnego narzędzia;

- przeprowadzenie badań pilotażowych z udziałem osób dwujęzycznych, które rozwiązują obie wersje testu – tu także należy się spodziewać uzyskania podobnych wyników w obu wersjach.

Wszystkie te działania mają przede wszystkim na celu sprawdzenie, czy test po adaptacji nadal spełnia kryteria *dobroci psychometrycznej*, które cechowały jego oryginalną wersję. Z punktu widzenia procesu walidacji kluczowe znaczenie ma tu aspekt trafności, a zatem odpowiedź na pytanie, czy polska wersja testu mierzy to, co powinna, wzięwszy pod uwagę założenia twórców oryginalnego testu.

Choć trudno tu udzielać bardzo uniwersalnych porad, wydaje się, że testy wykorzystywane w kontekście ZSK warto adaptować przede wszystkim wówczas, gdy zarówno zakres wiedzy wymaganej w kwalifikacji, jak i kontekst późniejszej aktywności związanej z jej uzyskaniem są zbliżone. Przykładem takiej korzystnej sytuacji mogą być np. testy wykorzystywane do weryfikacji wiedzy dotyczącej urządzeń technicznych czy oprogramowania. Zasady funkcjonowania określonego modelu tokarki czy systemu operacyjnego komputera, a także sposób ich wykorzystywania nie będą bowiem znacząco się różniły niezależnie od języka czy kraju, w którym są one używane. Podobnie niezmienny kulturowo charakter będzie miała np. wiedza dotycząca uniwersalnych zjawisk występujących w przyrodzie (np. ogólnej fizjologii człowieka). Jednocześnie jednak, nawet jeżeli wiedza ogólna zdaje się niezmienna, już jej stosowanie w określonym kontekście może podlegać bardzo istotnym modyfikacjom wynikającym z czynników specyficznych dla danego kraju. Dobrym przykładem może być tu np. testowanie wiedzy dotyczącej dietetyki. Co prawda, ogólne zasady odnoszące się do właściwego żywienia człowieka są stosunkowo uniwersalne, ale konkretne wytyczne będą zależały już od regulacji i zaleceń, czynników kulturowych, obyczajowych, różnic ekonomicznych, klimatu, dostępności produktów żywnościowych itd. W przypadku testu weryfikującego umiejętność doboru optymalnej diety proces adaptacji będzie więc bardzo złożony i często będzie wymagał rekonstrukcji większości oryginalnych zadań.

Przedstawione w tym rozdziale informacje mają charakter bardzo skrótowy. Są one jednak wystarczającą zachętą do starannego rozważenia wszystkich ryzyk, które wiążą się z pochopnymi próbami tłumaczenia testów bez uwzględnienia całej złożoności zagadnienia, a zwłaszcza jego niebagatelnymi kosztów. Wysiłek włożony w proces adaptacji testu nierzadko porównywalny jest z tym, którego wymaga skonstruowanie narzędzia od zera⁶.

⁶ Staranniejse omówienie problematyki adaptacji kulturowej narzędzi można znaleźć m.in. w pracy Hornowskiej (2001) oraz Anastasi i Urbiny (1999).

6. Konstruowanie testów wiedzy

6.1. Ogólne zasady istotne w konstrukcji zadań testowych

Zanim zostaną przedstawione szczegółowe uwagi odnoszące się do różnych typów zadań wykorzystywanych w testach wiedzy, omówiono, zagadnienia natury ogólnej, które mogą mieć istotne znaczenie dla jakości narzędzi walidacji opracowywanych przez IC.

Pierwsza z kwestii ma charakter praktyczny i jest to rekomendacja dotycząca sposobu zarządzania procesem tworzenia testów. Testy wiedzy w gotowej postaci często są narzędziami mało „efekownymi” – bywa, że ostateczna ich forma to kilka dokumentów zawierających arkusze z pytaniami, klucz odpowiedzi i kryteria oceny. Niestety można je bez trudu skopiować, sfotografować czy udostępnić w Internecie, nierzadko niszcząc tym samym ich diagnostyczną wartość. Sprawia to, że osoby niemające odpowiedniego doświadczenia i należytej wiedzy psychometrycznej na ogół nie doceniają ilości zasobów kadrowych, czasowych i finansowych, które są wymagane do opracowania takich narzędzi walidacji. Warto pamiętać, że siła testu nie tkwi w jego zewnętrznej formie, ale we własnościach psychometrycznych, które gwarantują, że uzyskane za jego pomocą informacje są wiarygodną podstawą do podejmowania decyzji walidacyjnej. Jeżeli IC chce zadbać, żeby opracowane przez nią narzędzia zapewniały tego rodzaju wiarygodność, niezbędne jest zapewnienie odpowiednich warunków w czasie ich tworzenia.

Należy mieć świadomość, że proces tworzenia testu – od jego projektowania, przez pisanie zadań, ich redakcję, badania pilotażowe aż po ostateczną weryfikację i wdrożenie – jest czasochłonny. Poza sytuacjami wyjątkowymi jest to zadanie o harmonogramie obejmującym raczej wiele miesięcy i lat niż kilka tygodni. Ma też często charakter iteracyjny, czyli pierwotnie opracowane założenia są weryfikowane pod wpływem kolejnych informacji uzyskiwanych w pilotażach.

Kolejną sprawą o kluczowym znaczeniu jest kwestia właściwego doboru zespołu zajmującego się konstrukcją testu. Zespoły takie zawsze powinny łączyć dwie grupy kompetencji – te dotyczące specyficznego obszaru podlegającego testowaniu, i te związane ze znajomością (teoretyczną i praktyczną) problematyki diagnozy i statystyki. Oznacza to, że na wszystkich etapach tworzenia testu powinny być zaangażowane dwie grupy ekspertów: eksperci dziedzinowi czy branżowi i specjaliści psychometrycy. W szczególności udział psychometrów nie może ograniczać się do obliczenia wyników po pierwszym zastosowaniu „gotowego” testu. Jeżeli założenia leżące u podstaw konstrukcji testu są błędne lub też stosowane w nim zadania mają niewłaściwą formę, bardzo często jedyną możliwą rekomendacją z punktu widzenia eksperta będzie ponowne rozpoczęcie konstrukcji testu. Oszczędzanie na zasobach kompetencyjnych w istocie doprowadzi do niskiej wartości ostatecznie opracowywanych rozwiązań, co skutkuje obniżeniem jakości nadawania kwalifikacji. Oczywiście zarówno zapewnienie zasobów kadrowych, jak i odpowiedniej ilości czasu ma

też swoje wymierne konsekwencje finansowe, o czym trzeba pamiętać już na etapie budżetowania tego rodzaju projektów.

Ostatnie zagadnienie ma charakter odmienny od zaprezentowanych uwag organizacyjno-zarządczych. Wiąże się ono bezpośrednio z definiowaniem celów, jakie przyświecają w procesie konstrukcji narzędzi wykorzystywanych w walidacji.

Testy wiedzy mają za zadanie pomóc stwierdzić, czy kandydaci osiągnęli określone w opisie kwalifikacji efekty uczenia się. Oczywiście jest, że wiedza dotycząca jakiegoś tematu, a także umiejętność jej wykorzystywania, może mieć bardzo zróżnicowany charakter. Czymś innym jest pamiętać definicję, czymś innym – dogłębnie ją rozumieć. Jeszcze innego typu wiedzą musi dysponować osoba potrafiąca krytycznie przeanalizować jakiś problem czy konkretny przypadek. Teorią, którą bardzo często przywołuje się w kontekście weryfikacji wiedзовych efektów uczenia się, jest tzw. klasyfikacja celów nauczania (inaczej taksonomia Blooma), zaproponowana ponad pół wieku temu przez amerykańskiego psychologa edukacyjnego⁷ Benjaminą Blooma. Bloom, pisząc o celach, które staramy się osiągnąć w procesie edukacji, wyróżnił w sferze poznawczej sześć hierarchicznie ułożonych poziomów zrozumienia zagadnień.

Taksonomia celów nauczania wg Blooma

1. Wiedza to możliwość odtwarzania opanowanego pamięciowo materiału, a więc znajomość terminów, pojęć, faktów czy odpowiedzi na wcześniej znane pytania.

Przykład: Wymień objawy diagnostyczne depresji.

2. Zrozumienie oznacza możliwość porządkowania, tłumaczenia czy też dokonywania porównań oraz syntezy posiadanej wiedzy.

Przykład: Opisz podobieństwa i różnice w objawach nadczynności i niedoczynności tarczycy.

3. Zastosowanie to umiejętność odwołania się do posiadanej wiedzy (np. faktów, reguł, zasad) w celu rozwiązania problemu.

Przykład: Jakie metody obróbki termicznej należy zastosować, aby zachować jak najwięcej wartości odżywczych i smakowych mięsa, szykując potrawy dla osoby będącej na diecie lekkostrawnej?

⁷ Psychologia edukacji to specjalność naukowa psychologii, skupiająca się na rozwoju emocjonalnym i poznawczym człowieka, kształceniu, metodach nauczania, doradztwie oraz ocenie osiągnięć i postępów z wykorzystaniem testów standaryzowanych. Nie jest tożsama z psychologią wychowawczą, szkolną czy rozwojową.

4. Analiza to możliwość przeprowadzania rozumowania wymagającego wnioskowania na temat dowodów, przyczyn, a także konsekwencji określonych działań.

Przykład: Jakie mogą być pozytywne i negatywne konsekwencje częstej zmiany planów treningowych u zawodników uprawiających dyscypliny siłowe? Uzasadnij odpowiedź.

5. Synteza oznacza umiejętność integrowania szerokiego zakresu informacji w celu opracowania nowych sposobów organizacji wiedzy i proponowanie nowych rozwiązań.

Przykład: Na podstawie przedstawionego opisu zaproponuj modyfikację procesu produkcji, pozwalającą na zoptymalizowanie jakości przy zachowaniu zbliżonych kosztów produkcji.

6. Ewaluacja to umiejętność prezentacji i uzasadniania swoich pomysłów czy poglądów przy uwzględnieniu założonych celów, norm czy kryteriów. Ocena w zakresie kryteriów wewnętrznych.

Przykład: Przeanalizuj obecną politykę promocyjną wybranego polskiego miasta i oceń ją z punktu widzenia realizacji celów określonych w strategii rozwoju turystyki regionu.

Źródło: opracowanie własne na podstawie taksonomii Blooma (1956).

Choć klasyfikacja Blooma nie jest pozbawiona wad, stanowi bardzo pomocne narzędzie wspierające twórców testów wiedzy. Pomaga ona w podejmowaniu decyzji – zarówno tych ogólnych, dotyczących adekwatności poszczególnych metod weryfikacji efektów uczenia się, jak i bardziej szczegółowych, np. odnoszących się do optymalnej formy zadań testowych czy sposobu zadawania pytań nt. konkretnych zagadnień. Planując konstrukcję testu, należy analizować efekty uczenia się ujęte w opisie kwalifikacji z jej perspektywy, tak aby adekwatnie dobrać złożoność zadań do wymagań uwzględnionych w efektach uczenia się.

7. Rodzaje zadań/odpowiedzi stosowanych w testach wiedzy

Podstawowym i najbardziej ogólnym podziałem, który pozwala uporządkować zadania wykorzystywane w testach wiedzy, jest taksonomia wyróżniająca zadania zamknięte i otwarte. Zadania zamknięte to takie, w których poprawna odpowiedź jest w ścisły sposób zdefiniowana przez osobę tworzącą test. W zadaniach otwartych osoba testowana samodzielnie formułuje odpowiedzi. Dalsza część rozdziału została uporządkowana zgodnie z tą logiką. Najpierw omówiono poszczególne rodzaje zadań w formacie zamkniętym, potem – zadania otwarte.

7.1. Zadania zamknięte

Zadania w formacie zamkniętym składają się z części: głównej (nazywanej trzonem) oraz odpowiedzi, spośród których można wyróżnić prawidłową czy też najlepszą (zwaną czasem werstraktorem) i pozostałe – fałszywe czy też słabsze (zwane dystraktorami).

Podstawowe zalety, które wyróżniają zadania testowe przygotowane w tym formacie, to przede wszystkim łatwość ich sprawdzania, obiektywizm oceny oraz możliwość wykorzystania w ramach jednego testu dużej liczby zadań. Wszystkie te charakterystyki sprawiają, że testy przygotowane w tej formie – jeżeli tylko są skonstruowane poprawnie – z powodzeniem mogą spełniać kryteria dobroci psychometrycznej. Ze względu na zminimalizowanie interakcji zachodzących pomiędzy osobą testowaną a asesorem stosunkowo proste jest osiągnięcie bezstronności w czasie przeprowadzania testu. Pytania w formacie zamkniętym ograniczają ryzyko wystąpienia zniekształceń w ocenie poprawności odpowiedzi, co jest procesem wystandaryzowanym (podlegającym zresztą względnie prostej automatyzacji, pozwalającej tworzyć też tego rodzaju testy w wersjach skomputeryzowanych, bez utraty jakości pomiaru). Wreszcie możliwość wykorzystania dużej liczby zadań (przykładowo w godzinnym teście może być ich nawet 60) w istotny sposób zwiększa rzetelność (precyzję) pomiaru. Jest to korzystne z punktu widzenia trafności uzyskiwanych wyników – duża liczba pytań może się odwoływać do różnych obszarów wiedzy i kontekstów jej użycia, co zmniejsza ryzyko zarówno przeszacowania, jak i niedoszacowania poszczególnych aspektów kompetencji kandydata.

Warto zaznaczyć, że wbrew potocznym opiniom, w których pytania w formacie zamkniętym są kojarzone wyłącznie z mechanicznym, pamięciowym przyswajaniem materiału, testy wykorzystujące tego rodzaju zadania są w istocie bardzo elastycznym narzędziem weryfikacji poziomu wiedzy. Nie tylko umożliwiają sprawdzanie stopnia zapamiętania treści, ale również pozwalają skutecznie zweryfikować bardzo złożone konteksty jej użycia, np. wymagające zastosowania odległych analogii czy umiejętności krytycznej oceny zjawisk. W zależności od sposobu konstrukcji i doboru odpowiedzi pytania w formacie zamkniętym mogą odpowiadać właściwie wszystkim „pięciom” opisanej klasyfikacji celów nauczania Blooma.

Istotną funkcję w kształtowaniu pytań w formacie zamkniętym pełni dobór dystraktorów, czyli odpowiedzi błędnych. Dzięki ich właściwemu wykorzystaniu możliwe jest zarówno kształtowanie poziomu trudności, jak i złożoności pytań. Dystraktory mogą na przykład być wykorzystywane do sprawdzenia, jak bardzo skutecznie kandydat czy kandydatka potrafi rozróżniać pojęcia zbliżone – czy brzmieniowo, czy znaczeniowo. W przypadku zadań wymagających rozumowania odpowiedni dobór dystraktorów umożliwia sprawdzenie, czy osoba badana nie popełnia typowych błędów w rozumowaniu, interpretacji czy aplikacji weryfikowanej wiedzy. Jednocześnie warto podkreślić, że dobierając odpowiedzi błędne, nie należy ulegać pokusie formułowania dystraktorów łudząco podobnych do odpowiedzi poprawnych. Możliwość wyboru poprawnej odpowiedzi w teście nie powinna wynikać ze spostrzegawczości osoby testowanej, ale z opanowania przez nią wiedzy i umiejętności. Pytania układane „złośliwie”, celowo „ukrywające” odpowiedź poprawną wśród dystraktorów będą w teście przede wszystkim źródłem szumu i obniżą rzetelność uzyskiwanego pomiaru. Mogą też budzić (uzasadnione) wrażenie, że proces walidacji nie jest prowadzony w sposób sprawiedliwy i rzetelny, co na pewno nie jest zjawiskiem korzystnym dla wizerunku instytucji certyfikującej oraz jakości nadawanych kwalifikacji.

7.2. Zadania z wielokrotnym wyborem

Formatem najchętniej wykorzystywanym w testach wiedzy są bez wątpienia zadania wielokrotnego wyboru, czyli takie, gdy osoba testowana musi odnaleźć jedną poprawną odpowiedź spośród kilku dostępnych opcji. W praktyce najczęściej są stosowane formaty pytań, w których do wyboru są 4 lub 5 odpowiedzi. Wartość ta wydaje się uzasadniona, jako że z jednej strony zmniejsza problem zgadywania, z drugiej zaś nie powoduje zagubienia w nadmiarze odpowiedzi i nie wydłuża nadmiernie czasu wypełniania testu.

Pomimo licznych zalet i (być może zwodniczej) prostoty pytania z wielokrotnym wyborem są dość wymagające i czasochłonne w konstrukcji. Aby ułatwić nieco proces tworzenia takich zadań, opracowano szereg dobrych praktyk i zasad, o których warto pamiętać w czasie tworzenia tego rodzaju pytań:

Najważniejsze dobre praktyki i zasady związane z tworzeniem pytań z wielokrotnym wyborem

- Jeżeli to tylko możliwe, w trzonie zadania powinno się pojawiać kompletne pytanie, czy też instrukcja, a nie np. zdanie zakończone wielokropkiem czy wyrażenie wieloznaczne.
- Trzon pytania powinien jak najdokładniej określać problem, z którym ma się zmierzyć osoba testowana.

- Formułując pytania, należy unikać tych treści, które są zbędne i nie stanowią informacji koniecznych do rozwiązania zadania.
- Ze względu na czytelność tworzonych zadań i pytań niekorzystne jest układanie odpowiedzi, w których systematycznie pojawia się ten sam zwrot.
- Należy zadbać o taką formę językową, w której jak największa część pytania czy instrukcji znajduje się w trzonie zadania.
- Bardzo ostrożnie powinno się stosować przeczenia, zwłaszcza gdy mają się pojawiać w części głównej zadań. Ze względu na ryzyko błędnego odczytania warto też pamiętać o graficznym ich wyróżnieniu, np. kapitalikami lub przez podkreślenie, tak aby nie umknęły one uwadze testowanych osób.
- Niemal bezwarunkowo należy unikać podwójnych przeczeń, których interpretacja nastrocza odpowiadającym poważnych trudności, co utrudnia rzetelny pomiar poziomu testowanej wiedzy.
- Układając dystraktory, powinno się zadbać, aby zarówno ich forma, jak i treść upodabniały je do odpowiedzi poprawnych. Poszczególne opcje powinny mieć podobną długość. Należy też unikać sytuacji, w których, na przykład, odpowiedzi poprawne cechuje większa szczegółowość, złożoność albo też wyróżnia je obecność specjalistycznej terminologii.
- Jeżeli w poszczególnych odpowiedziach pojawiają się liczby, warto zadbać o ich precyzyjną rozłączność, np. zakresy liczbowe poszczególnych opcji nie powinny się nakładać.
- Jeżeli to tylko możliwe, poszczególne odpowiedzi należy układać w porządku logicznym, np. chronologicznym rosnącym, malejącym czy alfabetycznym.
- Bardzo istotne jest, aby forma poszczególnych odpowiedzi (np. zgodność gramatyczna formy czasownika pod względem rodzaju czy liczby) nie sugerowała, która z odpowiedzi jest poprawna.
- Wszystkie dostępne opcje, zarówno prawidłowa, jak i dystraktory, powinny tworzyć poprawną językowo całość z trzonem zadania.
- Zadania, w których pojawiają się odpowiedzi: „żadne z powyższych” oraz „wszystkie z powyższych”, należy stosować szczególnie ostrożnie. Wyjątkowo chętnie są one wykorzystywane jako dystraktory. W przypadku ich użycia w teście warto zadbać, aby to one czasami stanowiły odpowiedź poprawną. Pytania tego rodzaju nad wyraz często cechuje też niejednoznaczność, wskazane jest więc formułowanie ich z wyjątkową ostrożnością/uwagą.
- Przy układaniu testu z zadaniami z wielokrotnym wyborem warto zadbać o to, aby poprawne odpowiedzi rozkładały się pomiędzy dostępnymi do wyboru opcjami (np. A, B, C lub D) w możliwie równych proporcjach. Żadna z nich nie powinna dominować.

7.3. Zadania wymagające dopasowania odpowiedzi

Zadania z doborem odpowiedzi (ang. *matching*) czasem są traktowane jako jeden z typów pytań w wielokrotnym wyborze. Ich poprawna konstrukcja rządzi się jednak nieco odmiennymi zasadami, dlatego omówiono je niezależnie. W zadaniach tego typu osoba testowana ma połączyć ze sobą wyrażenia prezentowane w dwóch kolumnach zgodnie z podaną regułą. Reguła ta może odwoływać się do różnego typu relacji zachodzących pomiędzy prezentowanymi elementami. Na przykład można prosić o wskazanie poprawnych definicji dla przedstawionych pojęć, połączenie przyczyn i skutków, części przynależnych do określonych bardziej ogólnych zbiorów itd. Przykładowe zadanie tego typu przedstawiono w ramce.

Przykład zadania na dobieranie odpowiedzi

W lewej kolumnie opisano objawy zaburzeń psychicznych i neurologicznych. Proszę wybrać z prawej kolumny ich poprawne nazwy. Odpowiedzi proszę udzielać, wpisując literę odpowiadającą wybranej nazwie w oznaczone miejsce obok opisu danego objawu.

problemy z rozumieniem lub produkcją mowy	_____	A. agnozja
problemy z rozpoznawaniem obiektów	_____	B. apraksja
trudności z wykonywaniem czynności	_____	C. afazja
niemożność odczuwania przyjemności	_____	D. anhedonia

Kiedy w testach wiedzy warto stosować tego rodzaju zadania? Są one przydatne przede wszystkim w sytuacjach wymagających zweryfikowania szerokiego zakresu wiedzy faktograficznej przy wykorzystaniu pojedynczego zadania. Najistotniejszą ich wadą – kluczową z perspektywy optymalizacji procesu walidacji – jest stosunkowo duża czasochłonność w czasie rozwiązywania. Takie zadania trudno jest też przekształcić tak, aby weryfikowały coś więcej niż tylko zapamiętywanie informacji. W sytuacji, w której potrzebna jest weryfikacja rozbudowanej wiedzy dotyczącej rozumienia definicji, przepisów czy procedur, stosowanie tego rodzaju zadań może być uzasadnione.

Najważniejsze dobre praktyki związane z konstrukcją zadań na dobieranie odpowiedzi

- Instrukcja zadania powinna w jednoznaczny sposób określać regułę, którą rozwiązujący powinni stosować w dopasowaniu odpowiedzi. Zbyt ogólna instrukcja („Połącz ze sobą właściwe elementy”) może być źródłem licznych nieporozumień. Znacznie lepiej jest wskazać wprost, jaka relacja powinna łączyć bodźce („Wskaż właściwy paragraf dla każdego z przestępstw opisanych w kolumnie po lewej stronie”).

- Reguła łącząca elementy powinna być stała w obrębie każdego zadania – łączenie w ramach jednego zadania różnych kategorii bodźców wprowadza chaos, a jednocześnie osłabia trafność pomiaru (zazwyczaj część odpowiedzi zupełnie do siebie nie pasuje, co pozwala je wykluczyć nawet przez osoby nieposiadające żadnej wiedzy merytorycznej).
- Należy unikać takiej konstrukcji bodźców, która umożliwia znalezienie poprawnych odpowiedzi opartych na wskazówkach gramatycznych czy językowych. Podstawą trafnego dopasowania bodźców powinna być wiedza osoby egzaminowanej, a nie zgodność rodzaju czy liczby.
- W zadaniach, w których jedna z części bodźców jest bardziej rozbudowana, np. definicje czy przykłady, należy umieszczać je w kolumnie po lewej stronie. Jeżeli to możliwe, prawa strona powinna zawierać krótszą część odpowiedzi.
- Podobnie jak w przypadku zadań typu „prawda – fałsz” korzystne jest unikanie dosłownych cytatów ze źródeł często wykorzystywanych w procesie kształcenia – nawet w przypadku testowania wiedzy faktograficznej lepiej badać stopień jej zapamiętania, a nie jedynie umiejętność rozpoznania określonych wyrażeń.

7.5. Zadania typu „prawda – fałsz”

Pytania typu „prawda – fałsz” pod wieloma względami są najprostszym rodzajem zadań, które można wykorzystywać w testach wiedzy. W zadaniach tego typu prezentuje się listę stwierdzeń i w przypadku każdego z nich osoba egzaminowana musi ocenić, czy prezentowane zdanie jest prawdziwe, czy nie. Takie pytania doskonale sprawdzają się nie tylko podczas weryfikacji wiedzy, ale również w problemach decyzyjnych, w których opisuje się jakąś sytuację i czyjąś reakcję na nią, a następnie zadawane jest pytanie, np. czy jest to reakcja poprawna, zgodna z przepisami, oczekiwana itp. Udzielanie odpowiedzi w tego typu zadaniach jest stosunkowo szybkie. Jeżeli nie są one bardzo złożone, można założyć, że egzaminowani w ciągu minuty będą mogli rozwiązać nawet kilka tego rodzaju problemów. Sprawia to, że przy wykorzystaniu zadań „prawda – fałsz” w stosunkowo krótkim czasie można zweryfikować względnie szeroki zakres wiedzy respondentów.

Niestety, prostota takich zadań ma też negatywne konsekwencje. Przede wszystkim odpowiadający mają tu aż pięćdziesiąt procent szans na przypadkowe udzielenie poprawnej odpowiedzi, nawet jeżeli nie mają rzeczywistej wiedzy. Oznacza to, że rzetelne oszacowanie poziomu kompetencji osób egzaminowanych wymaga przygotowania bardzo dużej liczby pytań – często osiągnięcie satysfakcjonującej precyzji pomiaru wymaga minimum 70–80 pytań. Istotnym problemem może być też w tym przypadku skonstruowanie zadań, w których będzie weryfikowane nie tylko zapamiętanie wiedzy faktograficznej czy reguł, lecz także jej zrozumienie czy trafne zastosowanie.

Jeżeli w teście zastosuje się pytania w formacie prawda – fałsz, warto pamiętać o przestrzeganiu następujących zasad:

Najważniejsze zasady dotyczące tworzenia pytań w formacie „prawda – fałsz”

- Liczba zadań powinna być na tyle duża, by możliwe było skuteczne odróżnienie osób faktycznie posiadających oczekiwaną wiedzę od odpowiadających poprawnie przez przypadek.
- Prezentowane stwierdzenia powinny być prawdziwe lub nie w sposób niebudzący wątpliwości, bez konieczności uzupełniania ich przez osoby wypełniające test o dodatkowe zastrzeżenia czy warunki.
- Z tych samych powodów w prezentowanych stwierdzeniach należy unikać wielkich kwantyfikatorów („zawsze”, „nigdy” etc.). Jako że reguły czy prawa pozbawione jakichkolwiek wyjątków występują sporadycznie, odpowiadający traktują obecność kwantyfikatorów jako istotną wskazówkę, że stwierdzenie jest fałszywe. Takie zadania często stają się przyczyną interpretacyjnych sporów (np. „Czy rzeczywiście reguła ta nie posiada absolutnie żadnych wyjątków?”).
- Stosowane słownictwo i konstrukcja zdań powinny być możliwie jak najprostsze. W przeciwnym wypadku poprawność odpowiedzi zaczyna bardziej zależeć od poziomu kompetencji językowych i inteligencji kandydatów niż od poziomu ich wiedzy.
- Poszczególne pytania powinny koncentrować się na pomiarze zrozumienia jednego stwierdzenia, prawa itd. Łączenie dwóch wątków w ramach jednego pytania prowadzi zwykle do niepotrzebnych komplikacji, które obniżają jakość pomiaru.
- Należy unikać stosowania w zadaniach stwierdzeń, które są dosłownym cytatem z przepisów, podręczników czy innych dokumentów wykorzystywanych jako materiały dydaktyczne. Nawet w przypadku weryfikacji wiedzy faktograficznej czy terminologicznej udzielenie poprawnej odpowiedzi nie powinno opierać się na prostym rozpoznaniu znajomego tekstu.
- W miarę możliwości należy przygotowywać zadania, które będą weryfikować nie tylko samo zapamiętanie, ale również bardziej zaawansowane poziomy przetwarzania treści, takie jak zrozumienie i umiejętność zastosowania wiedzy. Wiedza w kontekście zawodowym pełni zazwyczaj funkcję służebną wobec podejmowanych decyzji i wykonywanych czynności. Tak więc w przypadku zdecydowanej większości kwalifikacji bardziej niż umiejętność odtworzenia zapamiętanych faktów, pojęć czy reguł w obszarze zainteresowania pozostają kompetencje w zakresie jej wykorzystania.
- Podobnie jak w przypadku zadań z wielokrotnym wyborem niekorzystne jest stosowanie przeczeń w przedstawionych stwierdzeniach. Weryfikacja stwierdzeń z podwójnym przeczeniem niepotrzebnie nastrocza kandydatowi dodatkowych trudności, co prowadzi do obniżenia rzetelności pomiaru.

7.6. Zadania wymagające wypełnienia luk w tekście

Podobnie jak omówione zadania z dopasowaniem opcji, także te z lukami doskonale nadają się do weryfikowania stopnia zapamiętania wiedzy faktualnej. W tego typu zadaniach osoba odpowiadająca ma uzupełnić braki w prezentowanym tekście przez wpisanie odpowiedniego słowa lub właściwej frazy. W zadaniu może być wymagane podanie jednej lub większej liczby odpowiedzi (taki właśnie charakter ma przykładowe zadanie w ramce).

Przykład zadania z luką

Zgodnie z ustawą z dnia 1 marca 2018 r. o przeciwdziałaniu praniu pieniędzy oraz finansowaniu terroryzmu instytucjami obowiązany-
mi są przedsiębiorcy prowadzący działalność
w rozumieniu ustawy, inni przedsiębior-
cy świadczący usługę lub usługę pośrednictwa
w, niebędący innymi instytucjami obowiąza-
nymi, oraz oddziały przedsiębiorców zagranicznych prowadzących taką
działalność na terytorium Rzeczypospolitej Polskiej.

Niewątpliwą przewagą takich zadań, wobec wszystkich omawianych wcześniej ich typów, jest to, że od osoby odpowiadającej wymagają samodzielnego wygenerowania poprawnej odpowiedzi. Wykonanie zadania nie jest wspierane przez możliwość zgadywania czy rozpoznania poprawnej odpowiedzi, a zatem uzyskane wyniki stanowią bardziej trafną miarę rzeczywistego poziomu wiedzy. Choć jest to ogromną zaletą, to – jednocześnie – większa dowolność odpowiedzi dostarczanych przez osoby egzaminowane jest powodem licznych kłopotów. Przede wszystkim proces sprawdzania odpowiedzi udzielanych przez kandydatów jest znacznie bardziej czasochłonny i narażony na interpretacyjne wątpliwości. W przeciwieństwie do pozostałych typów zadań weryfikacja poprawności wykonanego zadania z luką nie poddaje się łatwo procesowi automatyzacji, co może stanowić przeszkodę w wykorzystywaniu ich w przypadku procedur walidacji, zakładających zastosowanie systemów komputerowych w procesie testowania. Wyjątkiem są tu takie testy, w których wybór poprawnej odpowiedzi nie budzi żadnych wątpliwości (np. podanie konkretnego numeru paragrafu, nazwy miejscowości czy formy czasownika pasującego do kontekstu zdania).

Jako że zadania wymagające wypełnienia luk cechuje bardziej otwarta konstrukcja, w przypadku planowanego ich wykorzystania wyjątkowego znaczenia nabierają badania pilotażowe poprzedzające utworzenie ostatecznej wersji zadań testowych. W pilotażu powinny uczestniczyć obie strony biorące udział w testowaniu – zarówno przedstawiciele grup, którzy w przyszłości będą się ubiegać o uzyskanie kwalifikacji, jak i osoby, które będą odpowiedzialne za sprawdzanie wyników testów. W przypadku osób podchodzących do walidacji istotne jest sprawdzenie, czy konstrukcja zadań nie skłania badanych do udzielania odpowiedzi sensownych i pasujących do kontekstu, ale nieprzewidzianych przez twórców testu. Takie sytuacje wskazują na konieczność korekt w sposobie

konstrukcji zadań. W przypadku osób oceniających odpowiedzi należy sprawdzić, czy klucz odpowiedzi obejmuje wszystkie pojawiające się typy reakcji i czy sposób oceniania we właściwy sposób odzwierciedla jakość wiedzy wyrażonej w odpowiedziach.

Zadania „z luką” mogą przyjmować bardzo różne formy. Poniższe wskazówki pozwolą jednak uniknąć kilku najczęściej występujących problemów:

- Luki wymagające uzupełnienia powinny dotyczyć jedynie słów i wyrażen mających istotne znaczenie merytoryczne. Konieczność uzupełniania luk obejmujących trywialne elementy zdania (np. zaimki osobowe) może być uzasadniona wyłącznie w testach językowych lub sprawdzających znajomość tekstów wymagających dosłownego zapamiętania.
- Należy też unikać usuwania z tekstu tak dużej liczby fragmentów, że jego znaczenie staje się nieczytelne. Taka praktyka prowadzi nieuchronnie do wieloznaczności zadania, kłopotów z oceną odpowiedzi oraz obniżenia rzetelności i trafności pomiaru.
- Nie należy konstruować zadań w sposób, który sugeruje poprawną odpowiedź, dostarczając istotnych wskazówek gramatycznych (np. co do rodzaju rzeczownika, który ma pojawić się w zdaniu).
- Jeżeli to tylko możliwe, część wymagająca uzupełnienia powinna znajdować się na końcu, a nie w początkowej części zdania.
- Należy upewnić się, że tworzone zadania mają faktycznie tylko jedną poprawną odpowiedź. W procesie weryfikacji istotne znaczenie mają tu zarówno eksperci dziedzinowi, jak i wiedza uzyskana w badaniach pilotażowych.

7.7. Pytania i zadania z otwartym formatem odpowiedzi

Pytania z otwartym formatem odpowiedzi (które dla wygody można nazwać pytaniami otwartymi) stanowią rodzaj zadań, w których osoba egzaminowana ma największą swobodę w udzielaniu odpowiedzi i prezentowaniu posiadanej wiedzy. Sposób reagowania jest tu zdeterminowany właściwie jedynie instrukcją oraz, jeżeli mają zastosowanie, ograniczeniami czasowymi lub technicznymi (np. maksymalna objętość tekstu).

Do tej kategorii zalicza się zarówno pytania, w których zawarta jest prośba o udzielenie odpowiedzi krótkich i zwięzłych (np. „Uzasadnij w punktach główne korzyści i ryzyka zdrowotne związane z zastosowaniem diety wegetariańskiej”), aż po takie, gdy jest wymagane napisanie dłuższego tekstu (np. eseju) lub skonstruowanie złożonej wypowiedzi ustnej (np. „Przygotuj i przedstaw w atrakcyjny sposób historię powstania styczniowego, korzystając z udostępnionych materiałów wizualnych. Twoimi słuchaczami są uczniowie III klasy liceum. Na przygotowanie wypowiedzi masz pół godziny. Twoja prezentacja powinna trwać nie mniej niż 5 i nie więcej niż 7 minut”).

Można uznać, że dowolność formy odpowiedzi – wpisana z definicji w ten rodzaj zadań – z perspektywy procesu walidacji stanowi też największą ich siłę i słabość. Do najistotniejszych zalet zadań otwartych należy to, że nadają się one doskonale do weryfikacji bardziej złożonych form zrozumienia i zastosowania wiedzy.

W dłuższej wypowiedzi pisemnej czy ustnej istnieje szansa na sprawdzenie poziomu zrozumienia tekstu, zdolności analizy i krytycznej oceny, umiejętności oceny złożonego kontekstu sytuacyjnego i jego uwzględnienia w proponowanych rozwiązaniach. Tu także weryfikowana może być wiedza dotycząca określonych konwencji językowych czy środków retorycznych, których zastosowania wymaga dłuższa wypowiedź.

W pytaniach tego typu bardzo utrudnione jest też zgadywanie poprawnych odpowiedzi – treść polecenia często nie dostarcza żadnych wskazówek w tym zakresie. Dłuższe wypowiedzi – ustne czy pisemne – nierzadko mogą służyć do jednoczesnej weryfikacji kilku różnych efektów uczenia się. Przykładowo poza dostarczeniem informacji o poziomie wiedzy mogą oceniać stopień opanowania szeregu umiejętności (np. komunikacyjnych, związanych z umiejętnością prezentacji, poprawnym redagowaniem tekstu itd.). [Dokładniejsze omówienie tego aspektu wykracza jednak poza zakres niniejszego opracowania. Osoby zainteresowane oceną umiejętności zachęcamy do lektury publikacji zawierającej krótki przegląd najpopularniejszych metod stosowanych w procesach walidacji (Gmaj i in., 2016) lub też podręczników zorientowanych na problematykę pomiaru dydaktycznego lub psychometrii – zwłaszcza ich fragmenty dotyczące pomiaru jakości wykonania zadań (np. Anastasi i Urbina, 1999)].

Niestety, poza licznymi zaletami, pytania w formacie otwartym mają szereg poważnych ograniczeń. Po pierwsze, ich stosowanie często jest bardzo czasochłonne. Weryfikacja wiedzy, umiejętności i kompetencji społecznych wymaga oceny bardzo wielu aspektów funkcjonowania. Nawet w obrębie samych efektów uczenia się związanych z wiedzą kandydatki czy kandydata nierzadko będzie wymagać kilka odrębnych obszarów. W tej sytuacji poświęcenie np. 45 min na przygotowanie eseju może być nieracjonalne. Problemem tej formy weryfikacji jest też jej wyrywkowość. Nie sposób, korzystając z dwóch czy trzech pytań (a często tyle udaje się zadać np. w czasie egzaminu ustnego), zweryfikować stopień opanowania bardzo szerokiego zakresu treści. Jako że priorytetem w procesie walidacji jest ocena, czy wiedza kandydatów jest wystarczająca i kompletna, taki sposób sprawdzenia trudno uznać za optymalny. Priorytetem w czasie testów realizowanych w tej formie powinna być weryfikacja tych aspektów kompetencji kandydata, które nie mogą być skutecznie sprawdzone przy wykorzystaniu innych typów zadań (np. umiejętności formułowania bardziej złożonych wypowiedzi ustnych czy pisemnych). Zadania otwarte (często realizowane w formie ustnej) odgrywają wtedy cenną rolę uzupełniającą w stosunku do pozostałych metod weryfikacji. Problematyczne może być też uzyskanie w tego rodzaju zadaniach satysfakcjonującego poziomu obiektywizmu i rzetelności, nawet w przypadku starannie opracowanych kryteriów oceny. Zarówno w przypadku testów przeprowadzanych w formie ustnej, jak i podczas oceny wypowiedzi pisemnych rekomenduje się skorzystanie z pomocy co najmniej dwóch asesorów, co pozwala poprawić rzetelność uzyskiwanych wyników, ale też znacząco zwiększa koszty i czasochłonność procedur.

Pytania w formacie otwartym są pozornie proste w przygotowaniu – instrukcja często ogranicza się tu do wskazania obszaru tematycznego czy zagadnienia, na

które ma się wypowiedzieć osoba egzaminowana. Jednocześnie jednak dobre przygotowanie kryteriów oceny oraz zweryfikowanie ich działania w czasie pilotażowego wdrożenia procedury znacznie niweluje tę złudną oszczędność czasu.

Należy też pamiętać, że w przypadku dłuższych wypowiedzi istotne znaczenie dla ich jakości mają czynniki związane z kompetencjami językowymi, kapitałem kulturowym, a także – w przypadku testów ustnych – zmienne osobowościowe osób piszących (osoby ekstrawertyczne czy pewne siebie będą wypadały przed asesorami lepiej). W takim przypadku tworzone procedury stają się obciążone błędem systematycznym – ocena wypowiedzi ustnej lub pisemnej, np. eseju, nie jest już wyłącznie funkcją posiadanej przez kogoś wiedzy, lecz także innych czynników, często niemających żadnego znaczenia dla uzyskiwanej kwalifikacji.

Podsumowując przedstawione uwagi, w trosce o jakość pomiaru i ze względów ekonomiczno-logistycznych powinno się unikać wykorzystywania pytań otwartych, jeżeli nie są one niezbędne w procesie walidacji. Kiedy pytania otwarte wydają się rozwiązaniem odpowiednim? Głównie wtedy, gdy opis danej kwalifikacji rynkowej stawia przed kandydatami wymagania związane z przeprowadzaniem złożonych, krytycznych rozumowań, umiejętności rozważania i uzasadnienia zjawisk czy decyzji pod wieloma względami, np. trener klasy mistrzowskiej w piłce siatkowej. Jednocześnie jednak należy pamiętać, że taki sposób aplikacji wiedzy, silnie osadzony w kontekście funkcjonowania zawodowego i zbliżony do realiów pracy zawodowej, np. dobór optymalnych oddziaływań fizjoterapeutycznych uwzględniający wiek i choroby współwystępujące u danego pacjenta, często może być skuteczniej zweryfikowany innymi metodami nietestowymi, ale np. przez wykorzystanie takich metod jak studia przypadku czy próbki pracy.

Rekomendacje dotyczące zastosowania w testach pytań w formie otwartej

- Pomimo otwartego charakteru zadanie powinno być klarownie zdefiniowane. W instrukcji warto odwoływać się do takich wyrażen, jak: „Porównaj...”, „Omów argumenty za i przeciw...”, „Przedstaw przyczyny...” itd. Czasowniki te powinny odpowiadać poziomowi rozumienia zagadnień określonego w efektach uczenia się.
- Sposób formułowania instrukcji powinien też korespondować z przygotowanym kluczem odpowiedzi. Przykładowo warto zastanowić się, czy osoba ma wskazać „optymalne oddziaływanie i uzasadnić je, odwołując się do badań”, czy też ma „zapropionować minimum trzy różne oddziaływania skuteczne w leczeniu”. Zarówno długość wypowiedzi, jak i liczba przykładów, argumentów czy uzasadnień powinny wynikać wprost z instrukcji.
- W przypadku testów realizowanych w formie ustnej istotne jest, aby określić dokładnie zakres dodatkowych wyjaśnień czy pytań pomocniczych. Dopuszczalne zachowania asesorów powinny być ściśle zdefiniowane, w przeciwnym wypadku nie ma szansy na zapewnienie obiektywizmu i wystandaryzowanie procesu oceny. Wiąże się to ściśle z koniecznością przygotowania instrukcji i starannego szkolenia asesorów korzystających z tego rodzaju narzędzi pomiarowych.

Przykład instrukcji dla asesora do testu w formie ustnej

Polecenie dla kandydata:

Proszę wymienić wszystkie znane Panu/i objawy depresji opisane w kryteriach diagnostycznych klasyfikacji ICD-10.

Instrukcja dla asesora:

W przypadku wymienienia przez kandydata niekompletnej listy asesor powinien dopytać „Czy to już wszystkie znane Pani/u objawy wymienione w ICD?”.

Uwaga: Asesorowi nie wolno udzielać jakichkolwiek podpowiedzi czy wskazówek.

8. Sposób przeprowadzania testów

Wziąwszy pod uwagę sposób przeprowadzenia (administrowania), testy można podzielić na:

- indywidualne lub panelowe (grupowe),
- realizowane w formie papierowej lub elektronicznej.

Odnosząc się do pierwszego rozróżnienia, można powiedzieć, że każdy z opisanych typów sprawdzania wiedzy może być realizowany w formie indywidualnej, ale nie każdy w modelu panelowym. W zasadzie tylko testy pisemne mogą mieć formę egzaminowania grupowego. W przypadku testu wykonywanego grupowo przede wszystkim należy zadbać o to, aby poszczególni uczestnicy sobie nie przeszkadzali i fakt testowania grupowego nie generował dodatkowych źródeł błędów pomiaru. Przy pomiarze grupowym istnieje także ryzyko, że osoby egzaminowane mogą spisywać odpowiedzi od innych osób. Powinno się bezwzględnie zadbać o to, aby takie sytuacje nie wystąpiły. Podczas pomiaru grupowego należy także upewnić się, że wszystkie osoby przystępujące do walidacji dobrze zrozumiały instrukcję i ich szanse są wyrównane.

W zasadzie wszystkie testy mogą być przeprowadzane zarówno w formie tradycyjnej, wymagającej zaznaczania odpowiedzi ołówkiem lub długopisem (tzw. testy papier-ołówek) i w obecności komisji walidacyjnej, jak i elektronicznej. Ta ostatnia jest najtrudniejsza w przypadku konieczności zapisywania odpowiedzi przez osoby mające małe doświadczenie z używaniem klawiatury. Forma elektroniczna testu znacząco jednak ułatwia (i obiektywizuje) proces oceniania testu.

Na koniec warto wspomnieć, że test powinien mieć określoną szatę graficzną. Jest to element standaryzacji testu, to znaczy ujednolicenia warunków badania. Pytania, polecenia, zadania powinny być formułowane nie tylko tak samo w zakresie treści, lecz także pod względem formy. Z reguły test wiedzy zaczyna się od miejsca, w którym osoba wpisuje swoje dane (lub kod badania), następnie zapoznaje się z instrukcją, dalej – nie jest to obligatoryjne – natrafia na trening lub przykładowe pytania i odpowiedzi, aby ostatecznie dojść do pytań właściwych. Należy pamiętać o tym, że format graficzny testu nie może utrudniać udzielania odpowiedzi. Sposób odpowiadania powinien być jednoznacznie określony, a miejsce do wpisywania odpowiedzi dostatecznie obszerne, by każdy miał szansę na udzielenie odpowiedzi. Wydruk testu musi być dobrej jakości, w formacie A4 (jeśli test składa się z kilku kartek, powinny być zszyte lub w formie książeczki, a strony numerowane). Zaleca się, aby czcionka była typowa, o prostym i czytelnym kroju, czarna, rysunki zaś – jeśli nie ma wymogu dotyczącego koloru z uwagi na specyfikę materiału – powinny być czarno-białe, wyraźne, numerowane. Zalecane jest wertykalne formatowanie kafeterii odpowiedzi (odpowiedzi do wyboru w kolejnych wierszach, a nie w tej samej linii).

8.1. Stosowanie testów u osób o specjalnych potrzebach

W zaprezentowanych rozdziałach wiele uwagi poświęcono standaryzacji, czyli m.in. ujednoliceniu materiału testowego i warunków testowania. Dzięki temu, jak wspomniano, zapewnia się sprawiedliwość traktowania, a także przekonanie, że odmienne wyniki odzwierciedlają różny stan wiedzy, a nie różne traktowanie podczas egzaminu. Nie zawsze jednak takie same warunki muszą zapewniać sprawiedliwe traktowanie. W oczywisty sposób dzieje się tak wtedy, gdy do testu przystępują osoby z jakimś rodzajem niepełnosprawności, który rzutuje na możliwość prowadzenia weryfikacji metodą testową w określonej formie. Na przykład osoby z niepełnosprawnościami sensorycznymi mogą spędzić więcej czasu na samym czytaniu poleceń, szczególnie gdy zostały one wydrukowane małą czcionką w przypadku osób słabowidzących i trudnym językiem (zdania wielokrotnie złożone, niejednoznaczne sformułowania) w przypadku głuchych. Z punktu widzenia możliwości prowadzenia (sprawiedliwego) testów w tych grupach możliwe są dwa rozwiązania tego typu problemów:

1. opracowanie jak najbardziej uniwersalnych i jak najmniej wykluczających zasad weryfikacji,
2. opracowanie specjalnych modyfikacji w procesie testowania, które pozwalają na przystosowanie weryfikacji do specjalnych potrzeb kandydatów.

Pierwszy z zalecanych sposobów można uznać za zalecany punkt wyjścia konstruowania różnych procedur walidacyjnych. Przede wszystkim należy zadbać, aby dany test nie dyskryminował lub nie faworyzował jakichś grup osób. Wątek ten był już podnoszony przy opisie kwestii związanych z obiektywnością – idealny test wiedzy pozwala na jej ocenę, a nie innych cech kandydata, np. takich jak stopień biegłości języka polskiego w przypadku głuchych. Typowe zabiegi stosowane po to, aby nie dyskryminować osób o słabszych umiejętnościach językowych, niewidomych i słabowidzących lub głuchych czy słabosłyszących, a także osób starszych lub o gorszym ogólnie stanie zdrowia to:

- jednoznaczne i proste językowo, powszechnie zrozumiałe instrukcje czy polecenie;
- możliwość uzyskania powtórzenia instrukcji lub polecenia;
- wyraźny format graficzny poleceń, w tym odpowiednio duża czcionka o łatwym do czytania kroju, dobra jakość wydruku;
- dwutorowość przekazu – polecenia można przeczytać, ale też ich wysłuchać, można też zapewnić tłumaczenie na polski język migowy (PJM);
- rezygnacja z ograniczeń czasowych w udzielaniu odpowiedzi;
- możliwość korzystania z takich pomocy, jak np. słowniki czy kalkulatory.

Nie zawsze te zabiegi będą wystarczające. Nawet bardzo duża czcionka dla osoby niewidomej nie będzie wystarczająca i konieczne będzie przeprowadzenie egzaminu w formie ustnej lub zastosowanie alfabetu Braille'a, lub wykorzystanie oprogramowania zamieniającego tekst na mowę. Tego rodzaju zmiany już są nazywane modyfikacją testu. Zmieniają one na tyle formę testu, że wyniki osób, u których test był przeprowadzony w dwóch różnych warunkach, stają się mało porównywalne. Niemniej z punktu widzenia weryfikacji efektów uczenia się obie formy egzaminu mogą być równoważne i równie dobre. Modyfikacje te mogą polegać na:

- dostosowaniu formy zadawania pytań i formułowania poleceń, np. użycia alfabetu Braille'a lub tłumaczenia na PJM odczytania przez osobę prowadzącą test;
- dostosowaniu formy udzielania odpowiedzi, np. wypowiadanie odpowiedzi zamiast jej zapisywania;
- pominięciu pewnych części egzaminu (choć w tym wypadku konieczne jest równoczesne zaproponowanie innych form egzaminowania – niezależnie od rodzaju specjalnych potrzeb opisane w kwalifikacji efekty uczenia się muszą zostać należycie zweryfikowane);
- możliwości skorzystania z dodatkowej pomocy, np. kalkulatora (pomocę tę nie mogą jednak czynić egzaminu łatwiejszym, lecz tylko lepiej przystosowanym dla niektórych grup osób);
- zmodyfikowaniu procedury prowadzenia egzaminu (np. wydłużenie czasu jego trwania).

Tworzenie modyfikacji testu znacząco komplikuje proces opracowania materiałów egzaminacyjnych. Modyfikacje w procedurze wymagają dobrego przemyślenia i z reguły przetestowania tego, czy – z jednej strony – spełniają swoją funkcję, a z drugiej – czy nie zniekształcają wyników egzaminu. Dlatego w pierwszej kolejności warto starać się opracować taki test, który będzie jak najbardziej uniwersalny.

8.2. Procedura administrowania testu wiedzy. Rola użytkownika testów

Procedura administrowania to zarządzanie „życiem” testu. Proces ten rozpoczyna się od jego opracowania, później stosowania, dalej następuje nanoszenie w nim zmian i poprawek. Jednak najistotniejszy etap administrowania to sposób jego używania, czyli prowadzenie walidacji za pomocą tego narzędzia. Na każdy proces używania testu w praktyce powinny składać się takie etapy, jak:

- Podjęcie decyzji o formie używanego testu. Taka decyzja oczywiście może być podjęta tylko wtedy, gdy istnieją alternatywne wersje testu (np. komputerowa, typu papier-ołówek). Jeśli dysponuje się pulą zadań i pytań, na podstawie której układany jest test, pierwszym etapem będzie losowanie zbioru tych zadań. Wyjątkowo w przypadku egzaminów ustnych losowanie zadań może być dokonane

przez kandydata w trakcie weryfikacji. Zaleca się, aby dobór zadań z dostępnej puli miał charakter losowy.

- Wyjaśnienie kandydatom celu, sposobu przebiegu i warunków weryfikacji (podanie instrukcji). W szczególności należy podać takie informacje, jak czas trwania testu i sposób udzielania odpowiedzi. Instrukcja ta powinna być dobrze opisana i taka sama dla wszystkich osób – to jeden z elementów zapewnienia standardowych warunków badania. Te informacje powinny być znane asesorum i innym osobom włączonym w przeprowadzanie walidacji.
- Przeprowadzenie weryfikacji – testowanie. To proces zadawania pytań lub zadań ustnie albo pisemnie oraz zbierania odpowiedzi od osób podchodzących do walidacji.
- Kodowanie wyników testu – ocena odpowiedzi.
- Interpretacja wyników. Ocena, czy zostały potwierdzone efekty uczenia się dla danej kwalifikacji rynkowej.
- Poinformowanie o wyniku testu.

W ramce wskazano informacje, które powinna zawierać instrukcja do testu dla kandydatów.

Instrukcja do testu dla kandydata

Instrukcja do testu musi zawierać wszystkie istotne informacje, z którymi powinien zapoznać się kandydat przed rozpoczęciem testowania. W szczególności osoba testowana powinna wiedzieć:

- z jakiego rodzaju zadaniami i pytaniami będzie konfrontowana;
- z ilu pytań lub zadań będzie składał się test;
- ile czasu będzie miała na udzielenie odpowiedzi lub rozwiązanie zadań;
- w jaki sposób może udzielać odpowiedzi;
- czy może, a jeśli tak, to w jaki sposób, zmieniać odpowiedzi w zadaniach, które już zakończyła, i pytaniach, na które udzieliła odpowiedzi;
- na co ewentualnie powinna zwrócić uwagę, odpowiadając;
- jakich zachowań powinna unikać lub też jakie są zabronione (np. przy testowaniu grupowym zabrania się zaglądania do odpowiedzi innych osób).

Na koniec instrukcji, która może być zarówno odczytana osobom podchodzącym do testu, jak i przekazana na piśmie, warto się upewnić, że wszyscy kandydaci zrozumieли instrukcję, i zachęcić do zadawania pytań w przypadku wątpliwości. Od momentu zakończenia czytania instrukcji z reguły zaczyna się główna część testowania, w trakcie której rozdawane i udostępnianie są pytania oraz zadania do wykonania. Jeśli wprowadzono takie ograniczenie, odmierzany jest czas.

Bardzo ważne, aby wszyscy kandydaci otrzymali ten sam zestaw informacji i mieli równe szanse na zapoznanie się z nimi i ich przyswojenie. Jest to spełnienie postulatu o standaryzacji warunków testowania.

W przypadku zadań, w których osoba testowana ma wykonać jakąś czynność, można zaplanować wcześniejszy trening dla tego działania. Jest to szczególnie zasadne wtedy, gdy zadanie jest całkowicie lub częściowo nowe dla osób egzaminowanych. Trening ten powinien być krótki i sprowadzać się do próby przyswojenia podstawowych elementów danego zadania. Celem treningu nie jest uzyskanie biegłości (wyuczenie) wykonywania określonego zadania, bo z reguły to jest przedmiotem samego testowania. Przykładowo, jeśli test ma formę komputerową, pierwsze pytania (lub jedno) mogą służyć nauczaniu osoby testowanej, w jaki sposób ma udzielać odpowiedzi (np. których klawiszy powinna używać).

Podobną rolę jak trening odgrywa włączenie do instrukcji przykładowych zadań lub pytań ze wskazaniem sposobu odpowiadania. W odróżnieniu od treningu przykład nie pozwala na aktywne ćwiczenie – dostarcza tylko informacji, jak powinno wyglądać zachowanie kandydata w trakcie testu, czego się po niej/nim oczekuje.

W przypadku testów komputerowych przykład może być zaprezentowany w formie animacji, a w przypadku testów typu papier i ołówki jako opis lub ilustracja. Ważne, aby wszyscy kandydaci mieli takie same szanse na ewentualny trening lub zapoznanie się z przykładami.

Instrukcja dla użytkowników testu, tj. osób przeprowadzających test lub asesorów, powinna przewidywać także ich reakcje w trakcie trwania testu. Przede wszystkim należy tu uwzględnić sposoby udzielania odpowiedzi na pytania osób testowanych, w tym przede wszystkim dotyczących materiału testowego. Należy rozstrzygnąć, czy asesor może objaśniać zadania lub pytania w trakcie testu. Jeśli tak, należy zdefiniować zakres przekazywanych informacji (podpowiedzi). Wszystkie osoby testowane powinny być potraktowane w ten sam sposób, czyli mieć szansę otrzymać takie same podpowiedzi lub – jeśli tak zdecydowano w procedurze przeprowadzania testu – żadna z osób nie może otrzymać żadnych podpowiedzi.

Planując procedurę badania testem, warto zwrócić jeszcze uwagę na kwestię informowania osób biorących udział w badaniu, jaki status ma zgadywanie odpowiedzi. Ma to oczywiście zastosowanie głównie w przypadku pytań z zamkniętym formatem odpowiedzi.

Zgadywanie jest specyficznym czynnikiem zniekształcającym wyniki testów i źródłem wariancji błędu. Wynika to z faktu, że jeśli osoby podchodzące do testu nie znają właściwych odpowiedzi na niektóre pytania, to udzielają odpowiedzi losowych. Zmienność wyników związana ze zgadywaniem ma charakter niesystematyczny,

dlatego obniża rzetelność pomiaru i nie wnosi niczego do trafności testu. Z tego powodu zgadywanie jest niekorzystne i powinno się dążyć do kontroli tego czynnika.

Tendencję do zgadywania można kontrolować za pomocą instrukcji testowej. W pierwszym z wariantów kandydatom można zalecać zgadywanie w każdym z przypadków, gdy nie znają prawidłowej odpowiedzi. W efekcie następuje wyrównanie pomiędzy osobami z tendencją do zgadywania i czynnik ten nie wpływa już na zmienność otrzymanych wyników. Ponieważ wszyscy testowani zgadują i mają takie same szanse, to czynnik ten powoduje, że wyniki wszystkich osób są podniesione o wynik wynikający ze zgadywania. Taka procedura prowadzi do spadku rzetelności wskutek zwiększenia się liczby odpowiedzi losowych, co jest główną wadą tego rozwiązania. Uwzględnienie czynnika zgadywania (np. przez obniżenie wyniku o poziom tzw. poprawki na zgadywanie) w wysokości progu odcięcia pozwala jednak na wzięcie pod uwagę zgadywania w procesie wnioskowania o wyniku testowym. Procedura ta wydaje się optymalna (poinformowanie osób testowanych, że jeśli nie znają prawidłowej odpowiedzi, to mogą zgadywać). Dzięki niej praktycznie eliminuje się braki w odpowiedziach i wyrównuje się tendencje między osobami, a tym samym zachęca się osoby niepewne swojej wiedzy, aby jednak udzieliły odpowiedzi. Warto dodać, że alternatywna procedura, tzn. zakazywanie zgadywania w ogóle, o tyle nie ma sensu, że nie można sprawdzić, czy osoby egzaminowane naprawdę się do niej zastosowały, a jej wprowadzenie tylko zwiększa różnice między osobami bardzo pewnymi siebie i skorymi do ryzyka a tymi, które są niepewne i boją się udzielać błędnych odpowiedzi.

W wariantcie drugim tendencję do zgadywania zmniejsza się, informując badanych, że są stosowane korekty wyników. W ten sposób wyrównuje się indywidualną tendencję do zgadywania i jednocześnie obniża się liczbę odpowiedzi losowych, przez co rzetelność pozostaje na tym samym poziomie. Możliwe są specjalne zabiegi statystyczne, pozwalające na dokonywanie korekty na zgadywanie. Zakłada się w niej, że kandydat albo zna prawidłową odpowiedź, albo zgaduje (odpowiada losowo). Otrzymany wynik zatem to liczba odpowiedzi prawidłowych bez zgadywania oraz liczba odpowiedzi odgadniętych. Liczbę odgadniętych odpowiedzi można oszacować na podstawie liczby odpowiedzi błędnych oraz liczby opcji odpowiedzi w zadaniach testowych (przy założeniu, że są jednakowo atrakcyjne).

9. Czas i miejsce stosowania testów wiedzy

Pomimo wymogu standaryzacji, tj. ujednolicenia warunków egzaminowania, zaleca się, aby testy wiedzy nie były zawsze stosowane w tej samej formie, a więc np. kolejność pytań, a także kolejność odpowiedzi w danym pytaniu może się zmieniać. Jest to odstępstwo od standaryzacji, które ma na celu zmniejszenie ryzyka w postaci wcześniejszego zapoznania się z pytaniami przez kandydata, ich zapamiętania, a także – jeśli test rozwiązywany jest jednocześnie przez grupę znajdującą się w jednym miejscu – zaglądania do odpowiedzi innych osób. Dobrze też, jeśli pytania do testu są losowane z większej puli gotowych pytań. Dzięki temu test nie ma zawsze tej samej formy i jest bardziej prawdopodobne, że osoby podchodzące do weryfikacji nie poznały wcześniej pytań i klucza odpowiedzi. Testy w formie papierowej powinny więc być generowane na potrzeby danego egzaminu, nie zaś powielane w niezmiennej formie przez dłuższy czas. Pomocą do obsługi takich testów mogą być specjalne programy komputerowe, które pozwalają na przechowywanie puli pytań, przeprowadzają losowanie pytań (np. wybranie 20 pytań z puli 100) oraz w przypadku pytań zamkniętych układają w losowej kolejności odpowiedzi. Programy te pomagają także w ocenianiu wyników testu. Dzięki skanowaniu odpowiedzi zapisywanych przez kandydata na specjalnym formularzu wyniki w teście są liczone w sposób automatyczny, co z jednej strony zmniejsza ryzyko pomyłki podczas liczenia, a z drugiej zwiększa obiektywizm procedury przeprowadzania walidacji metodą testu.

9.1. Testy w wersji komputerowej

Innym wyborem jest przeprowadzenie testowania w całości z użyciem aplikacji (wersja komputerowa). W takim wypadku pytania lub zadania są prezentowane na ekranie komputera, a odpowiedzi udzielane za pomocą klawiatury, myszy czy ekranu dotykowego. W przypadku wersji komputerowej odpowiedzi są zapisywane w zbiorze danych, zatem ich cyfrowa archiwizacja jest automatyczna. Program komputerowy może także losować kolejność pytań, a w pytaniach zamkniętych również kolejność opcji odpowiedzi. W przypadku egzaminowania z użyciem komputera można także zastosować bardziej złożony algorytm wyboru kolejnych zadań lub pytań (por. teoria odpowiedzi na pozycje testowe).

Zaletą badania z użyciem komputera jest także większa łatwość standaryzacji warunków. Na przykład prościej jest nałożyć ograniczenia czasowe na rozwiązywanie danego testu – po upływie pewnego czasu test automatycznie się wyłącza lub następuje przejście do kolejnego pytania, a brak odpowiedzi jest traktowany jako błąd. Używanie komputera niesie ze sobą też pewne wyzwania. Po pierwsze wymaga wyposażenia centrum egzaminacyjnego w odpowiedni sprzęt. Należy też zadbać o to, aby w trakcie wypełniania testu osoby egzaminowane nie mogły korzystać z podpowiedzi znalezionych choćby w Internecie. Sprzęt ten musi być więc odpowiednio skonfigurowany. Wyzwaniem jest również zróżnicowany poziom biegłości w posługiwaniu się nim. Osoby

egzaminowane, które mało lub wcale nie posługują się komputerem, mogą mieć znaczniejsze trudności w jego obsłudze, a także przeżywać większy stres podczas testowania. W obecnych czasach problem ten jednak traci na znaczeniu – umiejętność posługiwania się komputerem jest powszechna, choć nadal trudnością dla niektórych osób może być szybkie pisanie z użyciem klawiatury. Przed zastosowaniem testu z otwartym formatem odpowiedzi w formie komputerowej należy albo upewnić się, że osoby egzaminowane potrafią pisać komputerowo, albo zrezygnować z ograniczenia czasu w wypełnianiu, tak aby szybkość i biegłość w pisaniu maszynowym nie stały się czynnikiem faworyzującym lub dyskryminującym niektóre osoby.

W przypadku zadań lub pytań z zamkniętym formatem odpowiedzi komputerowa wersja testu ma znacznie więcej zalet niż wad i nie powinna budzić większych obaw. Należy wyraźnie podkreślić, że test zawsze powinien być prowadzony w takich samych (przynajmniej jak najbardziej zbliżonych) warunkach. Jeśli więc jest to komputerowa wersja testu, wszystkie osoby egzaminowane powinny mieć możliwość wypełniania testu przy użyciu komputera. Ważne, aby materiał testowy był podobnie eksponowany wszystkim osobom. W praktyce oznacza to, że osoby egzaminowane powinny wypełniać test na takim samym lub porównywalnym sprzęcie komputerowym (np. taka sama przekątna ekranu, ten sam sposób udzielania odpowiedzi, np. przy użyciu ekranu dotykowego lub myszy).

Zaletą testów komputerowych jest także to, że mogą odbywać się również za pośrednictwem Internetu (on-line), co oznacza, że osoba podchodząca do walidacji nie musi znajdować się fizycznie w siedzibie IC lub instytucji walidującej. Taka forma testowania jest szczególnie atrakcyjna w sytuacjach, gdy np. niepełnosprawność ruchowa lub duża odległość utrudnia lub wręcz uniemożliwia osobie testowanej dotarcie do miejsca przeprowadzania weryfikacji.

Do osoby testowanej zwykle jest wysyłana mailem lub przez specjalny system komputerowy IC informacja z linkiem do testu oraz otwierającym go hasłem i w czasie przez siebie wybranym lub – co częstsze – w czasie uzgodnionym z asesorem kandydat wypełnia test na własnym sprzęcie komputerowym. Programy służące do obsługi takiej formy testowania przeważnie pozwalają na kontrolę czasu wypełniania testu, a także czasowo blokują możliwość używania innych aplikacji podczas testowania, co minimalizuje ryzyko szukania podpowiedzi podczas egzaminu. Możliwe jest także zarejestrowanie takiego badania z użyciem kamery zamontowanej na komputerze osoby testowanej. Niewątpliwie jednak testowanie poza siedzibą IC czy IW, bez obecności asesora lub osoby przeprowadzającej test, niesie ze sobą ryzyko, że z jednej strony kandydaci wypełniają test w bardzo różnych (niestandardowych) warunkach, a z drugiej stają przed pokusą i szansą uzyskiwania podpowiedzi, co powoduje, że wyniki w teście przestają odzwierciedlać poziom realnej wiedzy czy umiejętności danej osoby. Należy więc zawsze rozważyć zalety i wady zastosowania takiego rozwiązania.

9.2. Czas rozwiązywania testu

Istotnym czynnikiem podczas planowania weryfikacji jest czas przeznaczony na testowanie, potrzebny na udzielenie odpowiedzi na poszczególne pytania lub rozwiązywanie zadań.

Wyróżnia się trzy typy testów w zależności od zastosowania (lub nie) ograniczeń w czasie rozwiązywania testu:

- *testy maksymalnego wykonania*, nazywane także *testami mocy*, w których zadania lub pytania różnią się poziomem trudności, czas na rozwiązanie nie jest ograniczony, a osoba testowana ma szansę zapoznać się ze wszystkimi przewidzianymi dla niej zadaniami czy pytaniami w tempie, jakie preferuje;
- *testy szybkości*, w których najistotniejszym kryterium oceny jest właśnie szybkość działania, a więc czas rozwiązywania jest ograniczony, osoba testowana nie ma szansy na rozwiązanie całej puli zadań lub odpowiedzenie na wszystkie pytania, a zadania lub pytania z reguły są na tym samym poziomie trudności (przeważnie są to zadania łatwe);
- *testy maksymalnego wykonania z ograniczeniem czasowym*, w których osoba testowana wykonuje zadania o zróżnicowanym poziomie trudności, ma na to ograniczony czas i jeśli zbyt wolno udziela odpowiedzi, może nie zdążyć ukończyć testu, a więc udzielić odpowiedzi na wszystkie przewidziane dla niej zadania czy pytania.

Wybierając typ danego testu i decydując o wyznaczeniu ograniczeń czasowych na jego rozwiązanie, przede wszystkim należy kierować się tym, co jest celem pomiaru w danym teście (określony rodzaj wiedzy, umiejętności czy kompetencji społecznych), a także uwzględnić charakter branży, do której odnosi się kwalifikacja. Decyzja o ograniczeniu czasu testowania musi być każdorazowo podjęta przy projektowaniu procedury testowej i nie ma jednej uniwersalnej porady w tym zakresie. Można wskazać szereg zalet i wad wynikających z ograniczenia czasu testowania i na tej podstawie określić (lub nie) ramy czasowe wypełniania testu.

Wśród zalet testów z ograniczeniem czasowym można wymienić:

- Znaczne ułatwienie procesu administrowania testu. Z góry wiadomo, ile czasu maksymalnie zajmie ta część egzaminu.
- Zapewnienie standardowych warunków testowania. Ograniczenie czasu powinno być zawsze takie samo dla wszystkich osób.
- Presja, która wpływa motywująco na osoby testowane – wszyscy czują się zmotwowani, żeby po rozpoczęciu testu niezwłocznie starać się udzielać odpowiedzi.
- Zwiększenie poziomu trudności testu. Nie tylko należy umieć wskazać lub sformułować prawidłową odpowiedź, ale też zrobić to szybko.

Z reguły w realnych sytuacjach zawodowych osoby testowane nie tylko muszą wykazać się pewną wiedzą, lecz także zrobić to odpowiednio szybko. Sytuacja ograniczenia czasowego jest więc z reguły bliższa tej naturalnej.

Ograniczenie czasu badania to też pewne ryzyka, a co najmniej wyzwania:

- Powoduje ono, że choć warunki są obiektywnie równe dla wszystkich, to subiektywnie ograniczenie może być różnie odczuwane. Osoby, które z jakichś względów wolniej udzielają odpowiedzi, działają w warunkach dla nich nieoptymalnych, a zatem test staje się dla nich trudniejszy, a ich wynik nie w pełni porównywalny z osobami, które działają szybko. Powodem, dla którego ludzie różnią się pod względem szybkości działania, są różnice wynikające z temperamentu, ale także wieku lub stanu zdrowia (np. trudności z czytaniem wynikające ze słabego wzroku). Stwierdzono np., że wiek jest wyraźnie ujemnie związany z czasem działania. Osoby starsze wolniej udzielają odpowiedzi.
- Stwarza presję, która z jednej strony może motywować do tego, aby szybko udzielać odpowiedzi, ale może też być źródłem nadmiernego stresu, wpływającego zaburzająco na działanie i powodującego, że osoby, które w warunkach bardziej komfortowych udzieliłyby poprawnej odpowiedzi, pod presją czasu zostaną fałszywie zakwalifikowane jako niemające danej wiedzy.
- Może powodować, że osoby testowane nie zdążą odpowiedzieć na wszystkie przewidziane dla nich pytania lub rozwiązać wszystkie zadania, a braki odpowiedzi będą traktowane tożsamo z odpowiedziami błędnymi. Ten problem jest szczególnie istotny wtedy, gdy zadania lub pytania w teście nie są ułożone w sposób rosnący pod względem trudności i po pytaniach trudnych, wymagających dłuższego namysłu, mogą pojawić się pytania łatwe, na które kandydat udzieliłby dobrej odpowiedzi, gdyby miał więcej czasu.

Podsumowując, przede wszystkim zaleca się, aby wszelkie ograniczenia w czasie pisania testu były podejmowane jako świadoma, przemyślana i uzasadniona decyzja konstruktora danego testu. Jej konsekwencje powinny być sprawdzone. Jeśli ograniczono czas egzaminowania, a intencją twórcy testu jest uzyskanie testu maksymalnego wykonania, a nie testu szybkości, należy się upewnić, czy czas nie jest zbyt krótki. Dobrze też sprawdzić, jak osoby testowane subiektywnie oceniały ten czas. Czy zgłaszały, że miały poczucie, że jest go za mało i nie zdążyły wykazać się wiedzą? Czy zgłaszały, że ograniczenie w czasie jest bardzo stresujące? Podejrzenia, że czas przeznaczony na egzaminowanie może być, przynajmniej dla części osób, zbyt krótki, musi skłaniać do rewizji założeń i zmiany warunków testowania.

Ostatnią kwestią związaną z warunkami administrowania testu jest możliwa liczba powtórzeń w używaniu testu u tej samej osoby. Możliwa jest sytuacja, że kandydat przed przystąpieniem do testu zna pulę potencjalnych pytań lub zadań. Ma wtedy szansę nie tylko poznać je wszystkie, lecz także prawidłowo rozwiązać i zapamiętać odpowiedzi. Podczas walidacji wypełnia więc test po raz kolejny, a liczba tych powtórzeń – jeśli pytania lub zadania są udostępnione przed walidacją – nie jest określona.

Takie rozwiązanie ma szczególnie sens wtedy, gdy:

- pula zadań lub pytań jest duża i praktycznie niemożliwe jest mechaniczne zapamiętanie klucza bez zrozumienia treści pytań i prawidłowych odpowiedzi;
- od kandydata oczekuje się bardzo wysokiego poziomu wykonania testu, a więc udzielenia poprawnych odpowiedzi na wszystkie lub niemal wszystkie pytania.

Taka sytuacja zdarza się np. podczas testów dla osób starających się o uprawnienia do kierowania pojazdem (egzamin na prawo jazdy). Celem tych egzaminów jest uzyskanie pewności, że egzaminowany zna kodeks ruchu drogowego i potrafi zachować się prawidłowo na drodze nie z reguły lub czasem, ale zawsze lub niemal zawsze. W takim przypadku nie ma powodu ukrywać przed egzaminowanym treści pytań, bo zakłada się, że jeśli – nawet wcześniej ćwicząc – potrafi udzielić poprawnych odpowiedzi, to świadczy to o posiadaniu odpowiedniej wiedzy, co jest przecież celem egzaminowania.

Ryzykiem związanym z udostępnianiem treści pytań przed weryfikacją jest pojawienie się „fałszywej wiedzy”. Oznacza to, że osoba podchodząca do testu nauczyła się na pamięć poprawnych odpowiedzi i zaznaczanie tych prawidłowych nie jest w jej wypadku wskaźnikiem posiadania wiedzy. Gdyby pulę pytań, które znała i trenowała odpowiedzi na nie, zastąpić nowymi, równoważnymi (mierzącymi tę samą wiedzę) pytaniami, to wynik przez nią uzyskany byłby znacząco odmienny. Ryzyko to skłaniałoby do zastosowania strategii odwrotnej – unikania wcześniejszego ujawniania pytań, niepozwalanie na powtórne wypełnienie tego samego testu oraz okresowej wymiany zadań lub pytań w teście, tak aby osoby testowane miały jak najmniejsze szanse na zapoznanie się z pytaniami przed weryfikacją.

10. Ocenianie odpowiedzi na zadania i pytania testowe

10.1. Budowanie wskaźników na podstawie wyniku testu

Cechą testów wiedzy jest to, że jakościowe odpowiedzi osób badanych, tj. sformułowanie własnej odpowiedzi w testach z tzw. otwartym formatem odpowiedzi lub wybranie odpowiedzi z odpowiedniej listy w testach z zamkniętym formatem odpowiedzi, mogą dość łatwo zostać przełożone na wskaźniki ilościowe. Zaletą wskaźników ilościowych jest łatwość ich agregowania, np. sumowania, a także możliwość precyzyjnej oceny w kategoriach potwierdzenia lub niepotwierdzenia danego kryterium, w przypadku kwalifikacji rynkowych – efektu uczenia się (ustalenie minimalnej liczby punktów – progu, od którego uznaje się, że dany efekt uczenia się jest potwierdzony). Aby odpowiedzi osoby testowanej przypisać odpowiednią liczbę punktów, należy określić algorytm dla takiej zamiany, nazywany w teście kluczem.

Należy pamiętać, że klucz w teście:

- zawsze jest określany na etapie konstruowania testu. Od początku trzeba określić, które lub jakie odpowiedzi w teście uzna się za poprawne, a które lub jakie za błędne. Sposób punktowania odpowiedzi musi zostać w pełni i szczegółowo opisany, zanim podda się testowaniu pierwszą osobę;
- powinien być jednoznaczny i taki sam dla wszystkich pomiarów (kandydatów). Odpowiedź powinna uzyskać taką samą liczbę punktów niezależnie od tego, kto jej udzielał oraz kto ją oceniał.

Jeśli zastosowanie klucza wymaga interpretacji, co mówi lub zapisała osoba egzaminowana, zaleca się, aby więcej niż jeden asesor oceniał te odpowiedzi i przydzielał punkty, a ostateczna ich liczba była np. średnią tych ocen. Jeśli nie ma takiej możliwości i ocena musi pochodzić od jednej osoby, powinna ona nie tylko przejść szczegółowe szkolenie, ale również proces jej oceniania powinien podlegać rewizji (*efekt egzaminatora* jest opisany w rozdz. 10.3).

Decyzja o przydzielonej liczbie punktów czy rozpiętości skali ocen może być dowolnie kształtowana przez zespół konstruujący test. Poprawną odpowiedź można równie dobrze punktować jako 1 pkt; 0,1 pkt, jak i 100 pkt. W praktyce jednak należy kierować się wygodą i zdrowym rozsądkiem. Jeśli odpowiedź może być oceniona tylko jako 0 pkt lub jako 100 pkt, tak szeroka skala nie ma żadnego uzasadnienia – w tym wypadku nie zakłada się istnienia punktów od 1 do 99. Stosowanie dodatkich liczb całkowitych to najczęściej przyjmowane rozwiązanie. Ocena odpowiedzi na skali 0 lub 1 jest zgodna z intuicją, a tak powstała skala oddaje rzeczywisty zakres możliwej zmienności.

Oczywiście ilościowe przekształcenie odpowiedzi nie musi odbywać się w formacie binarnym: 0 i 1. Odpowiedź osoby testowanej może być zakodowana na skali, w której możliwe są do uzyskania różne wartości, np. 0, 1, 2 i 3. Ważne jednak jest to, że decyzja o przyznaniu określonej liczby punktów musi być podejmowana na podstawie określonego i stałego schematu oceniania.

Przykłady schematu oceniania

Przykład 1.

W zadaniu z zamkniętym formatem odpowiedzi, w których możliwe są odpowiedzi a, b, c i d, można zdefiniować, że:

- zaznaczenie odpowiedzi c lub d zawsze oznacza przyznanie 0 pkt,
- zaznaczenie tylko odpowiedzi a lub tylko b prowadzi do uzyskania 1 pkt,
- łączne zaznaczenie odpowiedzi a i b przekłada się na 2 pkt.

Tak przyjęty klucz zakłada, że zaznaczenie odpowiedzi a i b jest dwa razy lepsze niż zaznaczenie tylko odpowiedzi a lub tylko b, ponieważ „odległość” pomiędzy 0 a 1 jest taka sama jak pomiędzy 1 i 2.

Przykład 2.

W zadaniu z otwartym formatem odpowiedzi osoba testowana ma wymienić jak najwięcej znanych jej sposobów poradzenia sobie z danym problemem, przy czym musi wskazać nie mniej niż dwa sposoby:

- jeśli wskazuje jeden tylko sposób, otrzymuje 0 pkt (nawet jeżeli jest to sposób trafny i jest to taka sama ocena jak wtedy, gdy nie wymieniła żadnego sposobu, wymieni kilka, ale żaden z nich lub tylko jeden będzie prawidłowy),
- jeśli wskazuje dwa sposoby, otrzymuje 1 pkt,
- opisuje trzy sposoby – 2 pkt,
- a gdy opisuje cztery lub więcej sposobów – 3 pkt.

W tym wypadku znacznie trudniej uznać, że osoba, która otrzymała 2 pkt, w porównaniu z osobą, która dostała 1 pkt, wie dwa razy więcej. Chociaż „odległości” pomiędzy 0 i 1, 1 i 2 oraz 2 i 3 są matematycznie takie same, to nie wiadomo, czy w taki sposób naprawdę kształtują się „odległości” pomiędzy osobami w poziomie wiedzy. Są to raczej pewne upraszczające założenia, które trzeba przyjąć i których należy być świadomym, gdy wnioskuje się na podstawie tak skonstruowanego wyniku.

Istotą testów wiedzy jest możliwość przekształcenia odpowiedzi osoby testowanej na wynik ilościowy. Podstawą, która je umożliwia, jest klucz odpowiedzi.

Czy może istnieć test wiedzy składający się tylko z jednego zadania lub pytania? Teoretycznie jest to możliwe. Celem konstruowania testu jest uzyskanie wskaźnika lub wskaźników, które pozwalają na potwierdzenie osiągnięcia danego efektu

uczenia się. W praktyce jednak może być dość trudno sformułować tylko jedno zadanie, które pozwala na taką weryfikację. Z reguły wiedza, którą musi wykazać się badany, jest raczej obszerna niż wąska, musi być użyta w różnych sytuacjach i kontekstach, a to oznacza, że wskaźnik dla danego efektu uczenia się musi stanowić agregację odpowiedzi na wiele pytań lub zadań. Dlatego drugim krokiem po sprecyzowaniu sposobu punktowania odpowiedzi w danym zadaniu i pytaniu jest ustalenie sposobu łączenia odpowiedzi z różnych pytań lub zadań.

Na tym etapie konstrukcji klucza niezbędna jest wiedza o trafności treściowej i kryterialnej pytań lub zadań. Sensowne jest łączenie ze sobą odpowiedzi na te pytania, o których wiemy, że odnoszą się wspólnie do danego efektu uczenia się. Jeżeli natomiast pytania w teście sprawdzają osiągnięcie różnych efektów uczenia się, lepiej na ich podstawie zbudować odrębnie zagregowane wskaźniki. A zatem generując zadania lub pytania w teście, trzeba wiedzieć nie tylko to, jak będą oceniane odpowiedzi, ale też jak na ich podstawie zostanie zbudowany wskaźnik (czyli zagregowana miara, która przykładowo powstaje na podstawie zsumowania odpowiedzi z kilku pytań), pozwalający na weryfikację danego efektu uczenia się.

Istotne jest, jak wiążą się ze sobą odpowiedzi na poszczególne pytania. Generalnie oczekuje się, że skoro odpowiedzi na różne zadania lub pytania opisują łącznie jakiś obszar wiedzy lub umiejętności, to powinno się zaobserwować tendencję do skorelowania ze sobą takich zadań lub pytań. Takie analizy można jednak przeprowadzić dopiero po zebraniu wyników od grupy osób i w przypadku testu składającego się z 20 pytań do analiz potrzebne będzie zebranie wyników od co najmniej 100 osób (zwykle potrzeba 5–10 razy więcej osób wypełniających test w stosunku do liczby pytań lub zadań).

Dla każdej pozycji testowej można wtedy obliczyć *moc dyskryminacyjną*. Moc dyskryminacyjna danej pozycji testowej informuje o tym, w jakim stopniu różnicuje ona daną populację pod względem cechy mierzonej przez test. Jest ona zatem wskaźnikiem tego, w jakim stopniu pozycja testowa mierzy tę cechę, którą mierzy cały test – informuje o tym, jak dobrą miarą tej cechy jest dana pozycja. Duża moc dyskryminacyjna pozycji testowych wskazuje na dużą homogeniczność testu – zazwyczaj wykazuje on wysoką rzetelność pomiaru. Rzetelność pomiaru danego testu jest więc uzależniona bezpośrednio od mocy dyskryminacyjnej pozycji. Do jej oszacowania stosuje się specjalne współczynniki korelacji.

Jak agregować, czyli łączyć ze sobą wyniki z kilku pytań? Najprostszym i najczęstszym sposobem jest sumowanie zakodowanych odpowiedzi z poszczególnych zadań lub pytań. Sumując, uznaje się, że odpowiedź na jakiegokolwiek pytanie nie świadczy o osiągnięciu lub nie efektu uczenia się i dopiero połączenie odpowiedzi z kilku pytań pozwala na zbudowanie takiego wskaźnika.

Dodając odpowiedzi z kilku pytań, należy zwrócić uwagę na ich rodzaj oraz zakres punktów możliwych do uzyskania z każdego z pytań. Te, które wchodzi w skład danego wskaźnika, w zasadzie powinny być tego samego formatu (np. wszystkie to pytania z formatem zamkniętym z czterema opcjami odpowiedzi do wyboru, w tym jedną poprawną), choć można sobie wyobrazić odstępstwa od tej reguły (np. gdy dodaje się do siebie punkty z pytań z zamkniętym formatem odpowiedzi oraz punkty z zadań z lukami). To pierwsze rozwiązanie jest w pełni poprawne

metodologicznie – sensowne jest sumowanie rzeczy podobnych. Na przykład jeśli chce się obliczyć odległość, nie można dodawać do siebie pomiarów wyrażonych w calach oraz metrach. W przypadku testów wiedzy można jednak poczynić odstępstwa od tej reguły, tzn. dodać do siebie punkty zebrane w teście z zamkniętym wyborem do punktów zebranych w pytaniach otwartych. Wymaga to jednak uzasadnienia merytorycznego, dlaczego wybrano takie rozwiązanie, i świadomości, że z punktu widzenia zasad konstruowania skal jest ono mniej typowe.

Poszczególne pytania mogą się różnić liczbą punktów, które można otrzymać za poprawną odpowiedź. Jeśli w pytaniu X można zdobyć 0, 1, 2 lub 3 pkt, a w pytaniu Y tylko 0 pkt lub 1 punkt, to po zsumowaniu tych pytań pytanie X silniej zaważy na całociowym wyniku. Jest to więc zagadnienie nadawania wag poszczególnym pytaniom.

Teoretycznie jest możliwe, że pytanie X rzeczywiście mierzy dany obszar wiedzy w szerszym zakresie i przypisywanie mu większego znaczenia może być uprawnione. Jeśli tak jednak by nie było, to opisany sposób sumowania pytań byłby błędny. Jeżeli odpowiedź na pytanie X jest równie dobrym wskaźnikiem wiedzy jak odpowiedź na pytanie Y, to osoba, która potrafi odpowiedzieć tylko na jedno z tych pytań, powinna dostać sumarycznie tyle samo punktów niezależnie od tego, na które pytanie odpowiada. W tym wypadku tak by jednak nie było. Osoba, która w pełni poprawnie odpowiedziała na pytanie X i nie odpowiedziała na pytanie Y, otrzyma 3 pkt na 4 pkt możliwe, natomiast osoba, która w ogóle nie zna odpowiedzi na pytanie X, ale zna na Y, tylko 1 pkt na 4 pkt.

Przypisywanie wag poszczególnym pytaniom musi być procesem teoretycznie (analiza efektu uczenia się) lub empirycznie (pewne pytania są lepsze diagnostycznie) uzasadnionym. Jeśli takiego uzasadnienia brakuje, należy tak zsumować wyniki z poszczególnych pytań, aby każde z nich w takim samym stopniu zaważyło na sumarycznym wyniku. W praktyce więc lepiej jest łączyć ze sobą pytania, które nie różnią się formatem i które mają ten sam zakres możliwych do uzyskania punktów.

10.2. Ocenianie odpowiedzi

Gdy definiuje się określoną regułę zliczania wyników w teście, kolejnym krokiem jest podjęcie decyzji, kiedy dany wynik uznaje się za świadczący o potwierdzeniu danego efektu uczenia się. Jest to więc zagadnienie wyznaczania progu minimalnej liczby punktów. Po jego przekroczeniu stwierdzi się, że dany efekt uczenia się został potwierdzony.

Należy wyraźnie podkreślić, że chociaż w ramach prowadzenia walidacji przyjmuje się, że dany efekt uczenia się musi być osiągnięty, by można było go potwierdzić w stu procentach, to nie oznacza, że wskaźnik, za pomocą którego jest dokonywana jego ocena, również musi osiągnąć wartość stu procent. Nie trzeba więc zakładać, że kandydat musi zdobyć sto procent punktów w teście wiedzy. Można sobie wyobrazić, że dany test pozwala na zmierzenie szerszego zakresu wiedzy niż ten, który potrzebny jest do potwierdzenia danego efektu uczenia się. Ponadto

rzetelność testu nigdy nie jest doskonała, a otrzymany wynik ujmuje zarówno wynik prawdziwy, jak i losową wariancję błędu.

Jaka liczba punktów osiągniętych w teście może być uznana za wystarczającą?

Na to pytanie nie ma niestety jednej uniwersalnej odpowiedzi. W praktyce próg ten może być zdefiniowany bardzo różnie i zależy od konstrukcji testu i związków jego wyników z zewnętrznym kryterium. Najlepszym sposobem na wyznaczenie takiego progu byłoby zebranie wyników w teście i porównanie ich z tymi, które w inny i zarazem wiarygodny sposób pozwalają potwierdzić dany efekt uczenia się. Jeżeli brakuje takich danych, próg ten musi zostać wyznaczony arbitralnie, na podstawie analizy treści pytań i określenia tego, w jakim stopniu i w jakim zakresie weryfikują one dany efekt uczenia się. Warto jednak, w trosce o jakość weryfikacji, po zebraniu odpowiednio licznej próbki danych poddać je analizom i w razie konieczności podwyższyć lub obniżyć poziom tego progu, nazywanego też *punktem odcięcia* (ang. *cut score*).

Ocena punktu odcięcia jest dokonywana przez wyznaczenie dwóch charakterystyk testu, tj. jego *czułości* (inaczej wrażliwości; ang. *sensitivity*) i *swoistości* (inaczej specyficzności; ang. *specificity*), wykorzystując do tego odpowiednie metody matematyczne i psychometryczne (Coombs, Dawes i Tversky, 1977).

Czułość i swoistość testu diagnostycznego to wartości opisujące zdolność testu do poprawnego wykrycia badanej cechy (czułość), jak też stwierdzenia jej braku (swoistość).

Stuprocentowa czułość czy wrażliwość testu oznaczałaby, że wszystkie osoby, których poziom wiedzy pozwala na potwierdzenie danego efektu uczenia się, zostaną wyłonię w wyniku testowania (ang. *correct-pass decisions*).

Stuprocentowa swoistość czy specyficzność świadczyłaby o tym, że wszyscy kandydaci, których wiedza jest niewystarczająca do potwierdzenia określonych efektów uczenia się na podstawie wyniku testu, zostaną wskazani jako ci, którzy nie potwierdzili tych efektów uczenia się dla danej kwalifikacji (ang. *correct-fail decisions*).

W praktyce niecelowe jest wyznaczenie progu odcięcia wyników zarówno w obrębie niskich, bardzo niskich, jak i bardzo wysokich wartości. Szczególnie próg położony nisko lub bardzo nisko świadczyłby, że test jest bardzo trudny, w którym już wynik np. na poziomie 20% czy 30% pozwala na potwierdzenie efektu uczenia się. Wykorzystanie takiego testu może stwarzać wrażenie, że z nieuzasadnionych powodów zastosowano zbyt trudną procedurę, która naraziła osoby egzaminowane na niepotrzebny stres. Test, tworzony w celu potwierdzenia danego efektu uczenia się, powinien zawierać zadania czy pytania nie tylko o adekwatnej treści (trafność), lecz także o odpowiednim poziomie trudności, który znacząco nie przekracza tego, który określa przypisany poziom PRK dla danej kwalifikacji rynkowej.

Należy podkreślić, że wyznaczanie progu odcięcia nie może być dokonane na podstawie wyników kandydatów (np. nie można przyjąć, że poziom zdawalności to górne 20% wyników), ponieważ poziom wiedzy i szans na posiadanie tej wiedzy wśród kandydatów jest nieznan. Osoby testowane mogły zdobyć kwalifikacje

poza systemem edukacji formalnej, a zakres tych doświadczeń, nieformalnej i pozaformalnej nauki nie jest znany.

Trudność pozycji to proporcja odpowiedzi poprawnych do wszystkich odpowiedzi – im zadanie łatwiejsze, tym wyższy jest ten współczynnik, a jego zakres rozciąga się od 0 do 1. Wynik 0 osiąga ta pozycja, na którą nikt nie udzielił prawidłowej odpowiedzi, a wynik 1 – wszyscy odpowiedzieli poprawnie. Zakłada się, że miara trudności pozycji wynika z cech jej samej i jest miarą opisującą wewnętrzne zróżnicowanie pozycji testowych. Różnice w ich trudności pozwalają na określenie różnic w poziomie nasilenia mierzonej cechy (wiedzy) u osób testowanych. Osoby mające małą wiedzę potrafią rozwiązać tylko pozycje łatwe, osoby mające średni poziom wiedzy radzą sobie z łatwymi i średnio trudnymi pozycjami, ale niewielka jest szansa, że rozwiążą pozycje trudne. Z kolei osoby o bardzo wysokim poziomie wiedzy mogą rozwiązać wszystkie zadania w teście – te łatwe, średnio trudne i trudne. Ważne jest więc, aby test zawierał zrównoważoną liczbę pozycji łatwych, średnich i trudnych, a średnia trudność pozycji oscylowała wokół wartości 0,5. Dzięki temu dobrze różnicuje osoby podchodzące do walidacji i nie jest ani za łatwy (taki test nie rozróżni tych z przeciętną wiedzą od tych z wysokim poziomem wiedzy), ani za trudny (pozwoli na rozróżnienie tych z przeciętną wiedzą od tych z niskim poziomem wiedzy).

Do walidacji mogą przystąpić kandydaci o różnym stopniu przygotowania. Jeśli np. do weryfikacji przystępują kandydaci, którzy są niewystarczająco przygotowani, to ich niskie wyniki mają źródło zarówno w cechach pozycji testowych (trudne pytania), jak i tym, że nie mieli szansy ich rozwiązać (dla osoby niemającej wystarczającej wiedzy w danym obszarze nawet łatwe pytania stają się trudne). Dlatego też – o ile nie ma pewności, że do testu przystępują osoby o zbliżonych szansach na zdobycie danej wiedzy – miary trudności pozycji oraz analiza zdawalności testu muszą być stosowane z dużą ostrożnością. W innym razie istnieje ryzyko, że:

- zostaną potwierdzone efekty uczenia się u osoby, która nie ma odpowiedniej wiedzy (zbyt nisko ustawiony próg odcięcia);
- nie zostaną uznane efekty uczenia się osobie, która dysponuje odpowiednią wiedzą (zbyt wysoko ustawiony próg odcięcia).

A zatem podstawowym punktem odniesienia przy wyznaczaniu progu odcięcia jest analiza kryterium testowego, którym w tym wypadku jest kryterium weryfikacji efektu uczenia się.

10.3. Przygotowanie osób oceniających, *efekt asesora* i osiągnięcie zgodności asesorów przy ocenie odpowiedzi otwartych

Osoba przeprowadzająca test i asesor oceniający jego wyniki (mogą to być dwie różne osoby) nie muszą być jednocześnie konstruktorem danego testu. Jeśli osoba przeprowadzająca test i asesor otrzymują gotowy test, ich obowiązkiem powinno być:

- poznanie założeń leżących u źródeł testu (np. zakresu wiedzy, opisu danej kwalifikacji, opisu efektów uczenia się);
- szczegółowe poznanie procedury testowania;
- szczegółowe poznanie sposobu oceny odpowiedzi (klucza);
- nabycie umiejętności obliczania wyniku w teście i wnioskowania o nim (w celu wydania decyzji walidacyjnej).

Ze względów praktycznych wiedza o danym teście powinna zostać spisana i udostępniona asesorom, umożliwiając im wcześniejsze przygotowanie się. Niekiedy proces samodzielnego uczenia się asesorów jest niewystarczający i konieczne jest szkolenie przede wszystkim w zakresie sposobów kodowania odpowiedzi. Taka sytuacja niemal zawsze zachodzi w przypadku oceny odpowiedzi na pytania otwarte (zarówno ustne, jak i pisemne). Asesorzy powinni mieć dostęp do klucza odpowiedzi i nauczyć się jego stosowania.

Przed rozpoczęciem testowania asesor powinien zapoznać się z kluczem kodowania odpowiedzi, mieć możliwość wyjaśnienia ewentualnych wątpliwości oraz poćwiczyć samodzielne kodowanie pod nadzorem bardziej doświadczonego asesora lub twórcy (konstruktora) testu. Wartościowe jest także udostępnienie przykładowych odpowiedzi (najlepiej pochodzących z prawdziwych, pilotażowych walidacji) i przypisanych im punktacji. Nie wystarczy więc tylko sformułowanie zasad oceny odpowiedzi na pytania testowe lub rozwiązania zadania.

W przypadku kodowania pytań otwartych, a więc przypisywania wartości liczbowych odpowiedziom osób testowanych, należy założyć, że prowadzący taką procedurę asesor był ekspertem w zakresie wiedzy, którą ocenia. Oznacza to, że ma wiedzę co najmniej w zakresie i na tym poziomie, który opisano w danej kwalifikacji rynkowej. Im mniejszy poziom biegłości asesora, tym większa potrzeba szczegółowego opisu klucza kodowania odpowiedzi.

Niezależnie od poziomu wiedzy eksperckiej asesorów zawsze istnieje potrzeba opracowania dokładnego klucza kodowania odpowiedzi. W pytaniach zamkniętych ma on postać łatwego do zastosowania algorytmu, natomiast w pytaniach z otwartym formatem odpowiedzi wymaga interpretacji i powinien zawierać jak największą liczbę przykładowych odpowiedzi ze wskazaniem, jaką liczbę punktów należy im przypisać, a także opis i wyjaśnienie ogólnych reguł oceniania. Przykład zawarto w ramce (s. 63).

*Przykład klucza kodowania odpowiedzi na pytanie otwarte***Co świadczy o tym, że kandydat wyjaśnia dany proces? Oznacza to, że:**

- 1) trafnie wskazuje na kolejne etapy i przekształcenia danego procesu (0–1 pkt),
- 2) wskazuje na wszystkie etapy we właściwej kolejności (0–1 pkt),
- 3) wyjaśnia mechanizm przechodzenia do kolejnych etapów (0–1 pkt),
- 4) odwołuje się do mechanizmów sprawczych w danym procesie, trafnie je identyfikując (0–1 pkt).

W przypadku pytań z otwartym formatem odpowiedzi wskazane jest, aby ocenianie odpowiedzi było dokonywane przez więcej niż jednego asesora. Dzięki temu zmniejsza się wpływ tzw. *efektu oceniającego* czy *egzaminatora* (ang. *rater effect*), czyli specyficznego stylu oceniania asesora i jego indywidualnej interpretacji klucza oceny, a cała procedura weryfikacji staje się bardziej obiektywna. Mianem *efektu asesora*⁸, czy *efektu egzaminatora*, określa się wiele czynników wpływających na ocenę niezwiązaną z rzeczywistym poziomem mierzonej wiedzy kandydata, ale z asesorem (Scullen, Mount i Goff, 2000). Najczęściej w specjalistycznej literaturze występują (Saal i in., 1980): *efekt halo* (inaczej *efekt aureoli* za: Wojciszke, 2000), łagodność i surowość oraz zgodność (rzetelność).

Uwzględnianie wyników ocen więcej niż jednego asesora może być realizowane w taki sposób, że asesorzy:

- dokonują niezależnych ocen, a na ich średniej opiera się ocenianie testu;
- oceniają niezależnie, a ocena testu opiera się na tej najczęściej przyznawanej. W tym wypadku konieczne jest, aby oceny pochodziły od nieparzystej liczby asesorów. Istnieje jednak ryzyko, że żadna z nich nie będzie najczęstszą, np. trójka asesorów proponuje trzy różne liczby punktów dla tej samej odpowiedzi;
- pracują razem i dokonują wspólnej oceny, uzgadniając ze sobą stanowiska i przekonując się wzajemnie do swoich punktów widzenia.

Trzeci rodzaj oceny można określić mianem komisyjnego. Dwa pierwsze sposoby to oceny przez tzw. sędziów niezależnych (asesorów, oceniających, egzaminatorów).

Zaletą komisyjnej procedury oceny jest możliwość dyskusji nad nią. Jest ona zatem pogłębiona i dobrze uzasadniona. Ryzykiem takiego sposobu oceniania jest jednak to, że może na niej zaważyć czynniki pozamerytoryczne. Wyższa pozycja w grupie jednego z asesorów, czyli np. większy autorytet czy lepsze umiejętności przekonywania innych, może spowodować, że oceny tej właśnie osoby zdominują prace komisji. Ocena

⁸ W teorii i w badaniach edukacyjnych znany jest *efekt oceniającego*, który w kontekście pomiaru edukacyjnego, np. egzaminów zewnętrznych, w Polsce znany jest jako *efekt egzaminatora*. Ponieważ w publikacjach dotyczących ZSK często pojawia się pojęcie „asesor”, to na potrzeby *Wskazówek* określono to zjawisko mianem *efektu asesora*. Więcej informacji o *efekcie egzaminatora* można znaleźć np. w publikacjach Dubieckiej, Szaleńca i Węziak (2006), http://www.ptde.org/file.php/1/Archiwum/XII/Efekt_egzaminatora.pdf oraz Henryka Szaleńca i współpracowników (2015), <http://pwe.ibe.edu.pl/opisy/efektEgzaminatora/Efekt%20egzaminatora.pdf>

dokonywana przez ekspertów niezależnych jest wolna od takiego ryzyka. Możliwe jest wreszcie połączenie tych dwóch metod: w pierwszym etapie asesory dokonują ocen niezależnych od siebie, a potem poddają je dyskusji, rezygnując z mechanicznych reguł przyznawania ostatecznej liczby punktów (średnia arytmetyczna lub najczęściej padająca wartość) i wypracowują wspólne rozwiązanie.

10.4. Wynik testu wiedzy i udzielanie informacji zwrotnej o wyniku

Wśród praw osób testowanych bez wątpienia można wskazać:

- prawo do uzyskania informacji o wyniku,
- prawo do zachowania poufności wyników (AERA, APA i NCME, 1999/2014).

Pierwsze oznacza, że kandydat może, a nawet powinien otrzymać informację zwrotną o wynikach testowania. Może mieć ona ilościowy i/lub opisowy charakter, co oznacza, że osoba testowana otrzymuje informację o liczbie osiągniętych punktów w teście wraz z informacją o progu odcięcia wyników, a także może zostać poinformowana o potwierdzonych lub nie efektach uczenia się wraz z kryteriami. Informacja może też dotyczyć tego, jak bardzo – na tle innych kandydatów, którzy napisali test – jest to wynik niski, przeciętny lub wysoki (można podać procentowo liczbę osób, które uzyskały takie same lub niższe wyniki). Informacja zwrotna nie musi, a nawet nie powinna (w trosce o zachowanie poufności materiału testowego) zawierać informacji o treści poszczególnych pytań oraz jak osoba testowana na nie odpowiedziała i które odpowiedzi są poprawne.

Przekazując informację zwrotną, zawsze należy kierować się zasadami wymienionymi poniżej, a zatem informacja zwrotna powinna:

- odnosić się do efektów testu, a nie do osoby. Przedmiotem oceny jest wiedza, a nie osoba testowana;
- skupiać się na opisanu faktów. Faktem w tym przypadku jest potwierdzenie lub nie efektów uczenia się dla danej kwalifikacji rynkowej. Informacja zwrotna powinna unikać interpretacji, czyli np. wnioskowania o powody uzyskania danego wyniku;
- być formułowana językiem unikającym ocen. Na przykład wynik w teście wyniósł x i jest to wartość pozwalająca na potwierdzenie efektu uczenia się dla danej kwalifikacji. Dodawanie, że jest to wynik dobry lub zły, zawierałoby niepotrzebny element oceny;
- być przekazana w taki sposób, by jej sens był zrozumiały. Należy unikać żargonu i zbyt trudnego słownictwa;
- być tak przekazana, żeby pozwoliła zrozumieć, co można zrobić, aby w przyszłości osiągnąć lepszy efekt;
- być przekazana bez zbędnej zwłoki po ocenieniu wyników testu;
- być wyrażona w sposób niepubliczny. Osoba testowana ma prawo zachować wiedzę o wyniku testu tylko dla siebie.

Jak wspomniano, kandydat ma prawo do zachowania poufności wyników. W sensie prawnym wyniki weryfikacji stanowią przedmiot przetwarzania danych osobowych. Przed przystąpieniem do weryfikacji osoba testowana wyraża zgodę na pozyskanie, przetwarzanie i przechowywanie jej danych osobowych. Taka zgoda jest konieczna, aby weryfikacja mogła być prowadzona. Jeśli dane pozyskiwane w tych procedurach są udostępniane innym instytucjom, np. w celu kontroli lub zewnętrznego zapewniania jakości, kandydat ma prawo wiedzieć, komu, na jakich warunkach oraz w jakim celu będą udostępnione te informacje. Zawsze, gdy jest to możliwe, należy dążyć do anonimizacji zbioru danych, na którym prowadzi się analizy mające na celu poprawę procesu testowania. Osoba testowana ma prawo oczekiwać, że dostęp do jej wyników testu jest chroniony przed osobami niepowołanymi. Wynik powinien być komunikowany w taki sposób, aby zmniejszyć ryzyko, że zapoznają się z nim inne osoby. Na przykład powinien być przekazany w zamkniętej kopercie lub w zaszyfrowanej hasłem korespondencji elektronicznej.

11. Podsumowanie: wskazówki dla instytucji certyfikujących dotyczące tworzenia i stosowania testów wiedzy w weryfikacji efektów uczenia się

Wewnętrzne zapewnianie jakości zarówno walidacji, jak i certyfikowania jest ustawowym obowiązkiem IC (ustawa o ZSK art. 50). Zgodnie z treścią ustawy o ZSK (art. 63) na system wewnętrznego zapewniania jakości (WSZJ) powinny się składać: zasady postępowania, procedury, stosowane metody oraz rozwiązania organizacyjne. WSZJ ma też jasno zdefiniowane cele, które powinny nie tylko gwarantować zapewnienie jakości walidacji i certyfikowania, lecz także być ukierunkowane na stałe ich doskonalenie. O zewnętrzne zapewnianie jakości nadawanych kwalifikacji dbają podmioty zewnętrznego zapewniania jakości (PZZJ) (rozdział 5 ustawy o ZSK)⁹.

Przesłanki zawarcia w WSZJ kwestii związanych z rzetelnością, trafnością i adekwatnością metod walidacji są w istocie bezpośrednią pochodną ogólnych zasad tworzenia dobrych narzędzi walidacji, np. testów, i właściwego ich wykorzystywania. W tym rozdziale uporządkowano kolejne etapy „cyklu życia” testu, czyli w kolejności: decyzję o jego powstaniu, konstruowanie, stosowanie, monitoring i doskonalenie go w czasie długotrwałego użytkowania. Warto podkreślić, że przedstawione dalej punkty stanowią ogólne zasady i ich stosowanie powinno uwzględniać kontekst instytucjonalny i wyzwania związane ze specyfiką danej kwalifikacji rynkowej. Założono, że będą stanowić raczej przedmiot namysłu i (auto)refleksji dla IC, a nie zestaw mechanicznie wdrażanych wskazówek.

11.1. Podejmowanie decyzji o tworzeniu testu

Decyzje o włączeniu testu wiedzy do metod walidacji w opisie kwalifikacji, a także zdefiniowanie zakresu efektów uczenia się weryfikowanych za jego pomocą powinny być podejmowane z uwzględnieniem planowania innych warunków niezbędnych do przeprowadzenia walidacji, w tym kosztów organizacji i materiałów czy wynagrodzenia dla osób konstruujących testy.

Dobór metod walidacji powinien być też adekwatny do charakteru efektów uczenia się w danej kwalifikacji. W szczególności ważne jest, aby unikać pokusy nadmiernie szerokiego stosowania testów w odniesieniu do efektów uczenia się związanych np. ze sprawnym wykonywaniem złożonych czynności. Wygoda i pozornie niskie koszty przeprowadzania testu wiedzy nie mogą uzasadniać ich nadużywania.

Krytycznej refleksji wymaga zakres stosowania testów wiedzy wykorzystujących zadania w formie otwartej oraz prowadzonych ustnie. Przyjęcie takiej formy

⁹ Ogólny przegląd najważniejszych informacji dotyczących sposobu zapewniania jakości oraz jego uwarunkowań prawnych można znaleźć w innych publikacjach przygotowanych przez IBE (np. Brzozowska i in., 2020; Dybaś, 2015).

weryfikacji, choć często pozwala zredukować inwestycje czasowe związane z przygotowaniem samego narzędzia, prowadzi do wytworzenia rozwiązań stawiających przed instytucją certyfikującą wyzwania organizacyjne (np. zapewnienie odpowiedniej liczby kompetentnych asesorów i ich wyszkolenie), finansowe (koszty pracy asesorów) oraz utrudnia uzyskanie satysfakcjonującego poziomu rzetelności walidacji.

Rekomenduje się także, aby w czasie planowania i realizacji procesu konstrukcji narzędzia oraz opracowywania wewnętrznego systemu zapewniania jakości zespół ekspertów dziedzinowych miał stałe wsparcie osób mających wiedzę i praktyczne doświadczenie z obszaru psychometrii, pomiaru edukacyjnego i statystyki. W niniejszym opracowaniu zdecydowano nie prezentować szczegółów dotyczących procedur analizy danych i statystyki. Stanowią one niezbędne elementy etapu konstrukcji testu, lecz wymagają złożonych kompetencji specjalistycznych, których w wystarczającym stopniu nie można nabyć po lekturze tej publikacji.

11.2. Konstruowanie testu

Wcześniejsze rozdziały szczegółowo omawiają wszystkie najistotniejsze standardy mające znaczenie dla jakości ostatecznie powstałego testu wiedzy i determinujące osiągnięcie satysfakcjonującego poziomu jego rzetelności i trafności. Patrząc na tę kwestię z perspektywy rozwiązań systemowych i proceduralnych, warto ponadto podkreślić trzy istotne aspekty. Po pierwsze, konstruowanie testu jest bardzo często procesem iteracyjnym, w którym – po uzyskaniu wstępnych danych – pierwotne założenia są weryfikowane i modyfikowane. Oznacza to, że etap konstrukcji trzeba planować realistycznie z zachowaniem odpowiednich rezerw czasowych i finansowych. Po drugie, zarówno trafność, jak i rzetelność pomiaru testowego mają charakter empiryczny. Nawet przy zachowaniu wysokich standardów metodologicznych takie parametry testu, jak precyzja pomiaru czy satysfakcjonująca trafność, nie są nigdy z góry zagwarantowane. Powinny być one potwierdzone na podstawie zgromadzonych danych przez zespół tworzący test. Łączy się to ściśle z ostatnim punktem: konstrukcja testu powinna zawsze uwzględniać badania pilotażowe, a następnie, po wdrożeniu zmian i przygotowaniu ostatecznej wersji, badania sprawdzające jakość weryfikacji efektów uczenia się testem.

11.3. Reguły dotyczące stosowania testów

Nawet najlepiej opracowany test działa tylko wówczas, jeżeli jest poprawnie i odpowiedzialnie stosowany. Najistotniejsze zagadnienia, które warto uwzględnić, to:

- Unikanie konfliktu interesów zachodzącego pomiędzy rolą osoby przygotowującej się do walidacji i osoby odpowiedzialnej za weryfikację. Kwestia ta nie

tylko powinna być bardzo klarownie zdefiniowana na poziomie regulacji obowiązujących w IC i IW, np. w regulaminie walidacji i WSZJ, ale również monitorowana i poddawana ewaluacji realizowanej przez zespół wewnętrznego zapewniania jakości.

- Odpowiednie przeszkolenie wszystkich osób wykorzystujących narzędzie opracowane przez IC. Newralgicznym etapem wdrożenia testów jest przekazanie go przez zespół konstruujący osobom, które będą stosować je na co dzień. Bardzo często pomiędzy tymi dwiema grupami pracowników zachodzą duże różnice w doświadczeniu i kompetencjach merytorycznych. Zespół wewnętrznego zapewniania jakości powinien opracować odpowiednie procedury i standardy szkoleń asesorów oraz osób przeprowadzających testy, a IC powinna uwzględnić je w WSZJ.
- Warto też pamiętać, że korzystanie z testu ma nie tylko aspekt pomiarowy i diagnostyczny, lecz także – bardzo istotny – wymiar etyczny. Proces szkolenia asesorów, osób przeprowadzających testy i odpowiadających za organizację w IC powinien obejmować oba aspekty.
- Reguły stosowania testu powinny też – w miarę możliwości – chronić narzędzie przed zbyt szybkim jego „zużywaniem się”, co oznacza m.in. dbałość o ochronę materiałów testowych, nieudostępnianie ich osobom postronnym czy nieupoważnionym. Jest to szczególnie istotne w przypadku stosowania narzędzia weryfikacji, opierającego się na stosunkowo niewielkiej liczbie zadań lub pytań czy wersji równoległych.
- Należy mieć świadomość, że upublicznienie treści testu, gdy jest możliwe pamięciowe opanowanie poprawnych odpowiedzi, oznacza brak możliwości zapewnienia rzetelnej walidacji przez IC, co może prowadzić do konieczności opracowania nowych wersji narzędzia.

11.4. Monitoring i doskonalenie

- Wziąwszy pod uwagę możliwości, które stwarzają współczesne rozwiązania informatyczne, rekomenduje się budowanie rozwiązań, pozwalających na łatwe monitorowanie wyników uzyskiwanych w testach, oraz reagowanie na nagłe (i niepodlegające racjonalnym wyjaśnieniom) zmiany w poziomie ich wykonania, które mogą sugerować np. upublicznienie materiałów testowych.
- Zaleca się stworzenie i wdrożenie procedur umożliwiających – zarówno osobom zaangażowanym w organizację i prowadzenie walidacji, jak i osobom podchodzącym do walidacji – zgłaszanie zaobserwowanych nieprawidłowości.
- Proponuje się okresową weryfikację aktualności stosowanych narzędzi względem stanu wiedzy, zmian w branży i praktyk rynkowych itp., do których odnosi się kwalifikacja rynkowa. Nie ma uniwersalnych reguł pozwalających określić, z jaką częstotliwością tego rodzaju ewaluacja powinna być dokonywana, ale

oczywiste jest, że np. wiedza, której oczekuje się od kompetentnego informatyka, zmienia się bez porównania częściej niż np. wymagania stawiane w tym zakresie stroicielowi fortepianów. Decyzje w tym obszarze powinni podejmować przede wszystkim eksperci dziedzinowi już na etapie tworzenia narzędzia. Odpowiednie zaprojektowanie procesu uaktualniania testu będzie gwarantować utrzymanie jego trafności w dłuższym czasie.

- Analogicznie jak w opisanym powyżej przykładzie trudno jest też wskazać ogólne reguły dotyczące częstotliwości przeprowadzania ewaluacji wewnętrznej stosowanych narzędzi testowych. Niewątpliwie im większa dynamika zmian w zakresie stanu techniki czy wiedzy i im częstsza rotacja pracowników zaangażowanych w przeprowadzanie testów, tym częściej powinien być realizowany bieżący monitoring.
- WSZJ może przewidywać, jak IC będzie postępować w sytuacjach kryzysowych, które stwarzają uzasadnione podejrzenie, że decyzje podejmowane w procesie walidacji są nierzetelne, błędne (np. odkrycie systematycznych błędów w procesie kodowania odpowiedzi testowych, duża częstotliwość łamania procedur przeprowadzania testów, pojawienie się uzasadnionego podejrzenia praktyk korupcyjnych prowadzących do „spalenia” testu).
- Ostatnim aspektem dotyczącym procedur kontrolnych jest budowanie pro jakościowej kultury w instytucjach certyfikujących. Niezależnie od wszystkich wdrażanych regulacji, przepisów sukces działań podejmowanych przez IC w bardzo dużym stopniu zależy od kształtowania właściwych postaw wśród pracowników. Odpowiedzialność ta leży przede wszystkim w gestii osób na stanowiskach zarządczych i kierowniczych. To one – zarówno swoimi działaniami, jak i decyzjami – mogą wpływać na kulturę i klimat panujące w instytucji. Te zaś mają duże znaczenie dla skuteczności i powodzenia przedsięwzięć opierających się na zaangażowaniu, rzetelności i etyce codziennej pracy osób odpowiedzialnych za planowanie, organizację i prowadzenie walidacji w IC.

Dodatkowe źródła informacji

American Educational Research Association, American Psychological Association i National Council on Measurement in Education (1999/2014). *Standards for educational and psychological testing*. Washington: American Educational Research Association. Pobrano z <https://www.testingstandards.net/#>
https://en.wikipedia.org/wiki/Standards_for_Educational_and_Psychological_Testing

Baza Dobrych Praktyk. Pobrano z <http://kwalifikacje.edu.pl/dobrepraktyki/>

Brunello, G., Schlotter, M. (2011). *Non Cognitive Skills and Personality Traits: Labour Market Relevance and their Development in Education & Training Systems, Discussion Paper No. 5743*. Bonn: The Institute for the Study of Labor (IZA). Pobrano z <http://ftp.iza.org/dp5743.pdf>

Brzozowska, A., Drzymulska-Derda, M., Dybaś-Stronkowska, M., Dymkowski, D., Hernik, K., Jędrzejczak, H., Krzemińska, S., Mrozkowiak, M. i Rycielska, L. (2020). *Rekomendacje w zakresie funkcjonowania instytucji certyfikujących w Zintegrowanym Systemie Kwalifikacji*. Warszawa: Instytut Badań Edukacyjnych. Pobrano z <https://kwalifikacje.edu.pl/wp-content/uploads/Rekomendacje-w-zakresie-funkcjonowania-IC-w-ZSK-PDF-21-MB.pdf>

Centralna Komisja Egzaminacyjna (2020). *Komunikat dyrektora CKE z 20 sierpnia 2020 r. w sprawie szczegółowych sposobów dostosowania warunków i form przeprowadzania egzaminu potwierdzającego kwalifikacje w zawodzie w roku szkolnym 2020/2021*. Pobrano z https://cke.gov.pl/images/_KOMUNIKATY/7_2021_Dostosowania_PP2012_2017.pdf

European Commission, Cedefop i ICF International (2014). *European inventory on validation of non-formal and informal learning 2014. Thematic report: validation methods*. Luxembourg: Publications Office.

Gmaj, I., Grzeszczak, J., Leyk, A., Pierwieniecka, R., Sławiński, S., Tauber, M. i Walicka, S. (2016). *Walidacja – nowe możliwości zdobywania kwalifikacji*. Warszawa: Instytut Badań Edukacyjnych. Pobrano z <https://kwalifikacje.edu.pl/walidacja-nowe-mozliwosci-zdobywania-kwalifikacji/>

Pokropek, A. (red.) (2015). *Modele cech ukrytych w badaniach edukacyjnych, psychologii i socjologii. Teoria i zastosowania*. Warszawa: Instytut Badań Edukacyjnych. Pobrano z <http://www.pwe.ibe.edu.pl/opisy/ksiazka/Modele%20cech%20ukrytych%20w%20badaniach>

Szaleniec, H., Kondratek, B., Kulon, F., Pokropek, A., Skórska, P., Świst, K., Wołodźko, T. i Żółtak, M. (2015). *Efekt egzaminatora w ocenianiu prac maturalnych z języka polskiego i matematyki*. Warszawa: Instytut Badań Edukacyjnych. Pobrano z <http://pwe.ibe.edu.pl/opisy/efektEgzaminatora/Efekt%20egzaminatora.pdf>

Katalog Metod Walidacji. Metoda testu teoretycznego. Pobrano z <http://walidacja.ibe.edu.pl/metody/pl/metody-walidacji/test-teoretyczny>

Katalog Metod Walidacji. Wskazówki dotyczące efektów uczenia się, które można sprawdzić za pomocą metody Test teoretyczny. Charakterystyki II stopnia Polskiej Ramy Kwalifikacji typowe dla kwalifikacji o charakterze zawodowym. Pobrano z: <http://walidacja.ibe.edu.pl/metody/pl/prk/97-charakterystyki-ii-stopnia-polskiej-ramy-kwalifikacji-typowe-dla-kwalifikacji-o-charakterze-zawodowym-test-teoretyczny>

Zintegrowany Rejestr Kwalifikacji: <http://www.rejestr.kwalifikacje.gov.pl>

Zintegrowany System Kwalifikacji: <http://www.kwalifikacje.gov.pl/>

Bibliografia

- Anastasi, A., Urbina, S. (1999). *Testy psychologiczne* (s. 18-270). Warszawa: Pracownia Testów Psychologicznych PTP.
- Bielecki, M., Kaczmarek, M. [niepublikowana ekspertyza]. *Konstruowanie i stosowanie testów wiedzy w celu weryfikacji efektów uczenia się*. Warszawa: Instytut Badań Edukacyjnych.
- Bloom, B. S. (1956). *Taxonomy of educational objectives. Cognitive domain* (Vol. 1, s. 201-207). New York: McKay.
- Brennan, R. L. (red.). (2006). *Educational Measurement. 4th Edition*. Westport: American Council on Education and Praeger Publishers.
- Brzozowska, A., Drzymulska-Derda, M., Dybaś-Stronkowska, M., Dymkowski, D., Hernik, K., Jędrzejczak, H., Krzemińska, S., Mrozkowiak, M. i Rycielska, L. (2020). *Rekomendacje w zakresie funkcjonowania instytucji certyfikujących w Zintegrowanym Systemie Kwalifikacji*. Warszawa: Instytut Badań Edukacyjnych.
- Cedefop; European Commission; ICF (2017). *European inventory on validation of non-formal and informal learning – 2016 update. Synthesis report*. Luxembourg: Publications Office. Pobrano z [https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32012H1222\(01\)&from=EN](https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32012H1222(01)&from=EN)
- Coombs, C. H. Dawes R. M. i Tversky A. (1977). *Wprowadzenie do psychologii matematycznej*. Warszawa: Państwowe Wydawnictwo Naukowe.
- Council of the European Union (2012). Council recommendation of 20 December 2012 on the validation of non-formal and informal learning. *Official Journal of the European Union*, C 398, 22.12.2012, 1-5.
- Dubiecka, A., Szaleniec, H., Węziak, D. (2006). *Efekt egzaminatora*. W: Publikacja pokonferencyjna: XII Konferencja diagnostyki edukacyjnej – o wyższą jakość egzaminów szkolnych (s. 98-115). Lublin: Polskie Towarzystwo Diagnostyki Edukacyjnej.
- Dybaś, M. (red.). (2015). *Zapewnianie jakości procesu nadawania kwalifikacji spoza systemów oświaty i szkolnictwa wyższego. Wskazówki do projektowania, organizowania i przeprowadzania walidacji oraz wewnętrznego zapewniania jakości*. Warszawa: Instytut Badań Edukacyjnych.
- Gmaj, I., Grzeszczak, J., Leyk, A. i in. (2016). *Walidacja – nowe możliwości zdobywania kwalifikacji*. Warszawa: Instytut Badań Edukacyjnych.
- Gmaj, I., Grzeszczak, J., Leyk, A., Pierwieniecka, R., Tauber, M., Walicka, S. (2017). *Zagraniczne systemy walidacji. Przegląd rozwiązań*. Warszawa: Instytut Badań Edukacyjnych. Pobrano z <https://kwalifikacje.gov.pl/images/Publikacje/zagraniczne-systemy-walidacji.pdf>

- Holland, P. W., Hoskens, H. (2002). *Classical Test Theory as a First-Order Item Response Theory: Application to True-Score Prediction From a Possibly Nonparallel Test*. Princeton: Research Publications Office, Educational Testing Service.
- Hornowska, E. (2001). *Testy psychologiczne. Teoria i praktyka*. Warszawa: Wydawnictwo Naukowe SCHOLAR.
- Jakubowski, M., Pokropek, A. (2009). *Badając egzaminy. Podejście ilościowe w badaniach edukacyjnych*. Warszawa: Centralna Komisja Egzaminacyjna.
- Livingston, S. A. (2004). *Equating test scores (without IRT)*. Princeton: Educational Testing Service.
- Majkut, P., Koniewski, M., Skórska, P. i Jakubowski, M. (2016). *Strategie badań nad pozapoznawczymi umiejętnościami uczniów na przykładzie Public Education Evaluation System, realizowanym przez Public Education Evaluation Commission w Arabii Saudyjskiej*. Konferencja Polskiego Towarzystwa Diagnostyki Edukacyjnej 2016.
- Marzano, R. J. (2001). *Designing a New Taxonomy of Educational Objectives. Experts in Assessment*. Thousand Oaks: CA Corwin Press, Inc.
- Pokropek, A. (red.). (2015). *Modele cech ukrytych w badaniach edukacyjnych, psychologii i socjologii. Teoria i zastosowania*. Warszawa: Instytut Badań Edukacyjnych.
- Saal, F. E., Downey, R. G. i Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, 88(2), 413–428.
- Scullen, S. E., Mount, M. K. i Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology*, 85(6), 956–970.
- Sławiński, S. (2016). *Słownik Zintegrowanego Systemu Kwalifikacji*. Warszawa: Instytut Badań Edukacyjnych.
- Shohamy, E. (1992). Beyond proficiency testing: A diagnostic feedback testing model for assessing foreign language learning. *Modern Language Journal*, 76(4), 513–521.
- Shohamy, E. (1993). *The power of tests: The impact of language tests on teaching and learning*. NFLC Occasional Paper. Washington, DC: National Foreign Language Center.
- Shohamy, E., Donitsa-Schmidt, S. i Ferman, I. (1996). Test impact revisited: Washback effect over time. *Language Testing*, 13(3), 298–317.
- Ustawa z dnia 22 grudnia 2015 r. o Zintegrowanym Systemie Kwalifikacji (t.j. Dz.U. z 2020, poz. 226).
- Wojciszke, B. (2000). Psychologia oceniania: mechanizmy, pułapki, środki zaradcze. W: A. Brzezińska, J. Brzeziński (red.), *Ewaluacja procesu kształcenia w szkole wyższej* (s. 211–237). Poznań: Wydawnictwo Fundacji Humaniora.

Ziewiec-Skokowska, G., Danowska-Florczyk, E. i Stęchły, W. (2016). *Opisywanie kwalifikacji nadawanych poza systemami oświaty i szkolnictwa wyższego. Poradnik*. Warszawa: Instytut Badań Edukacyjnych.

Autorzy

Maksymilian Bielecki – doktor n. humanistycznych, psycholog. Adiunkt w katedrze Neuronauki Poznawczej i Behawioralnej oraz wykładowca studiów podyplomowych SWPS Uniwersytetu Humanistycznospołecznego. Jako statystyk i psycholog współpracuje z wiodącymi instytucjami naukowymi, m.in. Warszawskim Uniwersytetem Medycznym i Instytutem Biologii Doświadczalnej im. M. Nenckiego. Naukowo interesuje się przede wszystkim problematyką pamięci, uwagi oraz eye trackingiem.

Jako konsultant w zakresie ekonomii behawioralnej, statystyki i psychometrii współpracował z wieloma firmami działającymi w obszarach HR, FMCG, finansów, ubezpieczeń i telekomunikacji. Jest współzałożycielem Centrum Innowacji Społecznych i Technologicznych HumanTech, działającego na Uniwersytecie SWPS.

Magdalena Kaczmarek – doktor n. humanistycznych, psycholog. Pracuje w SWPS Uniwersytecie Humanistycznospołecznym. Od kilkunastu lat zajmuje się psychologią różnic indywidualnych, diagnozą psychologiczną i pomiarem psychologicznym. Brała udział w licznych projektach naukowych oraz praktycznych, opracowując testy i kwestionariusze psychologiczne, w tym narzędzia do diagnozy poznawczej, cech osobowości oraz liczne kwestionariusze postaw. Zajmuje się także popularyzacją psychologii. Jest współzałożycielką Centrum Innowacji Społecznych i Technologicznych HumanTech, działającego na Uniwersytecie SWPS.

Ludmiła Rycielska – psycholog, od 2012 roku związana z Instytutem Badań Edukacyjnych w Warszawie, specjalistka w zakresie metodologii i analizy danych ilościowych. Jako ekspert ds. Zintegrowanego Systemu Kwalifikacji projektuje walidacje i wewnętrzne systemy zapewniania jakości dla instytucji certyfikujących i podmiotów zainteresowanych uzyskaniem uprawnień do certyfikowania. Posiada doświadczenie w tworzeniu i adaptacji kulturowej kwestionariuszy psychologicznych, układaniu pytań sondażowych i do dużych badań edukacyjnych, analizie psychometrycznej testów stosowanych w pomiarze edukacyjnym z wykorzystaniem IRT i KTT. Współtworzyła i testowała zadania na tablety dla dzieci 5-letnich na zlecenie Australian Council for Educational Research. Autorka artykułów naukowych i współautorka raportów z krajowych i międzynarodowych badań edukacyjnych. Współautorka *Rekomendacji z zakresu funkcjonowania instytucji certyfikujących w ZSK*.

Publikacja zawiera podstawowe zasady projektowania oraz stosowania pisemnych testów wiedzy do weryfikacji efektów uczenia się, które są wymagane dla danej kwalifikacji rynkowej. Przedstawia najważniejsze standardy tworzenia i późniejszego stosowania testów, a także wybrane problemy etyczne związane z testowaniem. Omówiono w niej także rodzaje testów ze wskazaniem ich najważniejszych wad i zalet, zasady konstruowania liczbowych wskaźników na podstawie wyniku testu oraz sposoby dokonywania oceny, czy dany efekt uczenia się został lub nie pozytywnie zweryfikowany.

Opracowanie jest skierowane do osób projektujących walidację, tworzących testy wiedzy na potrzeby weryfikacji oraz asesorów i osób odpowiadających za przebieg walidacji w uprawnionych do tego instytucjach certyfikujących, a także specjalistów opisujących kwalifikacje rynkowe.

ISBN 978-83-66612-64-8

Instytut Badań Edukacyjnych

ul. Górczewska 8

01-180 Warszawa

tel. +48 22 241 71 00

www.ibe.edu.pl