Edukacja 2012, 4(120), 52-71
ISSN 0239-6858

Zrównywanie wyników testowania. Definicje i przykłady zastosowania

Artur Pokropek*, Bartosz Kondratek*

Dojrzałe systemy testowania oraz większość nowopowstałych zawierają mechanizmy pozwalające na zrównywanie wyników z różnych sesji testowych w celu kontrolowania różnic w poziomie trudności różnych wersji testu. Artykuł przedstawia definicje zrównywania wyników wraz z przeglądem podstawowych planów zbierania danych stosowanych przy zrównywaniu. W celu ukazania podstawowych trendów w metodologii zrównywania testów na świecie przedstawiono 11 przykładowych systemów testowania, w których przeprowadzanie zrównywania jest wpisane w proces konstrukcji i raportowania wyników testu. Każdy test pokrótce omówiono i wskazano mechanizmy umożliwiające zrównywanie. Przegląd testów podzielono na trzy części w zależności od zastosowań badania testowego: narodowe systemy egzaminacyjne (SAT, ACT, PET, SweSAT), międzynarodowe systemy ewaluacyjne (TIMMS, PIRLS, PISA) oraz narodowe systemy ewaluacyjne (NAEP, EQAO, NAPLAN, NABC).

Słowa kluczowe: Zrównywanie wyników, plany zrównywania, badanie umiejętności

Definicja zrównywania

Zrównywanie testów polega na ustaleniu odpowiedniości, która pozwala na zamienne, równoważne posługiwanie się ich wynikami. Procedura ta dotyczy testów mierzących ten sam konstrukt i tworzonych zgodnie z tymi samymi specyfikacjami (*blueprint*). Konieczność zrównywania testów jest konsekwencją praktycznej niemożności stworzenia dwóch testów, które byłyby ściśle równoległe[1] (Holland, Dorans i Petersen, 2007). Za Aliną von Davier (2011, s. 1–2):

> Zrównywanie jest konieczne tylko ze względu na to, że standaryzowany pomiar edukacyjny korzysta z wielu form testów, które różnią się trudnością, mimo że są tworzone zgodnie z tymi samymi specyfikacjami […]. Zrównywanie można postrzegać jako procedurę statystycznej kontroli zmiennej zakłócającej (*confounding variable*), za którą przyjmuje się właśnie formę testu. Gdyby proces tworzenia testu był idealny, nie byłoby potrzeby zrównywania.

Aby łączenie wyników testowych (*linking*) mogło być uznane za zrównywanie (*equating*), musi zostać spełnionych wiele restrykcyjnych założeń. Można je znaleźć w podobnej formie u wielu autorów (np.

[1] Dwa testy są ściśle równoległe (*strictly parallel*), jeżeli każda osoba z populacji będzie miała taką samą wariancję błędu pomiaru oraz taki sam wynik prawdziwy w obu testach. Innymi słowy: dwa testy ściśle równoległe są w zupełności sobie równoważnymi (*perfectly equivalent, interchangable*) narzędziami pomiarowymi (Gruijter i van der Kamp, 2005).

Badanie wykonane w ramach projektu systemowego „Badanie jakości i efektywności edukacji oraz instytucjonalizacja zaplecza badawczego" realizowanego przez Instytut Badań Edukacyjnych i współfinansowanego ze środków Europejskiego Funduszu Społecznego (Program Operacyjny Kapitał Ludzki 2007–2013, priorytet III: Wysoka jakość systemu oświaty).

* Instytut Badań Edukacyjnych. Email: a.pokropek@ibe.edu.pl, b.kondratek@ibe.edu.pl

Kolen i Brennan, 2004; Lord, 1980); poniżej wymienimy je za Neilem Doranem i Paulem Hollandem (2000, s. 282–283):

> (a) tożsamy konstrukt (*equal construct*): testy mierzące różne konstrukty nie powinny być zrównywane
> (b) równa rzetelność (*equal reliability*): testy mierzące ten sam konstrukt, ale różniące się rzetelnością, nie powinny być zrównywane
> (c) symetria (*symmetry*): funkcja zrównująca wyniki w teście *Y* z wynikami w teście *X* powinna być odwrotnością funkcji zrównującej wyniki w teście *X* z wynikami w teście *Y*
> (d) równoważność (*equity)*: dla osoby rozwiązującej test nie powinno mieć żadnego znaczenia, którą z wersji testu rozwiązuje, gdy testy są zrównane
> (e) niezmienniczość względem populacji (*population invariance*): wybór (sub)populacji użytej do obliczenia funkcji zrównującej wyniki w testach *X* oraz *Y* nie powinien mieć znaczenia, tj. funkcja zrównująca używana do łączenia wyników w testach *X* oraz *Y* powinna być identyczna niezależnie od (sub)populacji rozwiązujących test *X* oraz *Y*.

W celu bliższego wytłumaczenia tych pięciu wymogów można się odwołać do komentarza zawartego w artykule *Equating test scores* (Holland i in., 2007). Wymogi (a) oraz (b) oznaczają, że zrównywane testy powinny być skonstruowane tak, aby były zgodne ze sobą pod względem treści oraz właściwości statystycznych. Wymóg (c) wyklucza możliwość zastosowania metod regresji do zrównywania testów. Wymóg (d) poniekąd tłumaczy konieczność wymogu (a) – jeżeli testy mierzyłyby różne konstrukty, to osoby mające je rozwiązywać preferowałyby podejście do tego testu, w którym – w ich mniemaniu – miałyby szansę uzyskać lepszy wynik. Wymóg (e) można wykorzystać do uzasadnienia wymogów (a) oraz (b). Jeżeli testy byłyby tworzone zgodnie z różnymi wzorcowymi specyfikacjami, to funkcja zrównująca wyniki mogłaby się różnić w zależności od wyboru subpopulacji. Przykładowo, łącząc wyniki testu mierzącego umiejętność rozumowania na podstawie materiału niewerbalnego z wynikami testu mierzącego umiejętność rozumowania na podstawie materiału werbalnego, zapewne uzyskano by różne funkcje łączące w zależności od płci badanych.

Pewnego dopowiedzenia wymaga wymóg równoważności (d), gdyż formalnie pojawia się on w dwóch, istotnie różniących się wersjach (Kolen i Brennan, 2004):

(1) $\forall_\tau \mathbb{P}(eq_Y(X) \leq y|\tau) = \mathbb{P}(Y \leq y|\tau)$

(2) $\forall_\tau \mathbb{E}(eq_Y(X)|\tau) = \mathbb{E}(Y|\tau)$

gdzie $eq_Y: X \rightarrow Y$ jest funkcją zrównującą test *X* z *Y*

Pierwsza wersja równoważności (*equity*) stanowi, że dla każdego wyniku prawdziwego τ warunkowy względem τ rozkład wyników otrzymywanych w teście Y jest taki sam, jak w zrównanym z nim teście X. Natomiast druga wersja równoważności osłabia wymóg warunkowej równości dwóch rozkładów do warunkowej równości jedynie względem pierwszego momentu zwykłego (czyli wartości średniej) tych rozkładów. W szczególności wersja (2) nie wymaga równości między warunkowymi wariancjami, czyli równości warunkowego błędu pomiaru. Pierwsza wersja (1), sformułowana przez Frederica Lorda (1980), jest bardzo restrykcyjna i spotyka się ze słusznym komentarzem Michaela Kolena i Roberta Brennana (2004): „korzystanie z równoważności Lorda jako kryterium oznacza, że zrównywanie albo nie jest możliwe, albo nie jest potrzebne" (zob. też: van der Linden, 2011, jak i sam Lord, 1980). Większość metod zrównywania wyników stawia sobie wprost za cel spełnienie słabszej formy równoważności (*weak equity/first-order equity*).

Pogłębioną refleksję nad równoważnością w silnym sformułowaniu Lorda (1) mo-

żemy znaleźć u Wima van der Lindena (2011), który zwracając uwagę na lokalny (tj. zdefiniowany w zależności od wyniku prawdziwego τ) charakter równania, proponuje zrównywanie oparte na konstrukcji lokalnych funkcji zrównujących. Pojawia się tu ścisła zależność pomiędzy wymogiem równoważności (d), a wymogiem niezmienniczości względem populacji (e). Mimo iż zależność jest taka, że (e) implikuje (d), van der Linden sugeruje, że przybliżanie się do spełnienia wymogu niezmienniczości względem populacji również przybliża spełnienie wymogu równoważności. Ujęcie van der Lindena wskazuje też istotę problemu, czyli to, że pomiar edukacyjny jest obarczony błędem, co umyka w nielokalnych modelach zrównywania wyników. Zignorowanie tego faktu przy stosowaniu pojedynczej funkcji zrównującej $eq_Y(x)$ prowadzi do lokalnego obciążenia. Niestety, wydaje się, że lokalne funkcje zrównujące wyniki obserwowane nie mogłyby zostać praktycznie wykorzystane do raportowania zależności między zrównywanymi testami – np. przy zastosowaniu IRT oznaczałoby to różne przekształcenia na test Y dla osób o tym samym wyniku w teście X, jeżeli różniłyby się ich oszacowania θ. Niemniej koncepcja lokalnych funkcji zrównujących i związek pomiędzy wymogiem równoważności a niezmienniczością względem populacji dostarczają ważnych narzędzi empirycznej weryfikacji spełnienia założenia o równoważności.

Jinghua Liu i Michael Walker (2007, s. 115), dokonując przeglądu wymogów stawianych procedurze zrównywania testów przez Lorda, Doransa i Hollanda oraz Kolena i Brennana, zdecydowali się na wyszczególnienie dodatkowych trzech wymogów na podstawie pracy tych ostatnich:

> (f) takie same inferencje (*the same inferences*): testy powinny mieć wspólne cele pomiarowe i powinny być zaprojektowane do wyciągania takiego samego typu wniosków
>
> (g) taka sama populacja docelowa (*the same target population*)
>
> (h) takie same charakterystyki/warunki pomiarowe (*the same measurement characteristics/conditions*): testy powinny mieć taką samą specyfikację, być przeprowadzane w takich samych warunkach oraz być równoważne pod względem właściwości psychometrycznych.

Zauważalna jest pewna redundancja zbioru wszystkich ośmiu wymienionych wymogów niezbędnych do przeprowadzenia zrównywania wyników testowych. Wydaje się jednak, że sformułowanie wszystkich wymogów explicite daje jaśniejszy obraz tego, czym jest zrównywanie wyników w teorii. Natomiast w praktyce niektóre z wymogów mogą być trudne do weryfikacji, np. wymóg (d). W kwestii wagi poszczególnych wymogów toczy się dyskusja, którą w skrócie omawia w swojej pracy zespół Hollanda (2007). Natomiast w kwestii praktycznej weryfikacji wymienionych wymogów warto odwołać się do Liu i Walkera (2007), którzy zastosowali interesujący zestaw kryteriów zrównywalności (*equatability*) Scholastic Assessment Test (SAT) wersji funkcjonującej do 2004 roku z nową wersją, która funkcjonuje od 2005 roku. Znamienne jest, że zadanie zrównywania wyników zostało podjęte w obliczu znacznej zmiany w zakresie wzorcowych specyfikacji testu, co przy konserwatywnym traktowaniu wszystkich wymogów stawianych przed zrównywaniem mogłoby zostać uznane za argument dyskwalifikujący możliwość dokonania zrównania. Zaproponowane przez nich kryteria zrównywalności były następujące:

- podobieństwo konstruktu (*construct similiarity*): weryfikowane zarówno przez stopień podobieństwa treści, jak i statystyczne właściwości testu
- empiryczna relacja pomiędzy nowym i starym testem: weryfikowana przez współczynnik korelacji między dwoma

testami w odniesieniu do współczynnika rzetelności każdego z testów (wyznaczającego górną granicę dla takiej korelacji)
- precyzja pomiaru: weryfikowana zarówno poprzez współczynnik rzetelności, jak i przez lokalne miary błędu pomiaru umiejętności
- niezmienniczość w podgrupach (*subgroup invariance*): weryfikowana przez relację między średnimi wynikami w zależności od istotnych zmiennych grupujących oraz przez analizę postaci funkcji łączącej wyniki w zależności od istotnych zmiennych grupujących.

Zasadniczym problemem, jaki zrównywanie wyników musi rozwiązać, jest rozdzielenie efektu trudności testu od efektu umiejętności uczniów wykonujących test. Są dwa podstawowe sposoby rozdzielenia tych dwóch efektów dla umiejętności uczniów zdających różne formy testu:

- Plany wykorzystujące „wspólne osoby" (*common examinees, common persons*), gdy próba złożona z tych samych osób rozwiązuje zrównywane testy lub zrównywane testy są rozwiązywane przez osoby należące do losowo równoważnych prób
- Plany wykorzystujące „wspólne zadania" (*common items*), gdy różne próby osób rozwiązują jednocześnie zbiory takich samych zadań testowych.

Plany zrównywania

Aby testy mogły zostać zrównane, trzeba przyjąć plan zrównywania. Niemal w każdym podręczniku zrównywania wyników testowych można znaleźć opis czterech planów: (a) plan grup równoważnych, (b) plan pojedynczej grupy, (c) plan zrównoważony, (d) plan nierównoważnych grup z testem kotwiczącym. Pierwsze trzy należą do kategorii *common examinees*, ostatni do *common items*. Przedstawiony poniżej opis planów zrównywania opiera się na pracy zespołu Aliny von Davier (2004), aczkolwiek bardzo zbliżone opisy można też znaleźć w pracach Kolena i Brennana (2004), Kolena (2007) i Samuela Livingstona (2004).

Plan grup równoważnych (*equivalent groups design, EG*) opiera się na dwóch założeniach:

- Istnieje pojedyncza populacja osób $\mathcal{P}$, które mogą rozwiązać każdy z testów *X* oraz *Y*.
- Z populacji $\mathcal{P}$ dobierane są dwie niezależne próby losowe; osoby z jednej próby rozwiązują test *X,* z drugiej rozwiązują test *Y*.

Schematycznie plan EG można przedstawić w następujący sposób:

Populacja	Próba	Test *X*	Test *Y*
$\mathcal{P}$	S_1	✓	
$\mathcal{P}$	S_2		✓

Losowanie prób S_1 oraz S_2, technicznie rzecz ujmując, zazwyczaj nie jest doborem prostym losowym, ale odbywa się poprzez tzw. spiralne rozdawanie dwóch testów (*spiraled sampling*) lub losowanie grupowe np. szkół lub oddziałów klasowych (Kolen, 2007). Dyskusję porównującą spiralne rozdawanie testów z prostą próbą losową można znaleźć w publikacji von Davier, Hollanda i Thayera (2004).

Plan pojedynczej grupy (*single group design*, SG) opiera się na dwóch założeniach:

- Istnieje pojedyncza populacja osób $\mathcal{P}$, które mogą rozwiązać oba testy *X,* oraz *Y*.
- Z populacji $\mathcal{P}$ dobierana jest jedna próba losowa; wszystkie badane osoby rozwiązują najpierw jeden, potem drugi test.

Przyjmując oznaczenie X^{I} oraz Y^{II} w celu wskazania, że test *X* jest przeprowadzany jako pierwszy, a test *Y* jako drugi, plan SG można przedstawić w następujący sposób:

Populacja	Próba	X^I	Y^{II}
$\mathcal{P}$	S_1	✓	✓

Przewaga planu SG nad planem EG polega na korzystaniu z powtarzanych pomiarów, co potencjalnie zwiększa moc statystyczną procedury zrównywania (Livingston, 2004), jeżeli tylko procedura zrównywania korzysta z zebranej informacji o korelacji między testami *X* oraz *Y*. Nieodłączną konsekwencją wprowadzenia powtarzanych pomiarów jest niebezpieczeństwo występowania istotnego efektu kolejności, który w planie SG nie jest kontrolowany.

Plan zrównoważony (*counterbalanced design*, CB) stanowi odpowiedź na potrzebę kontroli efektu kolejności pisania testu w planie SG; opiera się na dwóch założeniach:

- Istnieje pojedyncza populacja osób $\mathcal{P}$, które mogą rozwiązać oba testy *X* oraz *Y* w dowolnej kolejności.
- Z populacji $\mathcal{P}$ dobierane są dwie niezależne próby losowe: osoby z jednej próby rozwiązują najpierw test *X*, potem test *Y*, osoby z drugiej próby rozwiązują testy w odwrotnej kolejności.

Populacja	Próba	X^I	Y^I	X^{II}	Y^{II}
$\mathcal{P}$	S_1	✓			✓
$\mathcal{P}$	S_2		✓	✓	

Można zauważyć, że plan CB zawiera w sobie dwa plany SG ($X^I - Y^{II}$ i $X^{II} - Y^I$) oraz dwa plany EG ($X^I - Y^I$ i $X^{II} - Y^{II}$) – ma to odzwierciedlenie w metodach zrównywania wyników w tym planie, które mogą się różnić sposobem wykorzystywania każdego z tych zawartych wewnątrz CB planów (von Davier, Holland i Thayer, 2004).

Plan nierównoważnych grup z testem kotwiczącym (*nonequivalent groups with anchor test design*, NEAT) opiera się na dwóch założeniach:

- Istnieją dwie populacje osób: $\mathcal{P}$ oraz $\mathcal{Q}$, które mogą rozwiązać odpowiednio testy: *X* oraz *Y*, ponadto wszystkie osoby mogą rozwiązywać kotwicę *A*.
- Dwie próby losowe są dobierane niezależnie – jedna z $\mathcal{P}$, druga z $\mathcal{Q}$.

Populacja	Próba	*X*	*Y*	*A*
$\mathcal{P}$	S_1	✓		✓
$\mathcal{Q}$	S_2		✓	✓

Można zauważyć, że plan NEAT zawiera w sobie dwa plany SG ($X - A$ i $Y - A$).

Plan NEAT można formalnie podzielić w zależności od tego, czy zbiór zadań wchodzących w skład testu *A* jest odrębnym testem od *X* oraz *Y* (kotwica zewnętrzna – *external anchor*), czy też *A* stanowi podzbiór zadań testów *X* oraz *Y*, które są oceniane jako element składowy wyników w tych testach (kotwica wewnętrzna – *internal anchor*).

Klasyfikacja metod zrównywania wyników

Po wyborze planu zrównywania musi dojść do wyboru metod zrównywania. Na najogólniejszym poziomie można dokonać podziału metod zrównywania wyników testowych w zależności od tego czy:

- zrównywanie odbywa się na skali wyników obserwowanych, czy wyników prawdziwych
- zrównywanie odbywa się z bezpośrednim odwołaniem do modelu pomiarowego, czy nie.

Większość technik wykorzystywanych do zrównywania wyników testowych należy do kategorii zrównywania wyników obserwowanych (*observed score equating*), gdzie przez

wynik obserwowany rozumie się klasyczny sumaryczny wynik w teście. Nacisk na przeprowadzanie zrównywania na poziomie wyników obserwowanych jest konsekwencją tego, że w przeważającej większości takie właśnie wyniki są wykorzystywane do raportowania rezultatów testowania. Zrównywanie wyników obserwowanych może zostać przeprowadzone bez konieczności odwoływania się w modelu statystycznym do sparametryzowanego mechanizmu leżącego u podstaw obserwowanych wyników, ale także z wykorzystaniem takiego modelu, tj. z wykorzystaniem IRT (*IRT observed score equating*).

W obrębie podejścia opartego na modelach IRT pojawia się możliwość zrównywania wyników prawdziwych (*IRT true score equating*). Przez wynik prawdziwy danego ucznia rozumie się tu wartość oczekiwaną z wyniku obserwowanego tego ucznia. Aby zrównanie zostało przeprowadzone na skali wyników prawdziwych klasycznej teorii testów, konieczne jest oszacowanie parametrów modelu pomiarowego leżącego u podstaw obserwowanych odpowiedzi. Zrównywania na skali wyników prawdziwych nie można zatem przeprowadzić „ateoretycznie", jak w wypadku wyników obserwowanych. Omawiane zależności między metodami zrównywania wyników schematycznie przedstawione są w Tabeli 1.

Po przeglądzie teoretycznych aspektów zrównywania przejdźmy do jego praktycznych zastosowań w systemach egzaminacyjnych oraz międzynarodowych i narodowych systemach ewaluacyjnych.

Zrównywanie w wybranych systemach egzaminacyjnych[2]

Stany Zjednoczone są pionierem w dziedzinie nowoczesnych technik testowania, dlatego w niniejszym przeglądzie pojawiają się one na pierwszym miejscu. Rozwiązania z USA przedstawione zostaną na przykładzie dwóch najstarszych amerykańskich testów rozwiązywanych przede wszystkim przez uczniów po 12. roku nauki. W obydwu wypadkach są to testy wysokiej stawki, których wyniki brane są pod uwagę przy rekrutacji na uczelnie wyższe. Omawiane testy wykorzystują dwa różne schematy zrównywania, które stanowią wzór dla innych testów przedstawianych w kolejnych częściach tego artykułu.

Scholastic Assessment Test (SAT)

Scholastic Assessment Test (SAT) to najstarszy, funkcjonujący po dziś dzień (z pewnymi zmianami) test osiągnięć szkolnych na świecie. Powstał w 1926 r. na zlecenie College Board, organizacji zrzeszającej uczelnie wyższe oraz inne organizacje edukacyjne. Pierwszy test, przeprowadzony w 1926 r., trwał 90 minut i składał się z 315 pytań mierzących znajomość słownictwa oraz podstawo-

[2] Część opisów testów wykorzystujących metodologie zrównywania (SAT, ACT, PET, SweSAT, EQAO, NAPLAN) powstała na podstawie: Pokropek (2011).

Tabela 1
Schematyczny podział metod zrównywania wyników testowych

	Zrównywanie wyników obserwowanych	Zrównywanie wyników prawdziwych
Metody niezależne od modelu pomiarowego	*(non-IRT) observed score equating*	–
Metody oparte na modelu pomiarowym	*IRT observed score equating*	*IRT true score equating*

we umiejętności matematyczne. W kolejnych latach test przechodził szereg zmian, żadna z nich nie była jednak zmianą fundamentalną. Zwiększano i zmniejszano liczbę pytań, eksperymentowano z różnymi rodzajami zadań i wprowadzano nowe dziedziny wiedzy do pomiaru (Lawrence, Rigol, van Essen i Jackson, 2002). Ostatnie znaczące zmiany wprowadzone zostały w 2005 roku i przy ich okazji przeprowadzono również bardzo interesujące badania nad zrównywalnością zmienionego testu (Liu i Walker, 2007). Obecnie test składa się z 9 sekcji testowych i jednej sekcji zrównującej (eksperymentalnej), a łączny czas testowania to 3 godziny i 45 minut. Trzy sekcje mierzą umiejętność czytania ze zrozumieniem (67 pytań). Kolejne trzy – umiejętności matematyczne (54 pytań), a następne – umiejętności wypowiedzi pisemnej (49 pytań). Sekcja zrównująca[3] w całości poświęcona jest jednej dziedzinie wiedzy (czytanie ze zrozumieniem, pisanie lub matematyka) i jest skonstruowana tak, by uczniowie nie wiedzieli, która sekcja należy do części zrównującej.

W SAT wynik ucznia określa się na podstawie 170 pytań z sekcji testowej. Odpowiedzi na zadania sekcji zrównującej nie są brane pod uwagę przy szacowaniu końcowego wyniku ucznia. Za każdą poprawną odpowiedź uczniowie zdobywają jeden punkt, za błędną odpowiedź w zadaniach zamkniętych – cząstkowe punkty ujemne: –1/4 w zadaniach z czterema możliwościami wyboru, –1/3 w zadaniach z trzema możliwościami wyboru i –1/2 w zadaniach z dwoma możliwościami wyboru. Wyniki są skalowane i zrównywane metodą ekwicentylową, a następnie przedstawiane na jednej zagregowanej skali z przedziału 600–2400 punktów oraz na trzech osobnych skalach: dla czytania ze zrozumieniem, matematyki i pisania z przedziału 200–800. Osobnej ocenie podlegają eseje znajdujące się w części mierzącej umiejętność pisania. Warto podkreślić, iż zarówno w SAT, jak i w ACT (test omawiany w drugiej kolejności) każdy esej sprawdzany jest niezależnie przez dwóch egzaminatorów oceniających go na skali od 1 do 6 punktów. Sumaryczny wynik testu pisania zawiera się zatem w przedziale od 2 do 12 punktów.

W początkowych latach istnienia SAT nie podejmowano prób zrównywania wyników. Sytuacja ta zmieniła się w 1941 r. Odtąd każda nowa wersja testu zawierała około 20% pytań z poprzedniej edycji. Wyniki kolejnych edycji zrównywane były rok do roku. Średnią skali ustalono na 500 punktów dla 1941 r. (w 1995 r. skala została ponownie wycentrowana, tak by rokiem bazowym był rok 1995 o średniej 500). W kolejnych latach procedura zrównywania ewoluowała, choć do dzisiaj stosuje się schemat zrównywania dla planu nierównoważnych grup z testem kotwiczącym (NEAT).

Do zrównywania używa się klasycznych metod zrównywania liniowego i nieliniowego (*Tucker, Levine observed score, chained linear oraz chained equipercentile*). Wybór metody zależy od psychometrycznych właściwości testów, które mają zostać zrównane.

American College Testing (ACT)

ACT jest drugim (po SAT) najpopularniejszym testem mierzącym osiągnięcia uczniów w szkole średniej. Pierwszy raz został przeprowadzony w 1959 r., a skonstruował go – w odpowiedzi na test SAT –wybitny teoretyk pomiaru Everett Franklin Lindquist. ACT do 2005 r. mierzył 4 dziedziny wiedzy: umiejętność posługiwania się językiem angielskim, znajomość matematyki, czytanie ze zrozumieniem oraz rozumowanie w naukach przyrodniczych. Wśród teoretyków pomiaru panuje przekonanie, iż zadania w teście ACT są łatwiejsze niż w SAT, lecz czasu na

[3] Dokładniej – sekcje zrównujące, gdyż w jednej edycji testu stosuje się kilka różnych sekcji zrównujących. Dla prostoty wywodu dalej zakładamy jednak, iż sekcja jest jedna.

ich rozwiązanie jest znacznie mniej. Uczniowie mają 45 minut na rozwiązanie 75 zadań z języka angielskiego, 60 minut na 60 pytań z matematyki, 35 minut na 40 zadań mierzących umiejętność czytania, 35 minut na poradzenie sobie z 40 zadaniami z sekcji przyrodniczej oraz 30 minut na napisanie eseju. Łącznie uczeń rozwiązuje test ACT przez 3 godziny i 25 minut.

Każde zadanie w teście punktowane jest na skali 0–1, zatem każde zadanie ma taką samą wagę przy szacowaniu skali wyników. W przeciwieństwie do SAT nie ma też punktów ujemnych. Zrównywanie odbywa się metodą ekwicentylową. Wyskalowane wyniki testu przedstawiane są na skali od 1 do 36 punktów, gdzie punkty są liczbami całkowitymi. Publikowane są również wyniki w podskalach: angielski, matematyka, czytanie ze zrozumieniem oraz rozumowanie w naukach przyrodniczych. Wyniki z poszczególnych przedmiotów przedstawiane są na skali od 1 do 18. Wynik całościowy jest średnią z czterech podtestów. Test pisania nie jest obowiązkowy i nie liczy się do sumarycznego wyniku. Uczniowie, którzy decydują się na test mierzący umiejętność pisania, otrzymują wynik na skali 2–12 oraz od 1 do 4 komentarzy.

Zrównywanie w teście ACT odbywa się na podstawie schematu z równoważnymi grupami. Aby przeprowadzić zrównanie dwóch testów z różnych lat, spośród wszystkich uczniów losowana jest reprezentatywna próba losowa. Uczniowie należący do wylosowanej próby, oprócz aktualnej edycji testu, rozwiązują też kilka nowych, wcześniej niepublikowanych arkuszy egzaminacyjnych. Jako że w populacji zrównującej uczniowie rozwiązywali zadania z testu właśnie przeprowadzonego oraz zadania z testów, które dopiero mają się odbyć w kolejnych sesjach, możliwe jest zrównanie wyników z testu już przeprowadzonego z kolejnymi edycjami. W teście ACT do zrównywania używana jest metoda ekwicentylowa wykorzystująca analityczne metody wygładzania rozkładów (Kolen, 1984; ACT, 2007).

Psychometric Entrance Test (PET)

W 1981 r. w Izraelu został powołany Narodowy Instytut Testowania. Jego zadaniem było stworzenie ogólnonarodowego standaryzowanego testu, którego wynik byłby brany pod uwagę przy rekrutacji na uczelnie wyższe. Efektem prac tej instytucji jest test Psychometric Entrance Test (PET).

PET ma mierzyć kognitywne oraz szkolne zdolności będące predyktorami sukcesu w karierze akademickiej. Od roku 1990 PET składa się z trzech sekcji: rozumowania werbalnego (*verbal reasoning*), rozumowania ilościowego (*quantitative reasoning*) oraz sekcji badającej znajomość języka angielskiego (Beller, 1994). Części mierzące rozumowanie są częściowo podobne do testów inteligencji. W przypadku sekcji werbalnej zdający rozpoznają antonimy i analogie, odczytują wyrazy z zakrytymi literami. W części matematycznej testu uczniowie muszą poradzić sobie z różnorodnymi problemami matematycznymi i odczytywaniem danych zaprezentowanych w różny sposób. Test nie wymaga znajomości programu matematyki ze szkoły średniej, odwołuje się tylko do podstawowych pojęć matematycznych. W teście z języka angielskiego dominującą rolę odgrywa czytanie ze zrozumieniem tekstów akademickich (Beller, 1994).

Należy dodać, iż test PET jest w zasadzie testem szybkości, gdyż na rozwiązanie jednego zadania z części rozumowania werbalnego zdający ma około 50 sekund, a na zadania dotyczące rozumowania ilościowego – 60 sekund. PET jest podzielony na 8 sekcji, trwa 3 godziny i 20 minut. W każdym teście dwie z ośmiu sekcji to ukryte sekcje zrównujące.

Wynik końcowy szacowany jest za pomocą dwóch sekcji rozumowania ilościowego (po 25 zadań każda), dwóch rozumowania werbalnego (30 zadań każda) oraz dwóch sekcji badających umiejętność posługiwania się językiem angielskim (27 zadań każda). Łącznie daje to 164 zadania (Allalouf i Ben Shakhar, 1998).

Schemat zrównywania jest analogiczny do schematu amerykańskiego SAT. W danej edycji izraelscy uczniowie wykonują 6 takich samych sekcji testowych, ale za to różne sekcje zrównujące, tym samym zrównywanie prowadzi się według planu nierównoważnych grup z testem kotwiczącym. Jedna sekcja zrównująca testu PET rozwiązywana jest zawsze przez około 1000 egzaminowanych. W sekcji zrównującej mogą zawierać się sekcje z wcześniej zdawanych testów. Dla każdego testu i dla każdej umiejętności wykorzystuje się proste zrównywanie liniowe (Beller, 1994; Rapp 1999).

Swedish Scholastic Assessment Test (SweSAT)

Egzamin, którego wynik decyduje o przyjęciu na szwedzkie uczelnie wyższe, powszechnie nazywany SweSAT (Swedish Scholastic Assessment Test), został wprowadzony w 1977 r. Na początku przeznaczony był dla kandydatów na studia, którzy zdecydowali się na nie aplikować po ukończeniu 25. roku życia, natomiast o przyjęciu młodszych osób decydowały wyniki nauki w szkole. Szybko jednak dostrzeżono zalety standaryzowanego testowania i SweSAT stał się egzaminem powszechnym.

Szwedzki test składa się z sześciu części: znajomość słownictwa (30 zadań rozwiązywanych w ciągu 15 minut); czytania ze zrozumieniem (24 zadania, na które uczeń ma 60 minut); czytanie ze zrozumieniem tekstów angielskich (24 zadania rozwiązywane w ciągu 50 minut); test matematyczny (20 zadań w 45 minut); umiejętność interpretowania danych (głównie wykresów, tabel i map: 20 zadań w 55 minut); wiedza ogólna (30 zadań, na które przeznaczono 25 minut). Wszystkie zadania w teście są zadaniami zamkniętymi, punktowanymi na skali 0–1. Cały test trwa 4 godziny i 10 minut (Stage i Igren, 2002).

Surowy wynik skalowany jest za pomocą metody ekwicentylowej i przekształcany na skalę z punktacją z przedziału od 0,0 do 2,0 punktów. Test zrównywany jest przy założeniu, iż populacje z roku na rok się nie zmieniają. Zrównywanie polega na przekształceniu wyników surowych metodą ekwicentylową przy uwzględnieniu płci, wieku oraz pochodzenia społecznego uczniów. Funkcja zrównująca wybierana jest w taki sposób, by z roku na rok wyniki egzaminacyjne w poszczególnych podgrupach utworzonych ze względu na wymienione zmienne nie różniły się[4] (Stage, 2004).

Od 1997 r. prowadzi się pracę nad zastosowaniem metod IRT oraz zewnętrznych i wewnętrznych kotwic w zrównywaniu testu. Przeprowadzono serie badań zrównujących; niestety nie dysponujemy informacją, czy zdecydowano się na wprowadzenie takiego sposobu zrównywania.

Problematyka zrównywania w wybranych międzynarodowych programach ewaluacyjnych

Trends in International Mathematics and Science Study (TIMSS)

TIMSS to międzynarodowe badanie osiągnięć edukacyjnych uczniów z matematyki

[4] Przedstawione w tym rozdziale informacje dotyczą sytuacji do 2004 r., z tego bowiem roku dysponujemy ostatnim anglojęzycznym źródłem informacji o zrównywaniu egzaminów w Szwecji. Nie wiemy, czy schemat zrównywania po 2004 r. zmienił się, czy pozostał w kształcie, w jakim prezentowany jest w tym artykule.

i przyrody (*science*) po czterech oraz ośmiu latach nauki. TIMSS został opracowany przez IEA (International Association for the Evaluation of Educational Achievement) po to, by umożliwić krajom w nim uczestniczącym międzynarodowe porównanie poziomu osiągnięć edukacyjnych oraz trendów ich zmian. Wyniki każdej edycji są wiązane z poprzednią. Dodatkowo przeprowadzanie badania na dwóch populacjach pozwala na monitorowanie zmian w kohortach – młodsza kohorta z wcześniejszej edycji badań staje się przedmiotem badań w kolejnej edycji jako starsza kohorta.

Badanie TIMSS zostało po raz pierwszy przeprowadzone w 1995 r., a kolejne rundy przeprowadzane są regularnie co 4 lata. W 2007 r. w badaniu uczestniczyło 59 krajów, łącznie 425 tys. uczniów. Ostatnie badanie zostało przeprowadzone w 2011 r., a jego wyniki opublikowano w 2012 r.

W TIMSS, tak jak we wszystkich międzynarodowych badaniach porównawczych, aby zachować trafność pomiaru, maksymalizuje się liczbę używanych w badaniu zadań, używając złożonego schematu doboru zadań. Zadania w TIMSS, osobno dla każdego poziomu nauczania, umieszczone są w 14 zeszytach. Każde z zadań pojawia się w dwóch zeszytach. Metodologia IRT oraz losowy przydział zeszytów dla uczniów pozwalają na łączenie wyników testowych w jedną skalę (Olson, Martin i Mullins, 2008). Badanie odbywa się na losowej próbce szkół i uczniów, zazwyczaj (bo istnieją różnice między krajami) losuje się około 150 szkół i 4000 uczniów do nich uczęszczających. W 2011 r. po raz pierwszy w badaniu uczestniczyła Polska.

Umiejętności mierzone są grupą ponad 300 zadań otwartych i zamkniętych (wspólnych dla matematyki i przyrody). Gdyby zastosować klasyczny schemat, w którym wszyscy uczniowie rozwiązują wszystkie zadania, łączny czas rozwiązania wszystkich zadań wyniósłby 8 godzin dla młodszych uczniów, a 10 godzin dla starszych. Złożony schemat doboru zadań pozwala ograniczyć ten czas do 72 minut w przypadku uczniów młodszych oraz do 90 minut dla uczniów starszych (dodatkowych 30 minut przeznaczonych jest na wypełnienie ankiety).

Skalowanie wyników odbywa się za pomocą dwuparametrycznego modelu IRT dla zadań otwartych oraz trójparametrycznego modelu dla zadań zamkniętych. W skalowaniu wykorzystuje się metodologię *plausible values* (dosłownie: „wiarygodnych wartości"), gdzie uczniom losuje się po 5 wartości z rozkładu *a posteriori* ich umiejętności przy uwzględnieniu odpowiedzi na wszystkie zadania testowe oraz odpowiedzi z kwestionariusza (Wu, 2005). Taka metodologia pozwala na precyzyjne oszacowanie nie tylko średnich wyników w całej populacji, ale i wyników dla podgrup oraz wariancji tych wyników, a także pozwala na dalsze analizy odnoszące osiągnięcia uczniów do ich cech kulturowo-społecznych, programów edukacyjnych itp.

Jednym z głównych celów badania TIMSS jest monitorowanie trendów. W centralnym miejscu zagadnień technicznych badania mieści się problematyka wiązania wyników z kolejnych edycji. Skala TIMSS została osadzona w badaniu przeprowadzonym w 1995 r., tak że średni wynik krajów biorących udział w badaniu wynosi 500 punktów, a odchylenie standardowe 100 zarówno dla młodszych, jak i starszych uczniów.

Przekształcenie mające na celu ulokowanie wyników na wspólnej skali odbywa się zgodnie z planem nierównoważnych grup z testem kotwiczącym przy użyciu modelowania IRT, łącznej kalibracji (*concurrent calibration*) bieżącego i poprzedzającego go cyklu oraz liniowej transformacji wyników.

Schemat przedstawiający takie procedury na przykładzie TIMSS 2007 i 2011 został przedstawiony na Rysunku 1. Kalibracja będąca wynikiem wcześniejszych badań (poprzednia kalibracja TIMSS 2007) stanowi punkt odniesienia. Edycje 2007 i 2011 zawierają zestaw tych samych zadań (pula zadań B), co umożliwia wspólną kalibrację po przeprowadzeniu badania w 2011 r.

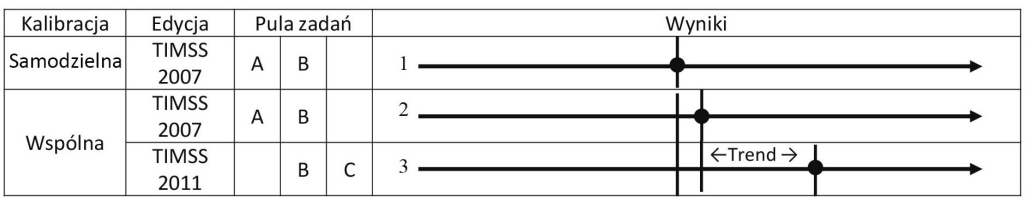
Średni poziom umiejętności uczniów biorących udział w badaniu w 2007 r. (kolumna: Wyniki) dla poprzedniej samodzielnej kalibracji będzie różnił się nieznacznie od wyników tych samych uczniów, które uzyskano przy pomocy wspólnej kalibracji. Jest to efekt poszerzenia puli zadań we wspólnej kalibracji o nowe zadania (z grupy C), których obecność wpływa na kształt skali uzyskanej z łącznej kalibracji.

Różnica między średnimi wynikami uczniów badanych w latach 2007 i 2011, oszacowana poprzez łączną kalibrację, wyraża faktyczną zmianę w poziomie umiejętności uczniów między edycjami badania, czyli trend. Aby wartość zmiany korespondowała z wcześniej ustanowioną skalą, wyniki łącznej kalibracji (zarówno dla roku 2007, jak i 2011) są liniowo przekształcane, tak, aby wyniki z łącznej kalibracji dla 2007 r. pokrywały się z wynikami uzyskanymi we wcześniejszej kalibracji. W ten sposób różnica między latami 2007 a 2011 pozostaje taka sama, z tym że liniowe przesunięcie lokuje je na wcześniej ustalonej skali (Olson, Martin i Mullins, 2008).

Progress in International Reading Literacy Study (PIRLS)

PIRLS jest międzynarodowym badaniem stawiającym sobie za cel pomiar biegłości w czytaniu wśród dzieci mających za sobą czwarty rok nauki. W Polsce PIRLS można traktować jako ocenę umiejętności uczniów, którzy kończą etap kształcenia zintegrowanego, ponieważ badaniem objęci zostali uczniowie klasy trzeciej. Podobnie jak TIMSS, PIRLS został opracowany przez IEA i jest swoistym dopełnieniem TIMSS. Pomiary PIRLS odbywają się cyklicznie co 5 lat. W 2006 r. przeprowadzono je w 40 krajach (w tym w Polsce). Kolejna edycja odbyła się w 2011 r. W większości krajów do badania losowano około 150 szkół, a z każdej szkoły dwa oddziały (lub jeden, jeżeli szkoła była jednoodziałowa). Taki schemat losowania miał docelowo doprowadzić do losowej, reprezentatywnej próby uczniów liczącej nie mniej niż 4000 osób. Taki schemat zastosowano również w Polsce, gdzie w 2006 r. test wykonało 4854 uczniów z 250 oddziałów klasy trzeciej w 148 szkołach podstawowych.

Konstrukcja i techniczne aspekty badania PIRLS są analogiczne do siostrzanego badania TIMSS. Wyniki testu są skalowane metodami IRT (modelem trzyparametrycznym dla zadań zamkniętych, modelem dwuparametrycznym dla zadań otwartych) przy użyciu metodologii *plausible values*. Skalę badania PIRLS 2006 zakotwiczono w wy-



Rysunek 1. Łączenie wyników dwóch edycji TIMSS (Na podstawie: Olson, Martin i Mullins, 2008, s. 242).

nikach badania z 2001 r. Tak więc średnia dla uczniów wszystkich krajów uczestniczących w badaniu w 2001 r. wynosiła 500 punktów, a odchylenie standardowe 100. Wyniki z 2006 r. przedstawiane są zatem na skali z 2001 r., co pozwala na bezpośrednie porównanie obydwu edycji badania.

W badaniu PIRLS, podobnie jak w TIMSS, stosuje się złożony schemat doboru zadań. Pozwala to znacząco poprawić treściową reprezentatywność testów. W każdej edycji PIRLS przygotowano 10 tekstów, a do każdego z tekstów zadawano przeciętnie 12 pytań. Średnio połowa z nich wymagała wyboru jednej z czterech odpowiedzi, połowa zaś krótkiej odpowiedzi pisemnej. Każdy z uczniów otrzymał losowo po dwa teksty, czyli rozwiązywał około 24 zadań testowych.

W badaniu PIRLS skonstruowanych jest kilka skal. Zadania zostały zaprojektowane tak, aby mierzyć następujące umiejętności: wyszukiwanie informacji (*focus on and retrieve explicitly stated information*); wyciąganie bezpośrednich wniosków (*make straightforward inferences*); wiązanie i interpretacja informacji (*interpret and integrate ideas and information*); ocena i interpretowanie informacji zawartych w tekście (*examine and evaluate content, language, and textual elements*). W 2006 r. do konstrukcji skal wykorzystano w sumie 174 zadania.

Sposób wiązania ze sobą poszczególnych wyników z kolejnych edycji badania przebiega dokładnie tak samo jak w opisanym poprzednio badaniu TIMSS. Odbywa się to zgodnie z planem nierównoważnych grup z testem kotwiczącym przy użyciu modelowania IRT, łącznej kalibracji (*concurrent calibration*) bieżącego i poprzedzającego go cyklu oraz liniowej transformacji wyników (Olson, Martin i Mullins, 2009).

Programme for International Student Assessment (PISA)

Program międzynarodowej oceny umiejętności uczniów PISA jest największym międzynarodowym badaniem edukacyjnym przeprowadzanym w ponad 60 krajach (w tym w Polsce) na reprezentatywnych losowych próbach uczniów, co trzy lata, począwszy od 2000 roku. Badanie to zarządzane jest przez Organizację Współpracy Gospodarczej i Rozwoju (OECD). Głównym celem PISA jest pomiar wiedzy i umiejętności niezbędnych do sprawnego funkcjonowania we współczesnym społeczeństwie w dziedzinach: matematyki, czytania oraz znajomości nauk przyrodniczych. W PISA pomiar skupia się na ocenie posługiwania się pojęciami i ich rozumienia oraz wykorzystywaniem wielu umiejętności. Z założenia mierzony jest poziom kompetencji niezbędnych uczniom w życiu dorosłym, na rynku pracy i do tego, aby w pełni funkcjonować we współczesnym społeczeństwie demokratycznym (OECD, 2012).

W każdym cyklu nacisk położony jest na jedną z wymienionych umiejętności. W 2000 r. szczegółowej ocenie poddano czytanie, w 2003 r. – matematykę, w 2006 r. – nauki przyrodnicze, a w 2009 r. ponownie czytanie. Umiejętności, w zależności od przedmiotu i edycji, mierzone są za pomocą od kilkudziesięciu do ponad stu zadań. Obok pomiaru umiejętności zbieranych jest wiele dodatkowych informacji. W każdym z 60 krajów uczeń wypełnia ankietę zawierającą baterię pytań dotyczących statusu społecznego rodziców, warunków życia, motywacji do nauki i aspiracji edukacyjnych.

W badaniu wykorzystuje się złożony schemat dystrybucji zadań oraz wielowymiarowe skalowanie Rascha z *plausible values*. Skala wyników została zakotwiczona w pierwszych edycjach badania, tak by średni wynik uczniów z krajów OECD wynosił 500 punk-

tów, a odchylenie standardowe 100 (rozwiązanie analogiczne do PIRLS i TIMSS). Do łączenia wyników z kolejnych edycji wybrany został dwustopniowy schemat wykorzystujący wspólne zadania (*common items*) oraz wspólnych uczniów biorących udział w testowaniu (*common persons*).

Łączenie wyników PISA między kolejnymi edycjami przedstawione zostanie na przykładzie skali mierzącej umiejętność czytania ze zrozumieniem edycji z lat 2006 i 2009. Łączenie wyników z dwóch edycji w tym przykładzie ma dwojaki charakter i odbywa się dwustopniowo. Najpierw zrównuje się skale za pomocą wspólnych zadań – ten element procedury odnosi się do ustalenia wspólnej skali między zadaniami łączącymi (*linking items*) w kolejnych edycjach. Następnie ustalana jest wspólna skala zadań łączących: dodatkowych zadań wykorzystanych w 2009 r. Opisywany schemat zrównywania przedstawiony został na Rysunku 2.

W procedurze zrównywania wyników w pierwszym kroku, kalibrowana jest próba PISA z roku 2009. Na jej podstawie szacowane są parametry zadań. Jako że w PISA używa się modelu Rascha, de facto jedynym estymowanym parametrem odnoszącym się do zadania jest jego trudność.

W 2009 r., aby zmierzyć umiejętność czytania ze zrozumieniem, wykorzystano 101 zadań; 26 z nich określone zostały jako zadania łączące, ponieważ użyto ich we wcześniejszej edycji badania. Uzyskane parametry zadań łączących z kalibracji dokonanej w 2009 r. przekształcane są (z dodaniem lub odjęciem stałej) tak, by średnia trudność zadań łączących była równa w obu edycjach. Średnia 26 zadań łączących w 2009 r. wyniosła -0,0885, a w 2006 r. było to 0,0021. Różnica między tymi dwoma kalibracjami wynosi zatem 0,0906. Gdyby w teście PISA w 2009 roku nie było dodatkowych zadań, na tym proces łączenia wyników mógłby się zakończyć: wskutek przesuwania liniowo wyników uzyskanych w kalibracji z 2009 r. o 0,0906 skala zostałaby dostosowania do skali z 2006 r. Dodatkowe zadania wprowadzone w 2009 r. (podobnie jak w przykładzie dla TIMSS) wymagają kolejnego kroku, który dostosowałby skalę nowo użytych zdań do skali z 2006 r.

W drugim kroku (*common persons "linking"*) szacowany jest poziom umiejętności uczniów – najpierw za pomocą kalibracji wszystkich zadań, potem jedynie za pomocą zadań łączących. Różnica w średnich umiejętnościach uczniów dla tych dwóch kalibracji wynosiła 0,1261.

<table>
<tr><td></td><td colspan="2"></td><td colspan="2">Zadania</td></tr>
<tr><td rowspan="3">Krok 1. wspólne zadania → ←</td><td rowspan="3">Uczniowie</td><td>2006</td><td>26 zadań łączących czytanie</td><td></td></tr>
<tr><td>2009</td><td>26 zadań łączących czytanie</td><td>75 nowych zadań mierzących umiejętność czytania</td></tr>
<tr><td colspan="3"></td><td colspan="2">Krok 2. wspólni uczniowie
→ ←</td></tr>
</table>

Rysunek 2. Schemat zrównywania dwu edycji badania PISA 2006 i 2009.

Wartości uzyskane w dwóch krokach są następnie dodawane, a skala powstała za pomocą kalibracji z roku 2009 (zadania nowe plus zadania łączące) przesunięta o uzyskaną sumę. Na koniec surowa skala wynikająca z domyślnych ustawień programu wykorzystanego do estymacji parametrów modelu jest liniowo przekształcana w skalę PISA.

Problematyka zrównywania w wybranych narodowych programach ewaluacyjnych

National Assessment of Educational Progress (NAEP, Stany Zjednoczone)

Program NAEP po raz pierwszy został wprowadzony w roku szkolnym 1969/70 i od tego czasu funkcjonuje w Stanach Zjednoczonych jako podstawowe narzędzie do pomiaru poziomu umiejętności uczniów do celów polityki edukacyjnej. Jest to projekt rządowy, za którego administrację i raportowanie wyników odpowiada National Center for Education Statistics (NCES), będący ramieniem Institute of Educational Statistics w U.S. Department of Education. Przez większość czasu program był prowadzony we współpracy z Educational Testing Service (ETS).

Zgodnie z wytycznymi ustawowo zapisanymi przez Kongres w 1988 r. NAEP raportuje w odstępie dwuletnim wyniki pomiaru umiejętności matematycznych oraz umiejętności czytania, a w odstępie czteroletnim – wyniki z nauk przyrodniczych oraz z pisania. Okazjonalnie badane są również inne umiejętności: z przedmiotowego zakresu sztuki, wiedzy o społeczeństwie, ekonomii, geografii czy historii Stanów Zjednoczonych. Testowanie NAEP przeprowadzane jest na reprezentatywnej próbie szkół, a jego wyniki nie są raportowane pojedynczym uczniom, szkołom czy dystryktom szkolnym. Do 1990 roku prawo zabraniało także raportowania wyników dla poszczególnych stanów (Beaton i Zwick, 1992). Zasadniczym celem programu NAEP jest ocena poziomu umiejętności podstawowych na poziomie całego kraju oraz na poziomie wybranych subpopulacji uczniów (np. ze względu na płeć oraz pochodzenie etniczne) lub typów szkół, a także ocena zmian w poziomie umiejętności na przestrzeni lat.

Badanie NAEP odbywa się w dwóch formach – badania głównego (*main* NAEP) oraz badania skoncentrowanego na analizie trendów długoterminowych (*long-term trend assessment*, NAEP LTTA). W badaniu głównym, które odbywa się co dwa lata, testy są konstruowane tak, aby odzwierciedlały aktualny stan programu nauczania. Od 1988 r. testowaniu są poddawani uczniowie z trzech równo oddalonych od siebie grup wiekowych: klasy 4 (9 lat), klasy 8 (13 lat) oraz klasy 12 (17 lat). Celem badania głównego jest dostarczenie danych do przeprowadzania porównań międzygrupowych w danym roku, także między uczniami z różnych poziomów edukacyjnych, jak również ocena zmian w poziomie umiejętności na krótszych od LTTA odcinkach czasu. Zrównywanie wertykalne w NAEP jest zapewnione poprzez występowanie wspólnych zadań w arkuszach rozwiązywanych przez uczniów w różnym wieku (Yamamoto i Mazzeo, 1992). Ze względu na liczbę zadań przekraczającą możliwości rozwiązania przez pojedynczego ucznia, zadania są rozprowadzone po populacji z wykorzystaniem zrównoważonego schematu blokowego (*balanced incoplete block*, BIB). Arkusz testowy dla pojedynczego ucznia składa się z dwóch 25-minutowych bloków. W zależności od przedmiotu wyniki są raportowane na skali o rozpiętości 0–300 lub 0–500 punktów. Nauczyciele wybranych przedmiotów oraz poziomów nauczania proszeni są o wypełnienie dodatkowych ankiet sprawdzających ich doświadczenie, stosowane metody nauczania, a także zbierających informacje o uczniach.

Badanie służące do analizy trendów długookresowych odbywa się w cyklu czteroletnim i ma za zadanie odniesienie wyników uczniów do pierwszych badań NAEP na skali umiejętności matematycznych oraz umiejętności czytania. Raport z przeprowadzonych w 2008 r. badań NAEP LTTA wyznaczał trend od 1973 r. (Rampey, Dion, Donahue, 2009). W badaniu LTTA biorą udział próby uczniów niezależnie losowane od badania głównego (Beaton i Zwick, 1992), a łączność z głównym NAEP polega na losowej równoważności grup. Ze względu na mniejszą liczbę zadań w porównaniu do głównego badania NAEP, trendy są wyznaczane niezależnie dla każdego poziomu nauczania i bez rozbicia na dodatkowe grupy (Yamamoto i Mazzeo, 1992). Arkusz testowy dla pojedynczego ucznia składa się z trzech 15-minutowych bloków. Wyniki są raportowane na skali o rozpiętości 0–500 punktów.

Badania długookresowego trendu NAEP korzystają z zadań wspólnych z wcześniejszymi zastosowaniami testu i są budowane zgodnie ze stałymi specyfikacjami, dzięki czemu zapewniona jest ścisła łączność z wcześniejszymi wynikami, mimo zachodzących w tym czasie przemian w programach nauczania. W 2004 roku LTTA jednak przeszedł znaczne przekształcenia, mające dostosować go do zmian w ogólnej metodologii badania NAEP (np. włączenia dostosowań dla uczniów ze specjalnymi potrzebami edukacyjnymi) oraz zwiększyć jego trafność. Ze względu na wprowadzone zmiany przeprowadzone zostały dodatkowe badania pomostowe (*bridge studies*), weryfikujące zgodność nowszej wersji LTTA z wcześniejszymi edycjami. Różne zmiany na przestrzeni lat wprowadzano również do głównego badania NAEP i im również towarzyszyły dodatkowe badania pomostowe (szczegółowe zestawienie badań pomostowych NAEP podają Nellhaus, Behuniak i Stancavage, 2009). W ostatnich latach prowadzone są intensywne badania nad przeprowadzaniem głównego badania NAEP z wykorzystaniem komputerów (Sandene i in., 2005). Badanie umiejętności pisania w klasach 8 i 12 w 2011 r. zostało w całości przeprowadzone komputerowo (NAGB, 2010).

Raportowane wyniki z badania NAEP są uzyskiwane z wykorzystaniem modelowania IRT. W zależności od formatu zadań stosowany jest dwuparametryczny lub trójparametryczny model logistyczny dla zadań ocenianych dychotomicznie, a dla zadań ocenianych na większą liczbę punktów stosuje się uogólniony model oceny częściowej. Kalibracja testów jest przeprowadzana z wykorzystaniem programów PARSCALE oraz BILOG, które zostały specjalnie dostosowane do potrzeb badania NAEP.

Konsekwencją podstawowych założeń NAEP jest konieczność sprowadzania wyników z różnych lat oraz z różnych poziomów edukacyjnych do wspólnej skali. Stosuje się w tym celu metodę łącznej kalibracji, w której zakotwiczenie testów za pomocą wspólnych zadań pozwala na oszacowanie rozkładów umiejętności uczniów z różnych populacji na wspólnej skali umiejętności. Surowe oszacowania uzyskiwane po kalibracji w programach statystycznych są następnie liniowo przekształcane do skali o docelowej średniej i odchyleniu standardowym. W niektórych wypadkach łącze między testami polega jedynie na losowej równoważności grup. Szczegółowy schemat i procedury łączenia zależą od przedmiotów, lat, w jakich testy były przeprowadzane, ewentualnie od zastosowania dodatkowych prób w badaniach pomostowych. Do wtórnych analiz wyników wykorzystywane są *plausible values* uzyskane na podstawie dopasowanego modelu IRT przy warunkowaniu ze względu na istotne zmienne kontekstowe. Szczegółowy opis wspomnianych

procedur skalowania i linkowania wyników znajduje się na stronach NCES poświęconych technicznym aspektom NAEP[5].

Testy Education Quality and Accountability Office (EQAO tests, Kanada, Ontario)

W Kanadzie nie istnieje jeden ogólnokrajowy system egzaminacyjny, jednak poszczególne prowincje prowadzą własne systemy ewaluacyjne i niezależnie testują swoich uczniów. Przykładem takiej prowincji jest Ontario. W 1996 r. uruchomiony został tam program ewaluacyjny EQAO, którego częścią jest testowanie uczniów (EQAO, 2011).

Testy EQAO mierzą umiejętności czytania, pisania oraz umiejętności matematyczne. Rozwiązywane są przez uczniów szkół podstawowych z klas 3 i 6. Uczniowie klasy 9 rozwiązują rozbudowany test osiągnięć szkolnych w zakresie matematyki, a uczniowie po 11. roku nauki – również rozbudowany test mierzący umiejętności czytania ze zrozumieniem oraz umiejętności tworzenia wypowiedzi pisemnych (Ontario Secondary School Literacy Test, OSSLT). Testy przeprowadzane są corocznie i są obowiązkowe dla wszystkich uczniów szkół publicznych. Uczniowie szkół prywatnych nie są zobowiązani do przystępowania do testów, lecz w większości przypadków uczestniczą w testowaniu (EQAO, 2011).

Testy mają w założeniu mierzyć, jaki poziom umiejętności uzyskują uczniowie w stosunku do obowiązującego w prowincji Ontario programu nauczania. W testach znajdują się zadania zamknięte, otwarte oraz krótkie wypowiedzi pisemne (mierzące umiejętność posługiwania się językiem angielskim). Testy po trzeciej klasie składają się z 36 zadań mierzących umiejętność czytania ze zrozumieniem, z 14 zadań mierzących umiejętność wypowiadania się w formie pisemnej oraz z 36 zadań z matematyki. Test mierzący umiejętności matematyczne przeprowadzany w klasie 9 składa się z około 30 zadań.

Wyniki każdego testowania dostarczane są uczniom, a średnie wyniki szkół są publicznie dostępne. Dodatkowo zdanie testu mierzącego umiejętności posługiwania się językiem angielskim przeprowadzanego w 11 klasie jest niezbędne do otrzymania certyfikatu wykształcenia drugiego stopnia – Ontario Secondary School Diploma (OSSD).

Wyniki uczniów prezentowane są na standaryzowanej skali, w której minimalny wynik wynosi 200, a najwyższe wyniki sięgają 400 punktów. Obok wyniku na skali przydzielane są oceny odzwierciedlające stopień opanowania przez ucznia danych umiejętności.

Do zrównywania wykorzystuje się schemat analogiczny do schematu amerykańskiego testu SAT: plan nierównoważnych grup z testem kotwiczącym. Różnica jest taka, że do procedury zrównywania wykorzystuje się trójparametryczny model IRT (3PLM). Parametry zadań, które znajdują się zarówno w bieżącym teście, jak i we wcześniejszej edycji, podczas estymacji modelu IRT są ustalane na wartościach otrzymanych we wcześniejszej edycji testu. W procesie estymacji parametrów bieżącego testu parametry zadań, które stanowią kotwicę, nie są estymowane, tylko przyjmują wartości oszacowane we wcześniejszej edycji. Jest to tak zwana metoda zrównywania ustalonych parametrów (*fixed parameters).* Warto zwrócić uwagę, iż test zrównywany jest również pionowo (*vertical scaling*), czyli wyniki uczniów z różnych poziomów kształcenia są bezpośrednio porównywalne.

[5] http://nces.ed.gov/nationsreportcard/tdw/analysis/

National Assessment Program – Literacy and Numeracy (NAPLAN, Australia)

W 2008 r. w Australii po raz pierwszy przeprowadzony został ogólnokrajowy egzamin mierzący umiejętności językowe oraz matematyczne: National Assessment Program – Literacy and Numeracy (NAPLAN). Test jest obowiązkowy dla wszystkich uczniów: przeprowadza się go w klasach: 3, 5, 7 oraz 9 (Freeman, 2009). Szczegółowy plan testu NAPLAN przedstawiony został w Tabeli 2.

Wyniki z każdego testu są skalowane za pomocą modelu Rascha. Wyniki uczniów oraz szkół generuje się z wykorzystaniem estymatora WLE (*weighted likelihood estimates*), a następnie przekształca do skali o średniej 500 punktów i odchyleniu standardowym 100. Wyniki na poziomie poszczególnych stanów oraz wyniki ogólnonarodowe uzyskuje się dzięki metodologii *plausible values*.

Każdego roku jest prowadzone tak zwane studium zrównujące, czyli badanie, w którym bierze udział losowa próbka uczniów również przystępujących do testu NAPLAN. Studium zrównujące odbywa się tydzień po testowaniu zasadniczym. W teście z tego badania znajdują się zadania, które pozwalają na zrównanie wyników z bieżącej edycji z wcześniejszymi edycjami testu. Zrównywanie odbywa się za pomocą modelu Rascha i podobnie jak w teście z Ontario, polega na estymowaniu parametrów pytań w taki sposób, by były zgodne z estymacją we wcześniejszej edycji testu. Inne podobieństwo NAPLAN do tekstu z Ontario to zrównywanie pionowe (*vertical scaling*), prowadzone oprócz zrównywania kolejnych edycji egzaminu. Dzięki temu wyniki uczniów z różnych poziomów kształcenia są bezpośrednio porównywalne (Cook, 2009).

National Assessment of Basic Competencies (Węgry)

Obok narodowego systemu egzaminacyjnego (egzamin zdawany przez uczniów po ukończeniu klasy 8, który jest przepustką do czteroletniej szkoły średniej) na Węgrzech wprowadzono system oceny kompetencji uczniów dla klas 6, 8 oraz 10. System ma przede wszystkim służyć ewaluacji elementów systemu oświatowego, analizie zmian w czasie, a w mniejszym zakresie również diagnozie kluczowych kompetencji poszczególnych uczniów.

Węgierskie testy zawierają zadania z matematyki oraz zadania sprawdzające umiejętność interpretacji tekstu (po dwa 45-minutowe bloki na każdą część). Zeszyty testowe zawierają po około 60 zadań z każdej dziedziny. Test umiejętności czytania i interpretacji sprawdza kompetencje kulturowe uczniów, umie-

Tabela 2
Plan testu NAPLAN

Poziom testowania/ Mierzona umiejętność	Klasa 3	Klasa 5	Klasa 7	Klasa 9
Czytanie	35 zadań (45 min)	35 zadań (50 min)	47 zadań (65 min)	47 zadań (65 min)
Język:				
a) gramatyka	25 zadań (40 min)	25 zadań (40 min)	30 zadań (45 min)	26 zadań (45 min)
b) ortografia	23 zadania (40 min)	23 zadania (40 min)	24 zadania (40 min)	28 zadań (40 min)
Matematyka:				
a) bez kalkulatora	35 zadań (45 min)	40 zadań (50 min)	32 zadania (40 min)	32 zadania (40 min)
b) z kalkulatorem	---	---	32 zadania (40 min)	32 zadania (40 min)

jętność interpretacji tekstów, wyszukiwania informacji, kojarzenia informacji z różnych źródeł. Test zawiera krótkie historie, nowele, narracje, artykuły z gazet, reklamy. Dodatkowo uczniowie (dobrowolnie) wypełniają ankietę dotyczącą ich pochodzenia społecznego (dane te pozwalają stworzyć indeks HBI – Home Background Index: Balázsi, 2006). Analiza tych danych za pomocą regresji liniowej pozwala wykryć zmienne społeczne, które najsilniej determinują osiągnięcia ucznia. Zadaniem indeksu HBI jest uchwycenie całego „bagażu społecznego", który uczeń przynosi ze sobą, przestępując progi szkoły, i który wpływa na wyniki ucznia. Celem samych zadań nie jest natomiast sprawdzenie, czy uczniowie posiadają wiedzę wymaganą na danym poziomie edukacji, lecz raczej sprawdzenie, czy potrafią użyć tej wiedzy, aby rozwiązać problemy, z którymi spotykają się na co dzień. Wszyscy uczniowie z danej populacji rozwiązują test, jednak wyniki statystyczne opracowywane są na poziomie krajowym tylko dla reprezentatywnej próby uczniów (20 uczniów z każdej szkoły).

Wyniki skalowane są za pomocą modeli IRT – zadania zamknięte modelem trzyparametrycznym, a zadania otwarte modelem dwuparametrycznym. Wyniki przedstawiane są w dwojakiej formie. Po pierwsze, przedstawiane są na skali o średniej 500 punktów i odchyleniu standardowym 100. Po drugie, ustalonych zostało 5 poziomów umiejętności zdefiniowanych zarówno przez pedagogiczne, jak i statystyczne kryteria (podobnie jak w badaniu PISA). Poziomy zostały zdefiniowane przez 4 progi, które wyznaczają granice między coraz bardziej złożonymi operacjami umysłowymi.

Do zrównania wyników wykorzystuje się równoważne grupy. Test kotwiczący przeprowadzany jest corocznie na próbce około 20 uczniów ze 150 szkół. Wylosowani uczniowie, oprócz testu z danego roku, wykonują tak zwany test bazowy, który jest tajny i niezmienny każdego roku. Wyniki tego testu oraz zastosowanie IRT pozwalają połączyć wyniki uczniów wykonujących testy w kolejnych latach (Balázsi, 2006) i monitorować zmiany w wynikach uczniów w kolejnych latach. Oprócz zrównania między latami, węgierskie testy zaprojektowane zostały w taki sposób, aby wyniki uczniów z różnych klas mogły być bezpośrednio porównywalne (osiąga się to również za pomocą omawianej już zewnętrznej kotwicy).

Podsumowanie

Najczęściej spotykanymi na świecie schematami zrównywania są: plan nierównoważnych grup z testem kotwiczącym oraz plan z równoważnymi grupami. Obydwa plany na dużą skalę zastosowano po raz pierwszy w Stanach Zjednoczonych: pierwszy w SAT, drugi w ACT. Jeżeli plan testowania nie zakłada powtórnego użycia zadań testowych, tak jak w Szwecji, poszczególne testy próbuje się zrównywać za pomocą metod skalowania wyników, które zakładają, iż z roku na rok populacje zdających nie zmieniają się w znaczącym stopniu.

Do zrównywania używa się metod opartych zarówno na klasycznych ekwicentylowych przekształceniach wyników obserwowanych, jak i na modelowaniu IRT. Należy zaznaczyć, iż metody ekwicentylowe stosowane są w starszych systemach egzaminacyjnych, gdzie nie bez znaczenia jest ciągłość stosowanej metody; w nowszych systemach egzaminacyjnych chętnie sięga się po modele IRT. W każdym z przywoływanych rozwiązań wyniki są skalowane (tj. uczniom nie są komunikowane ich wyniki surowe). Charakterystyczne jest także to, iż większość testów wysokiej stawki składa się z dużej, w porównaniu z polskimi warunkami, liczby zadań.

W polskim systemie egzaminacyjnym do chwili obecnej nie wprowadzono żadnego systemowego narzędzia umożliwiającego

coroczne zrównywanie egzaminów, brak również jakichkolwiek sygnałów, czy wprowadzenie takich rozwiązań jest w ogóle planowane. Zespół Pracowni Analiz Osiągnięć Uczniów w Instytucie Badań Edukacyjnych prowadzi badanie zrównujące wyniki egzaminów zewnętrznych, które jest próbą wypełnienia tej luki. Wspomniane badanie ma jednak charakter naukowy i jakkolwiek informacje z niego płynące są bardzo cenne dla systemu polskiej oświaty, to nie może ono zastąpić odpowiednich rozwiązań systemowych, na przykład takich jak zaprezentowane w tym artykule.

Literatura

ACT (2007). *Technical Manual*. Pobrano z: http://www.act.org/aap/pdf/ACT_Technical_Manual.pdf

Allalouf, A. i Ben Shakhar G. (1998). The effect of coaching on the predictive validity of scholastic aptitude tests. *Journal of Educational Measurement 35*(1), 31–47.

Balázsi, I. (2006). *National Assessment of Basic Competencies in Hungary*. Pobrano z: http://www.iaea2006.seab.gov.sg/conference/download/papers

Beaton, A. E. i Zwick R. (1992). Overview of the National Assessment of Educational Progress. *Journal of Educational Statistics. 17*(2), 95–109.

Beller, M. (1994). Psychometric and social issues in admissions to Israeli universities. *Educational Measurement: issues and practice 13*(2), 12–20.

Cook, J. (2009). An event start: innovative resources to support teachers to better monitor and better support students measured below benchmark. *ACER Research Conference*, series 3.

Davier von, A. A. (2011). A statistical perspective on equating test scores. W: A. A. von Davier (red.), *Statistical models for test equating, scaling, and linking* (s. 1–17). New York, NY: Springer-Verlag.

Davier von, A. A., Holland, P. W. i Thayer, D. T. (2004). *The kernel method of test equating*. New York, NY: Springer-Verlag.

Davier von, M. i Davier von, A. A. (2011). A general model for IRT scale linking and scale transformations. W: A. A. von Davier (red.), *Statistical models for test equating, scaling, and linking* (s. 1–17). New York, NY: Springer-Verlag.

Dorans, N. J. i Holland, P. W. (2000). Population invariance and the equatability of tests: Basic theory and the linear case. *Journal of Educational Measurement, 37*(4), 281–306.

EQAO (2011). *EQAO's technical report for the 2009–2010 assessments*. Toronto: Autor.

Freeman, C. (2009). First national literacy and numeracy tests introduced. *Research Developments 20*(20).

Gruijter, D. N. M. i van der Kamp, L. J. (2005). *Statistical test theory for education and psychology*. Pobrano z http://irt.com.ne.kr/data/test_theory.pdf

Holland, P. W., Dorans N. J. i Petersen N. S. (2007). Equating test scores. W: C. R. Rao i S. Sinharay (red.). *Handbook of statistics*, (t. 26) Psychometrics (s. 169–204). NY: Elsevier.

Kolen, M. J. (1984). Effectiveness of analytic smoothing in equipercentile equating. *Journal of Educational Statistics*, 9, 25–44.

Kolen, M. J. (2007). Data collection designs and linking procedures. W: N. J. Dorans, M. Pommerich, P. W. Holland (red.), *Linking and aligning scores and scales* (s. 31–55). New York, NY: Springer-Verlag.

Kolen, M. J., i Brennan R. L. (2004). T*est equating, scaling, and linking: Method and practice* (wyd. 2). New York, NY: Springer-Verlag.

Lawrence, I., Rigol, G. W., Van Essen, T. i Jackson, C. A. (2002). *A historical perspective on the SAT: 1926–2001*. College Board Research Report No. 2002–7. College Entrance Examination Board, New York.

Linden van der, W. J., (2011). Local observed-score equating. W: A. A. von Davier (red.), *Statistical models for test equating, scaling, and linking* (s. 201–223). New York, NY: Springer-Verlag.

Liu, J. i Walker M. E. (2007). Score linking issues related to test content changes. W: N. J. Dorans, M. Pommerich i P. W. Holland (red.), *Linking and aligning scores and scales* (s. 109–134). New York, NY: Springer-Verlag.

Livingston, S. A. (2004). *Equating test scores (without IRT)*. Princeton, NJ: ETS.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

NAGB (2010). *Writing framework for the 2011 National Assessment of Educational Progress*. National Assessment Governing Board, U.S. Department of Education, Washington, DC: U.S. Government Printing Office.

Nellhaus, J., Behuniak, P. i Stancavage, F. B. (2009). *Guiding principles and suggested studies for determining when the introduction of a new assessment framework necessitates a break in trend in NAEP.* NAEP Validity Studies, American Institutes for Research: Palo Alto, CA.

OECD. (2012). *PISA 2009 Technical Report*. Paris: OECD Publishing.

Olson, J. F., Martin, M. O. i Mullins, I. V. S. (2008). *TIMSS 2007 Technical Report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston: Boston College.

Olson, J. F., Martin, M. O., i Mullins, I. V. S. (2009). *PIRLS 2006 Technical Report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston: Boston College.

Pokropek A. (2011). Zrównywanie wyników egzaminów zewnętrznych w kontekście międzynarodowym. W: *XVII Konferencja Polskiego Towarzystwa Diagnostyki Edukacyjnej*, Kraków 2011.

Rampey, B. D., Dion, G. S. i Donahue, P. L. (2009). *NAEP 2008 Trends in Academic Progress* (NCES 2009–479). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education: Washington, D.C.

Rapp, J. (1999). *Linear and Equipercentile Methods for Equating PET, NITE*. Pobrano z: https://www.nite.org.il/files/reports/e266.pdf

Sandene, B., Horkay, N., Bennett, R., Allen, N., Braswell, J., Kaplan, B. i Oranje, A. (2005). *Online assessment in mathematics and writing: reports from the NAEP technology-based assessment project, research and development series* (NCES 2005–457). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Stage, C. (2004). *Notes from the Tenth International SweSAT Conference*. Umeå, June 1–3.

Stage, C. i Ígren, G. (2002). *The Swedish Scholastic Assessment Test (SweSAT)*. Department of Educational Measurement, Umeå Univ.

Wu, M. (2005) The role of plausible values in large-scale surveys. *Studies in Educational Evaluation 31*, 114–128.

Yamamoto, K., Mazzeo, J. (1992). Item Response Theory scale linking in NAEP. *Journal of Educational Statistics, 17*(2), 155–173.