**dr Artur Pokropek**
Instytut Badań Edukacyjnych
Zespół Edukacyjnej Wartości Dodanej

# Zrównywanie wyników egzaminów zewnętrznych w kontekście międzynarodowym

## Wstęp

Aby możliwym było porównywanie osiągnięć szkolnych uczniów, którzy rozwiązywali testy egzaminacyjne przeprowadzane w różnych administracjach, niezbędne jest wprowadzenie takich mechanizmów, które pozwolą na zrównanie wyników testowych. Procedury zrównywania pozwalają na wyrugowanie z końcowego wyniku losowych wahań trudności między formami jednego egzaminu. Jest to niezwykle ważne w takich testach jak egzamin maturalny, gdzie wynik decyduje o przyszłych losach edukacyjnych ucznia. Jeżeli w teście nie stosuje się procedur zrównujących, wyniki z dwóch różnych edycji nie są porównywalne. Gdy testy nie są zrównywane, nie można również analizować, czy jakość nauczania na danym poziomie kształcenia, a wraz z nią poziom realizacji celów kształcenia, zmieniają się w przeciągu kolejnych lat, czy też nie. Utrudniona jest tym samym ewaluacja pracy szkoły i nauczycieli.

Zarówno dojrzałe systemy testowania, jak i większość nowo powstałych systemów, wprowadzają do konstrukcji testów mechanizmy pozwalające na zrównywanie. Polski system egzaminacyjny na tle innych krajów wypada blado. Do tej pory nie wprowadzono (i na szczeblach decyzyjnych póki co nie planuje się wprowadzić) żadnego systemowego narzędzia umożliwiającego coroczne zrównywanie egzaminów. Konstrukcja polskich egzaminów, sposób ich standaryzacji i podejście do skalowania wyników są anachroniczne i często ocierają się o popełnianie błędów w sztuce testowania.

Niniejszy artykuł ma przybliżyć rozwiązania stosowane w testowaniu na świecie ze szczególnym uwzględnieniem problematyki zrównywania. Nie jest to przegląd kompletny, nie było to ambicją autora, a raczej przegląd, który ma pokazać najważniejsze trendy w tej dziedzinie. Zostanie tutaj przedstawionych sześć testów o różnych zastosowaniach, pochodzących z różnych części świata. Każdy test zostanie pokrótce omówiony i wskazane zostaną mechanizmy umożliwiające jego zrównywanie.

## Stany Zjednoczone

Stany Zjednoczone są pionierem w dziedzinie nowoczesnych technik testowania. Nie może zatem dziwić, iż w artykule dotyczącym technicznych aspektów testowania w perspektywie porównawczej 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 przypadkach 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 egzaminów w innych krajach przedstawianych w kolejnych częściach tego artykułu.

## SAT (Scholastic Assessment Test)

SAT to najstarszy funkcjonujący po dziś dzień (z pewnymi zmianami) test osiągnięć szkolnych na świecie. Powstał w 1926 roku na zlecenie *College Board*, organizacji zrzeszającej przede wszystkim uczelnie wyższe oraz inne organizacje edukacyjne. Pierwszy test przeprowadzony został w 1926 roku. Trwał 90 minut i składał się z 315 pytań mierzących znajomość słownictwa oraz podstawowe umiejętności matematyczne. W ciągu kolejnych lat 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. Ostatnia znacząca zmiana wprowadzona została w 1994 roku i w takim kształcie test przeprowadzany jest do dzisiaj.

Obecnie na rozwiązanie całego testu uczeń ma trzy godziny i czterdzieści pięć minut. Test składa się z 9 sekcji testowych i jednej sekcji zrównującej (eksperymentalnej). Trzy sekcje mierzą umiejętność czytania ze zrozumieniem (67 pytań). Kolejne trzy umiejętności matematyczne (54 pytań), a następne umiejętności wypowiedzi pisemnej (49 pytań). Sekcja zrównująca 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 prawidłową odpowiedź uczniowie uzyskują jeden punkt, za błędną odpowiedź w zadaniach zamkniętych punkty ujemne: -1/4 w zadaniach z czterema możliwościami wyboru, -1/3 z trzema możliwościami wyboru i -1/2 z dwoma możliwościami wyboru. Wyniki są skalowane i zrównywane metodą ekwicentylową i przedstawiane na jednej zagregowanej skali z przedziału 600-2400 oraz trzech osobnych skalach: dla czytania ze zrozumieniem, matematyki i pisania z przedziału 200-800. Osobno podaje się również oceny esejów znajdujących się w teście umiejętności pisania. Warto tutaj podkreślić, iż zarówno w SAT, jak i ACT (test omawiany w drugiej kolejności) każdy esej sprawdzany jest niezależnie przez 2 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 już w 1941 roku. Od tej chwili każda nowa wersja testu zawierała około 20% pytań z poprzedniej edycji. Wyniki kolejnych edycji zrównywane były roku do roku. Średnią skali ustalono na 500 dla roku 1941 (w 1995 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 nierównoważonych grup z testem kotwiczącym: *nonequivalent groups with anchor test (NEAT)*

*design*. Uproszczony schemat takiego zrównywania przedstawiony został w tabeli 1. W przedstawionym schemacie test A zrównywany jest za pomocą sekcji zrównującej („sekcja zrów.") z testem B. Sekcja zrównująca testu A w teście B traktowana jest jako jedna z 9 sekcji testowych. Test B ma za to inną sekcję zrównującą, za pomocą której można go zrównać z testem C, w którym to owa sekcja traktowana jest już jako pełnoprawna sekcja testowa (Rao i Sinharay, 2007).

**Tabela 1. Zrównywanie za pomocą planu nierównoważonych grup z testem kotwiczącym**

| Sesja/ populacja | Zadania testowe | | | | | |
|---|---|---|---|---|---|---|
| 1 | Test A | | *sekcja zrów.* | | | |
| 2 | | | Test B | | *sekcja zrów.* | |
| 3 | | | | | Test C | *sekcja zrów.* |

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.

## ACT (American College Testing)

ACT jest drugim (po SAT) najpopularniejszym testem mierzącym osiągnięcia szkolne uczniów po szkole średniej. Pierwszy raz przeprowadzony został w 1959 roku. Skonstruowany został, jako odpowiedź na test SAT, przez wybitnego teoretyka pomiaru Everetta Franklina Lindquista. ACT do 2005 roku mierzył 4 dziedziny wiedzy: umiejętność posługiwania się językiem angielskim, matematykę, czytanie oraz rozumowanie w naukach przyrodniczych. Wśród teoretyków pomiaru panuje przekonanie, iż zadania w teście ACT są łatwiejsze niż w SAT, lecz czasu na ich rozwiązanie jest znacznie mniej. Uczniowie mają 45 minut na zapoznanie się i rozwiązanie 75 zadań z języka angielskiego, 60 minut na 60 pytań z matematyki, 35 minut na rozwiązanie 40 zadań mierzących umiejętności 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, tym samym każde zadanie ma taką samą wagę przy szacowaniu skali wyników. W przeciwieństwie do SAT nie ma też żadnych punktów ujemnych. Skalowanie 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 oraz rozumowanie w naukach przyrodniczych. Wyniki z poszczególnych przedmiotów przedstawiane są na skali od 1 do 18. Wynik całościowy jest średnią z 4 podtestów. Uczniowie, którzy decydują się na test mierzący umiejętności pisania, otrzymują wynik na skali 2-12 oraz od 1 do 4 komentarzy. Wynik z testu pisania nie jest obowiązkowy i nie liczy się do sumarycznego wyniku.

Zrównywanie odbywa się na podstawie schematu zewnętrznej kotwicy z ekwiwalentnymi grupami (uproszczony schemat tego zrównania przedstawiony został w tabeli 2.). Aby przeprowadzić zrównanie dwóch testów z różnych lat, w przypadku ACT losowana jest reprezentacyjna próba losowa (na rysunku zaznaczona jako populacja zrównująca). W procesie zrównywania jedną formę rozwiązuje więcej niż 2000 uczniów. Uczniowie w tej próbie dostają kilka nowych form egzaminacyjnych oraz jedną wcześniej już zrównaną. Jako że w populacji zrównującej uczniowie rozwiązywali zadania z testu wcześniej przeprowadzonego (Test A) oraz zadania z testów, które dopiero mają się odbyć w kolejnych sesjach (Test B i C), możliwe jest zrównanie wyników z testu już przeprowadzonego (Test A) z testami, które zostaną wykorzystane w przyszłości. W teście ACT do zrównywania używana jest metoda ekwicentylowa wykorzystująca analityczne metody wygładzania rozkładów (Kolen 1984, Rao i Sinharay 2007).

**Tabela 2. Zrównywanie za pomocą planu zewnętrznej kotwicy z ekwiwalentnymi grupami**

<table>
<tr><td>Sesja/<br>populacja</td><td colspan="6">Zadania testowe</td></tr>
<tr><td>1</td><td colspan="2" rowspan="2">Test A</td><td colspan="2"></td><td colspan="2"></td></tr>
<tr><td>zrównująca</td><td colspan="2">Test A</td><td colspan="2">Test B</td><td colspan="2">Test C</td></tr>
<tr><td rowspan="2">2</td><td colspan="2"></td><td colspan="2" rowspan="2">Test B</td><td colspan="2"></td></tr>
<tr><td colspan="2"></td><td colspan="2"></td></tr>
<tr><td rowspan="2">3</td><td></td><td></td><td></td><td></td><td colspan="2" rowspan="2">Test C</td></tr>
<tr><td></td><td></td><td></td><td></td></tr>
</table>

## Izrael (*Psychometric Entrance Test*)

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

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 60 sekund na zadania dotyczące rozumowania ilościowego. PET jest podzielony na 8 sekcji. Każda sekcja trwa 30 minut (razem 3 godziny 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). Co łącznie daje 164 zadania (Allalouf, 1998).

Schemat zrównywania jest analogiczny jak w przypadku amerykańskiego SAT. W danej administracji izraelscy uczniowie piszą 6 takich samych sekcji testowych, ale za to różne sekcje zrównujące, tym samym wykorzystywane jest zrównywanie za pomocą planu nierównoważonych 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 form. Dla każdej formy i dla każdej umiejętności wykorzystuje się proste liniowe zrównywanie (Beller, 1994).

## Szwecja (*Swedish Scholastic Assesment Test*)

Egzamin, którego wynik decyduje o przyjęciu na szwedzkie uczelnie wyższe, powszechnie nazywany *SweSAT* (*Swedish Scholastic Assesment Test*) wprowadzony został w 1977 roku. Na początku przeznaczony był dla kandydatów na studia, którzy zdecydowali się na nie aplikować po osiągnięciu 25. roku życia, o przyjęciu młodszych uczniów 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); czytanie 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); ogólna wiedza (30 zadań, na które przeznaczono 25 minut). Wszystkie zadania w teście są zadaniami zamkniętymi i punktowane są na skali 0-1. Cały test trwa cztery godziny i dziesięć minut (Stage i Igren, 2002).

Surowy wynik skalowany jest za pomocą metody ekwicentylowej i przekształcany jest na skale z przedziału 0,0 do 2,0. Test zrównywany jest przy założeniu, iż populacje z roku na rok nie zmieniają się. 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[1] (Stage, 2004).

Od roku 1997 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.

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

## Canada – Ontario (*EQAO tests*)

W Kanadzie nie istnieje jeden ogólnokrajowy system egzaminacyjny, jednak poszczególne prowincje prowadzą systemy ewaluacyjne i testują swoich uczniów. Przykładem takiej prowincji jest Ontario. W 1996 roku 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 (trzecia i szósta klasa). Uczniowie 9 klasy rozwiązują rozbudowany test osiągnięć szkolnych w zakresie matematyki, a uczniowie po 11. roku nauki również rozbudowany test mierzący umiejętności czytania ze zrozumieniem oraz 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 podchodzenia do testów, lecz w większości przypadków również 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 (testy mierzące umiejętność posługiwania się językiem szwedzkim. 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 dziewiątej klasie 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 szwedzkim 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. Skalowanie odbywa się za pomocą trzyparametrycznego modelu IRT (3PL). Obok wyniku na skali przydzielane są oceny odzwierciedlające stopień opanowania przez ucznia danych umiejętności.

Do zrównywania wykorzystuje się schemat analogiczny jak w przypadku amerykańskiego testu SAT: plan nierównoważonych grup z testem kotwiczącym. Z tym, że do procedury zrównywania wykorzystano trzyparametryczny model IRT (3PL). Zadania, które znajdują się zarówno w bieżącym teście, jak i we wcześniejszej administracji są estymowane w taki sposób, by ich parametry były zgodne z estymacją we wcześniejszej administracji testu. W procesie estymacji testu bieżącego zadania, które stanowią kotwice, nie są *de facto* estymowane, tylko przyjmują wartości parametrów estymowanych we wcześniejszej administracji. Warto również zwrócić uwagę, iż test zrównywany jest również pionowo (*vertical saling*), czyli wyniki uczniów z różnych poziomów kształcenia są bezpośrednio porównywalne.

## Australia (*National Assessment Program – Literacy and Numeracy*)

W roku 2008 w Australii po raz pierwszy przeprowadzony został ogólnokrajowy egzamin mierzący umiejętności językowe oraz matematyczne. NAPLAN (*National Assessment Program – Literacy and Numeracy*). Test jest obowiązkowy dla wszystkich uczniów i przeprowadzany jest w klasie 3, 5, 7 oraz 9 (Freedman, 2009). Szczegółowy plan testu NAPLAN przedstawiony został w tabeli 3.

**Tabela 3. 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<br>a) Gramatyka<br>b) Ortografia | 25 zadań (40 min)<br>23 zadania (40 min) | 25 zadań (40 min)<br>23 zadania (40 min) | 30 zadań (45 min)<br>24 zadania (40 min) | 26 zadań (45 min)<br>28 zadań (40 min) |
| Matematyka<br>a) bez kalkulatora<br>b) z kalkulatorem | 35 zadań (45 min)<br>--- | 40 zadań (50 min)<br>--- | 32 zadania (40 min)<br>32 zadania (40 min) | 32 zadania (40 min)<br>32 zadania (40 min) |

Wyniki z każdego testu skalowane są za pomocom modelu Rascha. Wyniki uczniów oraz szkół generowane są metodą *weighted likelihood estimates* (WLE), a następnie przekształcane do skali o średniej 500 i odchyleniu standardowym 100. Wyniki na poziomie poszczególnych stanów oraz wyniki ogólnonarodowe uzyskiwane są dzięki metodologii *plausible values*.

W procedurze zrównywania w NAPLAN przyjęto schemat zewnętrznej kotwicy z równoważnymi grupami (tak jak w amerykańskim teście ACT). Każdego roku prowadzone jest tak zwane studium zrównujące, w którym bierze udział losowa próbka studentów, którzy również podchodzą do testu NAPLAN w danym roku. Studium zrównujące odbywa się tydzień po testowaniu zasadniczym. W teście zrównującym znajdują się zadania, które pozwalają na zrównanie pytań z bieżącej administracji z wcześniejszymi administracjami 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 one zgodne z estymacją we wcześniejszej administracji testu. Warto również nadmienić, iż podobnie jak testy w Ontario, oprócz zrównywania pomiędzy edycjami egzaminu z kolejnych lat, NAPLAN zrównywany jest też pionowo (*vertical saling*), czyli wyniki uczniów z różnych poziomów kształcenia są bezpośrednio porównywalne (Cook, 2009).

## Podsumowanie

Najczęściej spotykanymi schematami zrównywania w systemach egzaminacyjnych są: plan nierównoważonych grup z testem kotwiczącym oraz plan zewnętrznej kotwicy z ekwiwalentnymi grupami. Obydwa plany na dużą skalę użyte zostały po raz pierwszy w Stanach Zjednoczonych. Pierwszy z nich

w SAT drugi w ACT. Jeżeli plan testowania nie zakłada powtórnego użycia zadań testowych, tak jak w Szwecji, poszczególne administracje testów 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 zarówno metod opartych na metodach ekwicentylowych, jak i na modelowaniu IRT. Przy czym metody ekwicentylowe stosowane są w starszych systemach egzaminacyjnych, gdzie nie bez znaczenia jest ciągłość stosowanej metody; w młodszych systemach egzaminów zewnętrznych chętnie sięga się po modele IRT. W każdym z przywoływanych systemów egzaminacyjnych wyniki są skalowane (tj. uczniom nie są komunikowane ich wyniki surowe). Charakterystyczne jest to, iż większość testów wysokiej stawki składa się z dużej, w porównaniu z polskimi warunkami, liczby pytań. Patrząc na krótkie charakterystyki testów przedstawionych w tym artykule, czytelnik znający polski system egzaminacyjny musi stwierdzić, iż daleko naszemu rodzimemu systemowi egzaminacyjnemu do międzynarodowych standardów.

**Bibliografia:**

1. Allalouf, A. i G. Ben Shakhar (1998). „The effect of coaching on the predictive validity of scholastic aptitude tests". *Journal of Educational Measurement* 35 (1): 31-47.
2. Beller, M. (1994). „Psychometric and social issues in admissions to Israeli universities". *Educational Measurement: issues and practice* 13 (2): 12-20.
3. 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.
4. EQAO (2011). *EQAO's Technical Report for the 2009–2010 Assessments*. Toronto.
5. Freeman, C. (2009). „First national literacy and numeracy tests introduced". *Research Developments* 20 (20).
6. Kolen, M. J. (1984). „Effectiveness of analytic smoothing in equipercentile equating". *Journal of Educational Statistics* 9, 25-44.
7. Rao, C. R. i S. Sinharay. (2007). *Psychometrics.* 26-ed. North Holland.
8. Stage, C. (2004). Notes from the Tenth International SweSAT Conference. Umeå, June 1–3, 2004.
9. Stage, C. i G. Ígren (2002). *The Swedish Scholastic Assessment Test (SweSAT)*. Deptartment of Educational Measurement, Ume Univ.