Edukacja 2014, 2(127), 85–107
ISSN 0239-6858

Testy osiągnięć szkolnych TOS3: przykład narzędzia skonstruowanego z wykorzystaniem modelu Rascha

Aleksandra Jasińska, Michał Modzelewski
Instytut Badań Edukacyjnych*

Większość stosowanych w Polsce testów osiągnięć szkolnych pozbawionych jest mocnego uzasadnienia swojej trafności w postaci szczegółowej dokumentacji. Sytuacja ta wpływa negatywnie na rozwój metodologii konstrukcji tych narzędzi. Artykuł stanowi opis zestawu trzech standaryzowanych testów osiągnięć szkolnych TOS3 wykorzystanych w ramach dwóch badań. Testy te służą pomiarowi osiągnięć szkolnych z obszaru edukacji polonistycznej i matematycznej uczniów kończących I etap edukacyjny. W artykule przedstawiono proces konstrukcji testów osiągnięć z wykorzystaniem modelu Rascha (szczególnego przypadku jednoparametrycznego modelu IRT). Udokumentowano także trafność i rzetelność TOS3, wykorzystując wyniki dwóch reprezentatywnych badań ($N > 5000$). Artykuł pokazuje korzyści wynikające z wykorzystania modelu pomiarowego podczas budowy narzędzi. Opisane doświadczenia mogą być źródłem wskazówek dla twórców przyszłych testów osiągnięć szkolnych w Polsce.

Słowa kluczowe: pomiar dydaktyczny, pomiar osiągnięć szkolnych, psychometria, IRT, model Rascha, konstrukcja testów, rzetelność, trafność, TOS3.

Pierwsze lata nauki są powszechnie uznawane za kluczowe dla przyszłego sukcesu edukacyjnego dzieci (Aubrey, Godfrey i Dahl, 2006; Boland, 1993; Slavin, Karweit i Wasik, 1992). Możliwość rzetelnego opisania umiejętności uczniów na początkowych etapach kształcenia jest bardzo ważna – na poziomie konkretnej szkoły może ona sprzyjać indywidualizacji procesu nauczania, na poziomie samorządowym lub centralnym, może wspomagać prowadzenie skutecznej polityki edukacyjnej. Testy przeznaczone do takiej diagnozy muszą posiadać odpowiednie własności psychometryczne – ich trafność powinna być dobrze udokumentowana, a precyzja pomiaru wysoka.

Do niedawna jedyną możliwością wglądu w sytuację początkowych etapów edukacji w skali ogólnopolskiej był system egzaminów zewnętrznych. Od 2002 r. co roku przeprowadzany jest *Sprawdzian w szóstej klasie*, potocznie zwany „sprawdzianem szóstoklasisty". Nie jest to jednak wgląd satysfakcjonujący; *Sprawdzian* dostarcza informacji

Artykuł powstał w ramach projektów systemowych „Badanie jakości i efektywności edukacji oraz instytucjonalizacja zaplecza badawczego" oraz „Rozwój metody edukacyjnej wartości dodanej na potrzeby wzmocnienia ewaluacyjnej funkcji egzaminów zewnętrznych" prowadzonych w Instytucie Badań Edukacyjnych i współfinansowanych ze środków Europejskiego Funduszu Społecznego (Program Operacyjny Kapitał Ludzki 2007–2013, priorytet III: Wysoka jakość systemu oświaty).

* Adres do korespondencji: ul. Górczewska 8, 01-180 Warszawa. E-mail: a.jasinska@ibe.edu.pl

dopiero po zakończeniu sześciu lat edukacji, podzielonych na dwa trzyletnie cykle kształcenia.

W ciągu ostatnich kilku lat sytuacja w obszarze stosowanych w Polsce testów osiągnięć szkolnych (przeznaczonych dla szkół podstawowych) znacząco się zmieniła. Obok systemu egzaminacyjnego można wyróżnić jeszcze dwie jego gałęzie – sektor związany z badaniami edukacyjnymi oraz sektor prywatny, związany z działalnością wydawnictw edukacyjnych.

W latach 2006–2013 odbyło się kilka prowadzonych na szeroką skalę badań, których elementem był pomiar osiągnięć uczniów szkół podstawowych. W latach 2006–2011 odbyły się cztery ogólnopolskie *Badania umiejętności podstawowych uczniów szkół podstawowych* przeprowadzone w ramach projektu „Badanie umiejętności podstawowych uczniów trzeciej klasy szkoły podstawowej". W 2011 r. Polska wzięła udział w dwóch badaniach międzynarodowych przeznaczonych dla uczniów klas trzecich: badaniu kompetencji w dziedzinie matematyki i nauk przyrodniczych (TIMSS; Konarzewski, 2012) oraz po raz drugi (pierwszy raz w 2006 r.) w badaniu umiejętności czytania (PIRLS; Konarzewski, 2012). W 2009 r. rozpoczęło się podłużne *Badanie uwarunkowań wyników nauczania w szkołach podstawowych* prowadzone w ramach projektu „Badania dotyczące rozwoju metodologii szacowania wskaźnika edukacyjnej wartości dodanej (EWD)". Rok później rozpoczęło się, również podłużne, *Badanie szkolnych uwarunkowań efektywności kształcenia* (SUEK). Wszystkie one, dzięki wykorzystaniu różnych testów osiągnięć, pozwalają przyjrzeć się sytuacji w polskiej oświacie na przełomie I i II cyklu kształcenia. Obraz pierwszych lat nauki dopełnia *Badanie sześcio- i siedmiolatków na starcie szkolnym*, przeprowadzone w latach 2012–2013, będące diagnozą umiejętności polskich uczniów na progu edukacji szkolnej. Na uwagę zasługuje w szczególności fakt, że w ramach tego badania wykorzystano po raz pierwszy *Test umiejętności na starcie szkolnym* (TUNSS), jedyny, jak do tej pory, komputerowy, adaptatywny test osiągnięć szkolnych w Polsce (Karwowski i Dziedziewicz, 2012).

Wzrostowi zainteresowania badaczy oceną osiągnięć szkolnych uczniów towarzyszy rosnąca potrzeba przeprowadzania jakiejś formy diagnozy, przed którą stoją same szkoły. O jej nasileniu świadczy m.in. wysoki odsetek – 70–80%, w zależności od edycji (Pregler, 2013; Pregler i Wiatrak, 2011; 2012) – szkół podstawowych dobrowolnie uczestniczących w corocznym *Ogólnopolskim badaniu umiejętności trzecioklasistów* (OBUT), powadzonym w latach 2011–2014. Tezę tę potwierdzają także sukcesy komercyjnych, cyklicznych programów diagnostycznych, prowadzonych przez wydawnictwa edukacyjne[1].

Choć na potrzeby egzaminów zewnętrznych, badań naukowych i komercyjnych programów diagnostycznych konstruowane są liczne narzędzia służące do pomiaru osiągnięć szkolnych, to tylko nieliczne z nich posiadają dokumentację pozwalającą ocenić ich jakość. Wysoki standard w tym zakresie wyznaczają międzynarodowe badania osiągnięć, takie jak PIRLS i TIMSS, w ramach których powstają rozbudowane raporty techniczne (np. Martin i Mullis, 2013). Podręcznik dokumentujący proces tworzenia narzędzia oraz własności psychometryczne posiada również wspomniany wcześniej TUNSS (Karwowski i Dziedziewicz, 2012). Zdecydowana większość narzędzi służących do pomiaru osiągnięć szkolnych nie posiada, niestety, takiej dokumentacji. Można wyróżnić dwa przykłady negatywnych konsekwencji tego stanu

[1] Na przykład *Ogólnopolski sprawdzian trzecioklasisty z Operonem*, prowadzony od 2008 r., czy uruchomiony rok później *Ogólnopolski próbny sprawdzian szóstoklasisty*. Oba programy diagnostyczne prowadzone są przez Wydawnictwo Pedagogiczne Operon.

rzeczy. Z jednej strony zasadność wniosków wyciąganych na podstawie wyników dostarczonych przez nieudokumentowane testy osiągnięć może być zakwestionowana. Z drugiej strony, nie następuje rozwój metodologii konstrukcji testów osiągnięć szkolnych w naszym kraju – tworząc narzędzia badacze w Polsce, są zmuszeni do zaczynania „od zera", brak bowiem narastania doświadczenia w tym obszarze.

Niniejszy artykuł odnosi się do obu tych kwestii. Po pierwsze, stanowi opis narzędzia wykorzystanego w badaniach SUEK i EWD – czyli zestawu trzech testów osiągnięć szkolnych (TOS3), przeznaczonych dla uczniów kończących I etap edukacyjny (testu umiejętności czytania, świadomości językowej i umiejętności matematycznych). Po drugie, wraz z kwestiami poruszanymi w innym artykule (Jasińska i Modzelewski, 2012), stanowi szczegółowy zapis doświadczeń w zakresie konstrukcji narzędzi służących do pomiaru osiągnięć szkolnych z obszaru edukacji polonistycznej oraz matematycznej dla początkowych etapów nauczania.

Etapy konstrukcji testów osiągnięć szkolnych

Konstrukcja testów osiągnięć jest bardzo wymagającym przedsięwzięciem zarówno pod względem logistycznym, jak i teoretycznym. Dokładne zaplanowanie potrzebnych do podjęcia działań, a następnie przestrzeganie ustalonej procedury jest niezbędne w celu kontroli czynników, które mogłyby negatywnie wpłynąć na trafność tworzonego narzędzia – zarówno na jej aspekt treściowy, jak i kryterialny (Downing, 2006b; Kane, 2006).

Zgodnie z zaleceniami dotyczącymi tworzenia testów (Downing, 2006b) prace nad TOS3 rozpoczęto od określenia ogólnych założeń związanych z planowanym pomiarem – zdefiniowano badaną populację, przyjęto teorię pomiaru, zgodnie z którą miały zostać konstruowane narzędzia, a także określono główne umiejętności, których miał dotyczyć pomiar.

Badana populacja i cel pomiaru

Pierwsza wersja TOS3 powstała na potrzeby badania SUEK; w związku z jego celami oraz harmonogramem badana populacja została określona jako uczniowie, którzy zakończyli naukę na I etapie edukacyjnym (tj. znajdują się na początku klasy IV szkoły podstawowej) w roku szkolnym 2011/2012. Byli oni ostatnim rocznikiem kształconym według poprzedniej podstawy programowej, określonej w rozporządzeniu Ministra Edukacji Narodowej i Sportu z dnia 26 lutego 2002 r. (Dz. U. Nr 51, poz. 458, z późn. zm.). W badaniu EWD wzięli natomiast udział uczniowie, którzy zostali objęci reformą programową, co wymagało dostosowania testów do innej populacji. Szczegóły tej procedury zostały opisane w dalszej części artykułu.

Charakter badania podłużnego SUEK wymagał tego, by skonstruowane na jego potrzeby testy dostarczały jak najbardziej precyzyjnych wyników dla szerokiego zakresu umiejętności badanej populacji (tzw. testy szerokiego zasięgu)[2]. Z uwagi na to, że najbardziej interesującym poziomem analizy z punktu widzenia badania SUEK był poziom oddziałów szkolnych, priorytetem było rzetelne określenie efektów kształcenia właśnie na poziomie oddziałów. Osiągnięcie tak postawionych celów wymagało dobrania do testu zadań o trudnościach z całego zakresu skali.

Teoria pomiaru

Na potrzeby konstrukcji narzędzi zdecydowano się na zastosowanie teorii odpowiedzi

[2] Innymi typami testów są testy przesiewowe, które pozwalają precyzyjnie stwierdzić, czy uczeń osiągnął minimalny, ustalony przez ekspertów poziom umiejętności. Natomiast testy selekcyjne służą do jak najdokładniejszego zróżnicowania wyników uczniów o wysokich umiejętnościach (Jakubowski i Pokropek, 2009).

na zadanie (*item response theory*, IRT), a dokładniej modelu Rascha, jako podstawy teoretycznej pomiaru (Rasch, 1960, zob. też Kondratek i Pokropek, 2013). Na ten wybór wpłynęło wiele czynników. Po pierwsze, IRT pozwala pełniej opisać relacje między poziomem umiejętności ucznia a prawdopodobieństwem udzielenia przez niego poprawnej odpowiedzi na zadanie, niż ma to miejsce w klasycznej teorii testów (KTT). Dzięki temu, że IRT pozwala na szczegółowy opis właściwości psychometrycznych poszczególnych zadań, możliwe jest zastosowanie różnorodnych kryteriów ich doboru do ostatecznej wersji testu. Po drugie, w ramach IRT można określić pożądane właściwości pomiarowe dla docelowego testu w dowolnych zakresach skali, wykorzystując do tego celu krzywą informacyjną testu. Po trzecie, IRT umożliwia zastosowanie skomplikowanych schematów badawczych, w tym opisanego dalej próbkowania macierzowego, bez konieczności uciekania się do złożonych schematów zrównywania podczas analizy danych. Spośród dostępnych modeli IRT model Rascha jest najbardziej restrykcyjny. Zakłada bowiem, że wszystkie zadania mają równie dobre właściwości pomiarowe. Sprawia to, że surowa liczba punktów stanowi statystykę dostateczną do oszacowania poziomu umiejętności uczniów (Ayala, 2009). Zastosowanie modelu Rascha pozwala zatem na stworzenie narzędzia, dla którego wyniki surowe można łatwo przeliczyć na wyskalowane za pomocą tablicy przeliczeniowej, co sprzyja komunikowalności rezultatów pomiaru. Oczywiście będzie to możliwe, jeśli model z powodzeniem uda się dopasować do danych.

Forma narzędzia

Z uwagi na koszty oraz łatwość wykorzystania powstałych narzędzi, zdecydowano się na konstrukcję testów w formie papierowej, do samodzielnego wypełniania w ramach badania audytoryjnego. Określono także czas potrzebny na rozwiązanie ostatecznych wersji narzędzi, co w połączeniu z założeniem, że miały one być testami mocy[3], pozwoliło wstępnie przewidzieć potrzebną liczbę zadań w testach. Zdecydowano się także na stworzenie dwóch równoległych wersji testów z pulą zadań kotwiczących. Obie wersje były administrowane równomiernie w taki sposób, że połowa uczniów w oddziale rozwiązywała jedną wersję, a połowa drugą. Miało to, po pierwsze, na celu zmniejszenie błędu pomiaru związanego z odpisywaniem przez uczniów odpowiedzi. Po drugie, umożliwiło wydłużenie testu w celu lepszego pokrycia zadaniami treści i umiejętności szczegółowych nauczanych na danym etapie kształcenia. Zabiegi te pozwoliły na bardziej trafne oszacowanie wyników na poziomie oddziałów.

Koncepcja skal pomiarowych

Opis badanych umiejętności znajduje się w dalszej części artykułu. W tym miejscu warto wspomnieć, że na etapie projektowania narzędzi szczegółowo określono treści i umiejętności definiujące badane konstrukty. Podczas tego etapu skupiono się na analizie podstawy programowej kształcenia ogólnego, założeniach krajowych, międzynarodowych i zagranicznych badań umiejętności oraz na analizie wniosków płynących z projektu poświęconego nowej formule sprawdzianu dla klasy VI[4]. Opracowana koncepcja skal była przedmiotem konsultacji merytorycznych z ekspertami oraz została poddana zewnętrznej recenzji. Na podstawie koncepcji skal pomiarowych

[3] Test mocy jest testem, w którym czas rozwiązywania zadania nie jest powiązany z prawdopodobieństwem udzielenia przez ucznia prawidłowej odpowiedzi. W związku z tym, czas przewidziany na rozwiązanie testu mocy jest na tyle długi, by wyeliminować wpływ jego braku na udzielane odpowiedzi uczniów.

[4] „Nowa formuła sprawdzianu w klasie VI", projekt realizowany w latach 2007–2010 przez Centralną Komisję Egzaminacyjna, koordynator: Anna Pregler.

przygotowano plany testów, precyzujące, ile zadań mierzących poszczególne umiejętności szczegółowe powinno znaleźć się w teście. Były one gwarancją trafności treściowej docelowych narzędzi. Plany ostatecznej wersji testów, dostosowanych do populacji badania EWD, przedstawiono w Aneksie do artykułu.

Przygotowanie zadań do badania pilotażowego

Kolejne etapy związane były z przygotowaniem dużej puli zadań, spośród których wybrane miały zostać zadania do badania pilotażowego. Najpierw przygotowano wskazówki dla autorów zadań dotyczące ich konstrukcji oraz schematów oceniania. Wykorzystano do tego doświadczenia amerykańskie w zakresie konstrukcji testów (Downing, 2006a; Haladyna, Downing i Rodriguez, 2002). Szczególny nacisk położono na wymóg mówiący o tym, że pozyskiwane zadania mają być od siebie niezależne. Wykluczano sytuację, w której rozwiązanie jakiegoś zadania mogło być uzależnione od poprawnego rozwiązania innego zadania. Było to istotne z punktu widzenia wykorzystywanego modelu analizy danych, w którym przyjmuje się założenie o lokalnej niezależności poszczególnych pozycji testowych. Należy przy tym zauważyć, że zadania pogrupowane w wiązki (np. poprzez wspólne polecenie lub tekst, do którego się odnoszą) niekoniecznie łamią założenie o lokalnej niezależności (Baghaei, 2008). W teście świadomości językowej oraz teście matematycznym dopuszczano zadania, które miały wspólne polecenie (np. prośbę o znalezienie synonimu lub rozwiązanie podanych działań), ale poszczególne przykłady nie mogły być powiązane. Test umiejętności czytania był jedynym, w którym z założenia grupa zadań zawsze odnosiła się do tego samego tekstu. W tym jednak wypadku trudno o inne, dające się zastosować w praktyce, rozwiązanie.

Równolegle trwała rekrutacja autorów, podczas której kandydaci byli proszeni o przygotowanie próbki kilku zadań, zgodnej z wytycznymi. Kandydaci, których zadania zostały najlepiej ocenione, zostali zaproszeni do dalszej współpracy, w trakcie której zlecano im do opracowania zadania odwołujące się do określonych treści i umiejętności (według opracowanych planów testów). Autorzy byli proszeni o układanie zadań z bardzo szerokiego zakresu umiejętności (od zadań bardzo łatwych do zadań bardzo trudnych dla docelowej populacji). Przesyłane przez autorów zadania były na bieżąco recenzowane i, w razie wykrycia niedoskonałości, odsyłane do dopracowania. Pozyskane w trakcie tego etapu zadania były poddane dodatkowej ocenie i poprawkom na specjalnie zorganizowanych warsztatach z udziałem matematyków, polonistów, pedagogów wczesnoszkolnych, dydaktyków praktyków, koderów oraz członków zespołu badawczego. Ostatnim etapem prac nad zadaniami i zeszytami pilotażowymi była ich obróbka graficzna i skład. Łącznie na potrzeby pilotażu przygotowano 823 zadania składające się na trzy skale pomiarowe.

Plan badania pilotażowego

W badaniu pilotażowym szczególny nacisk położono na dobór zadań do zeszytów testowych oraz stworzenie planu testowania (określającego, którzy uczniowie mają rozwiązywać które zeszyty testowe[5]), który zapewni zrównoważone próbkowanie macierzowe, umożliwiające wspólne skalibrowanie zadań z jednego testu oraz jak najdokładniejsze oszacowanie parametrów psychometrycznych zadań i każdego z trzech testów. W związku z harmonogramem badania SUEK zdecydowano, że badanie odbędzie się na losowej próbie wyłonionej

[5] Ponieważ zadań do przetestowania było bardzo dużo (były one pogrupowane w kilkanaście zeszytów testowych dla każdego testu), nie było możliwości, by każdy uczeń rozwiązał wszystkie zadania.

z dwóch populacji uczniów, odmiennych od docelowej – na uczniach klas III i V szkoły podstawowej[6]. Przetestowanie zadań na tak zdefiniowanej populacji nie stanowiło problemu, bowiem pozyskane zadania miały z założenia wykazywać się zróżnicowaną trudnością. W celu zwiększenia fasadowej trafności pomiaru, zadania bardzo łatwe (według przewidywań autorów) zostały zgrupowane w zeszytach rozwiązywanych tylko przez uczniów młodszych, natomiast zadania najtrudniejsze w zeszytach przeznaczonych wyłącznie dla uczniów klas V. Dzięki temu zminimalizowane zostało ryzyko, że na sposób odpowiedzi uczniów na zadania wpłynąć mogła postrzegana przez nich niedostosowana trudność zadań (np. zlekceważenie testu zbyt łatwego lub zbyt trudnego).

Ostatecznie na potrzeby badania pilotażowego stworzono 44 zeszyty testowe. W każdym zeszycie znalazło się średnio 19 zadań. Plan testowania został tak przygotowany, że każdy z zeszytów dla danego testu współwystępował z jak największą liczbą innych zeszytów dla jednej grupy uczniów. Równoważył także prawdopodobieństwo rozwiązywania zeszytów w badanej próbie uczniów. Dodatkowo plan testowania uwzględniał różną kolejność rozwiązywanych przez uczniów zeszytów podczas kolejnych sesji testowych (by uniknąć wpływu zmęczenia uczniów na oszacowanie parametrów zadań). Każdy z uczniów miał przewidziane do rozwiązania cztery zeszyty z wybranego testu, rozwiązywał więc zatem średnio 76 zadań. Każde zadanie było natomiast rozwiązywane przez co najmniej 427 uczniów.

Analiza danych i kryteria doboru zadań

Dzięki zastosowanemu planowi testowania możliwa była łączna kalibracja parametrów dla wszystkich zadań w ramach każdego z testów. Za pomocą programu ACER ConQuest 2.0 (Wu, Adams, Wilson i Haldane, 2007) dla każdego z testów osobno wykonano serię analiz polegających na dopasowaniu do danych jednowymiarowego modelu Rascha. Zadania wielokategorialne (oceniane na dłuższej skali niż zero-jedynkowa) analizowane były w ramach modelu *partial credit* (Masters i Wright, 1997). Do oszacowania parametrów wykorzystano estymator brzegowej najwyższej wiarygodności (*marginal maximum likelihood*, MML). Należy zauważyć, że ze względu na algorytmy obliczeniowe zaimplementowane w wykorzystanym oprogramowaniu, analizy nie uwzględniały pogrupowania uczniów na oddziały i szkoły. Uczniowie w analizach traktowani byli więc jako prosta próba losowa z dwóch populacji uczniów – trzecio- i piątoklasistów[7].

Jako miary dopasowania zadania wykorzystano miary *infit* oraz *outfit* raportowane przez program. Gdy miary te osiągają wartość 1, przyjmuje się, że zadanie jest dobrze dopasowane do modelu. W praktyce dopuszcza się jednak pewne rozchwianie wartości tych statystyk (Ayala, 2009). Zadanie uznawano za odpowiednio dopasowane, jeżeli miary te mieściły się w zakresie od 0,8 do 1,2. Dodatkowo wspierano się analizą empirycznych i teoretycznych krzywych charakterystycznych dla zadań. Na każdym kroku eliminowano kilka zadań najgorzej dopasowanych do modelu Rascha. Procedura analizy dopasowania zadania wielokategorialnego była jednak odmienna od tej stosowanej przy zadaniach ocenianych dychotomicznie.

[6] Zadania zostały więc poddane badaniu pilotażowemu na uczniach trochę młodszych i trochę starszych niż docelowa populacja. Badanie było realizowane na ogólnopolskiej losowej próbie 80 szkół podstawowych w klasach III i V (łącznie przebadano 281 klas, co dało 5454 uczniów) w roku szkolnym 2010/11. Okienko testowe było na początku II semestru.

[7] Dla każdej z grup modelowana była odrębna średnia rozkładu w populacji. Więcej o warunkowaniu rozkładów umiejętności w populacji można znaleźć w artykule Margaret Wu (2005).

W przypadku wykrycia słabego dopasowania dla zadania ocenianego na dłuższej skali punktowej najpierw podejmowano próbę jego polepszenia poprzez zmianę definicji wybranych kategorii punktowych. Modyfikacje skal punktowych były zaplanowane już na etapie konstrukcji kluczy kodowych, gdy dla wybranych zadań otwartych zdefiniowano kody opisujące różne typy potencjalnych rozwiązań. Na etapie analizy poszczególnym kodom przypisano wartości punktowe (najczęściej od 0 do 2) i sprawdzano, czy założenie o wzrastającym poziomie umiejętności wymaganym dla uzyskania wyższej kategorii punktowej znajduje odzwierciedlenie w danych. Jeżeli średni poziom umiejętności uczniów uzyskujących daną kategorię punktową nie był rozróżnialny od tego, który osiągali uczniowie o wyższej lub niższej kategorii punktowej, to na podstawie analizy jakościowej zapisów opisujących tę kategorię punktową decydowano się na modyfikację przypisania punktów do kodów lub skróceniu skali punktowej, np. z 0–1–2 do 0–1. Decyzja o tym, które kategorie punktowe połączyć, była podejmowana indywidualnie dla każdego zadania. Po wprowadzeniu modyfikacji analizę powtarzano.

Na ocenę przydatności zadania do ostatecznej wersji narzędzia miała także wpływ jego trudność, sposób oddziaływania dystraktorów (w przypadku zadań zamkniętych) oraz wyniki dwóch analiz zróżnicowanego funkcjonowania zadania (*differential item functioning*, DIF; Kondratek i Grudniewska, 2013). Zarówno zadania zbyt łatwe, jak i zbyt trudne w odniesieniu do zakładanego poziomu umiejętności populacji badanych, były odrzucane. Analiza dystraktorów polegała na weryfikacji założenia o negatywnym związku danego dystraktora z poziomem umiejętności (wraz ze wzrostem umiejętności prawdopodobieństwo wybrania dystraktora powinno maleć) oraz sprawdzeniu, czy dana niepoprawna odpowiedź nie jest dystraktorem „martwym" (niewybieranym przez uczniów). Spośród zadań zamkniętych preferowane były te, których dystraktory nie budziły wątpliwości odnośnie do tych kryteriów.

Wykonano analizy sprawdzające zadania pod kątem efektu DIF ze względu na etap kształcenia (III lub V klasa) oraz płeć, usuwając zadania, w których (a) zaobserwowano zróżnicowane funkcjonowanie zadania w ramach wyróżnionych grup oraz (b) źródło tych różnic dało się przypisać do treści lub formy zadania na etapie analizy jakościowej.

Efektem wieloetapowych analiz było wyłonienie trzech zestawów zadań – dla umiejętności czytania pozostały 74 zadania (ok. 40% zadań poddanych pilotażowi), dla świadomości językowej 123 zadania (ok. 40%) oraz dla umiejętności matematycznych 181 zadań (ok. 56%). Spośród tych zestawów, zgodnie z założonymi planami testu, przygotowano ostateczne narzędzia wykorzystane w badaniu SUEK, tj. pierwszą wersję TOS3.

Wykorzystanie TOS3 w badaniu SUEK

Warunki pilotażu stanowiły swoisty test założeń stojących za modelem Rascha – badani uczniowie reprezentowali, względem docelowej populacji, dwa krańce spektrum umiejętności. Jedną z pożądanych właściwości modelu Rascha jest fakt, że uporządkowanie zadań ze względu na trudność jest takie samo dla uczniów, niezależnie od ich poziomu umiejętności. Pozwala to z jednej strony przewidywać, jak poradzą sobie z zadaniami uczniowie o odmiennym poziomie umiejętności niż uczniowie badani. Z drugiej zaś, umożliwia zweryfikowanie fundamentalnego dla TOS3 założenia, mówiącego o tym, że badany zakres umiejętności szkolnych jest rzeczywiście rozwijany w toku nauki szkolnej. Jeżeli wybrane na podstawie danych z pilotażu zadania faktycznie spełniają założenia

modelu Rascha, to powinny również sprawdzić się na etapie badania zasadniczego, gdy rozwiązywali je uczniowie rozpoczynający naukę w klasie IV. Analizy przeprowadzone na podstawie wyników badania zasadniczego potwierdziły jakość przygotowanych narzędzi (Jasińska i Modzelewski, 2012). Testy okazały się dopasowane do rzeczywistego poziomu umiejętności uczniów w tym wieku, a wybrane zadania potwierdziły swoje właściwości pomiarowe (udało się uzyskać dobre dopasowanie do modelu Rascha).

Zmiany w TOS3 na potrzeby badania EWD

Uczniowie uczestniczący w badaniu SUEK nie byli jeszcze objęci reformą programową[8]. Dlatego pierwsza wersja testów została dostosowana do wytycznych poprzedniej podstawy programowej. Całe narzędzie miało być jednak wykorzystane także w badaniu EWD, w którym populacja uczniów objęta badaniem była pierwszym rocznikiem kształconym według nowej podstawy programowej. W celu wykorzystania testów TOS3 w badaniu EWD sprawdzono, czy zadania składające się na narzędzie pomiarowe są zgodne z nowymi wytycznymi.

W przypadku testu umiejętności czytania oraz testu świadomości językowej stwierdzono, że wszystkie zadania są zgodne także z nową podstawą programową. Biorąc jednak pod uwagę fakt, że test świadomości językowej dla uczniów z badania SUEK okazał się testem umiarkowanie trudnym, a także był testem o najmniejszej liczbie pozycji testowych (Jasińska i Modzelewski, 2012), zdecydowano się uzupełnić o dodatkowe zadania. Z banku przetestowanych na tę okoliczność zadań[9] wybrano dwa łatwiejsze, o dobrych właściwościach psychometrycznych, odwołujące się do umiejętności słabiej reprezentowanych w pierwotnym teście.

Analiza testu matematycznego pokazała, że cztery zadania odwoływały się do treści, które nie zostały wyszczególnione w nowej podstawie programowej (porównywanie ilorazowe, pojęcie pola figur płaskich, rozpoznawanie wielokątów). Zadania te zastąpiono czterema innymi (wybrano je ze wspomnianego banku zadań), które odwoływały się do tych samych nadrzędnych grup treści, ale nie wymagały operowania pojęciami czy wiadomościami, które mogły nie zostać wprowadzone na lekcjach, z uwagi na wymagania nowej podstawy programowej.

Zmiany dokonane w testach były niewielkie, jednak biorąc pod uwagę także to, że testy TOS3 w badaniu EWD były rozwiązywane przez uczniów nauczanych zgodnie z inną podstawą programową, nie można przyjąć za pewne, że wymienione czynniki nie spowodowały zmiany mierzonych testami skal. Przez skalę można rozumieć uporządkowanie, zarówno względne, jak i bezwzględne, zadań mierzących daną umiejętność pod względem trudności.

Co mogłaby oznaczać zmiana w hierarchii zadań pomiędzy badaniami SUEK a EWD? Pierwszym z możliwych wyjaśnień jest oczywiście błąd pomiaru związany z procedurami realizacji badania. Mielibyśmy z nim na przykład do czynienia, gdyby w dużej części oddziałów, w wyniku ściągania lub złamania innych procedur prowadzenia badania, część zadań okazała się łatwiejsza niż w drugim badaniu. Zakładając jednak, że błąd pomiaru nie był ani duży, ani systematyczny w żadnym z badań, możemy przejść do drugiej możliwości. Zmiana

[8] Wprowadzoną rozporządzeniem Ministra Edukacji Narodowej z dnia 23 grudnia 2008 r. w sprawie podstawy programowej wychowania przedszkolnego oraz kształcenia ogólnego w poszczególnych typach szkół (Dz. U. z dnia 15 stycznia 2009 r. Nr 4, poz. 17 z późn. zm.).

[9] Bank zadań przetestowanych w badaniu pilotażowym powstał w ramach projektu „Badania dotyczące rozwoju metodologii szacowania wskaźnika edukacyjnej wartości dodanej (EWD)".

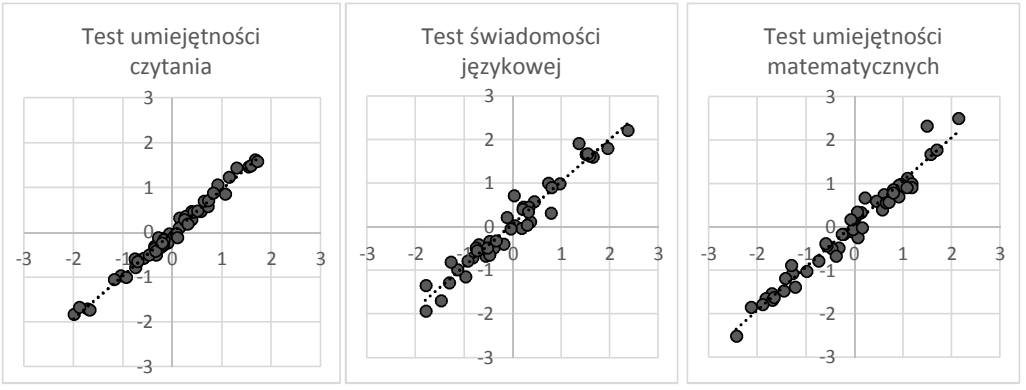
w obserwowanym uporządkowaniu zadań mogłaby świadczyć także o tym, że reforma programowa wpłynęła na systematyczne zróżnicowanie stopnia opanowania poszczególnych treści, mierzonych zadaniami testowymi, przez uczniów w ramach danej umiejętności. Posługując się przykładem z testu matematycznego, o takiej zmianie moglibyśmy mówić, gdyby w wyniku większej efektywności nauczania geometrii, zadania z tego obszaru okazałyby się łatwiejsze dla którejś z populacji. Kolejną możliwością, niezależną od poprzedniej, jest efekt, jaki mogło mieć dodanie lub usunięcie poszczególnych zadań z testu na definicję umiejętności. W szczególności, moglibyśmy oczekiwać, że nowe zadania „ściągnęłyby" interpretację skali w kierunku treści, które mierzą.

Czy zatem zadania wspólne dla obu edycji testu tworzą odmienne hierarchie ze względu na trudność? Porównanie parametrów odpowiadających sobie zadań przedstawiono na Rysunku 1, posługując się wykresami rozrzutu. Współczynniki korelacji liniowej Pearsona wynoszą odpowiednio: 0,993 dla testu umiejętności czytania, 0,973 dla testu świadomości językowej i 0,982 dla testu umiejętności matematycznych. Widzimy zatem, że są one niemal identyczne.

Odtworzenie się parametrów trudności zadań przemawia na rzecz stabilności skal będących przedmiotem pomiaru w obu badaniach. Warto jednak sprawdzić, czy w obu badaniach zadania opisane takimi parametrami są tak samo dobrze dopasowane do danych. Gdyby bowiem skale uległy zmianie, mogłoby się okazać, że niektóre zadania nie są już tak dobrze dopasowane do zdefiniowanego modelu, jak były w poprzednim badaniu. To z kolei mogłoby rodzić podejrzenia o zróżnicowanym funkcjonowaniu zadań w obu populacjach. Dla celów tej analizy posłużymy się wspomnianymi wcześniej miarami dopasowania *infit* i *outfit*. W Tabeli 1 podano statystyki opisowe dla tych miar: minimum, maksimum, pierwszy i dziewiąty decyl.

Zauważyć możemy, że statystyki opisowe dla miar dopasowania są zbliżone dla danych z obu badań. Analiza ta pokazała więc, że jakość dopasowania zadań do modelu w obu badaniach jest porównywalna. Dodatkowo zadania, które zostały dodane do testów na potrzeby badania EWD okazały się także dobrze dopasowane do danych (miary dopasowania mieszczą się w granicach 0,92–1,13).

Przedstawione tu wyniki pokazały, że skale mierzone testami osiągnięć TOS3 są stabilne, mimo zmian, jakie zostały dokonane w samych narzędziach i mimo różnych badanych populacji.



Rysunek 1. Porównanie parametrów trudności zadań z badania SUEK (oś pozioma) i EWD (oś pionowa).

Tabela 1
*Porównanie statystyk opisowych dla miar dopasowania zadań w badaniu SUEK i EWD**

Test	Badanie	*Outfit*				*Infit*			
		Min	Max	$k_{1.10}$	$k_{9.10}$	Min	Max	$k_{1.10}$	$k_{9.10}$
Umiejętność czytania	SUEK	0,75	1,26	0,88	1,19	0,85	1,17	0,92	1,10
	EWD	0,79	1,36	0,85	1,19	0,84	1,20	0,90	1,11
Świadomość językowa	SUEK	0,81	1,18	0,90	1,12	0,86	1,12	0,94	1,08
	EWD	0,78	1,21	0,91	1,13	0,83	1,13	0,95	1,09
Umiejętności matematyczne	SUEK	0,82	1,27	0,91	1,13	0,89	1,15	0,94	1,05
	EWD	0,79	1,29	0,91	1,15	0,90	1,19	0,94	1,06

*Oznaczenia: $k_{1.10}$ – pierwszy decyl; $k_{9.10}$ – dziewiąty decyl.

Ostateczna wersja testów

TOS3 są testami papierowymi, dostosowanymi do badania audytoryjnego. Tabela 2. przedstawia liczbę zadań wchodzących w skład poszczególnych testów. Każde z narzędzi ma dwie równoległe wersje, z pulą 15–16 zadań wspólnych (kotwiczących) dla obu wersji. Zarówno zadania kotwiczące, jak i zadania w każdej wersji testu są reprezentatywną próbką planu testu pod względem mierzonych treści i umiejętności, a obie wersje dla każdego testu mają porównywalną trudność.

Na każdy test składa się od 45 do 53 zadań (od 30 do 35 zadań na każdą wersję). Zadania są pogrupowane w 12 zeszytach testowych (po 6 na każdą wersję). Zadania z testu umiejętności matematycznych zostały wydzielone do odrębnych zeszytów (po 2 na każdą wersję), natomiast zadania z testu umiejętności czytania i testu świadomości językowej umieszczono we wspólnych zeszytach (po 4 na każdą wersję), z uwagi na większą ilość czasu potrzebą do przeczytania tekstów w teście sprawdzającym umiejętności czytania. Badany uczeń rozwiązuje więc 6 zeszytów testowych (najlepiej po 2 jednego dnia testowania), a na rozwiązanie jednego zeszytu przewidziane jest 35 minut.

Wszystkie zadania wchodzące w skład testów TOS3 są utajnione, tak aby narzędzia te mogły zostać wykorzystane w innych projektach badawczych[10].

Charakterystyka skal pomiarowych

Testy osiągnięć szkolnych, przygotowywane na potrzeby badań SUEK i EWD, skupiają się na pomiarze najważniejszych z punktu widzenia kształcenia w szkole podstawowej

[10] Z prośbą o udostępnienie narzędzi w celach naukowych należy zwrócić się do Instytutu Badań Edukacyjnych w Warszawie.

Tabela 2
Liczba zadań w testach osiągnięć szkolnych

Test	Liczba zadań w teście	Liczba zadań w każdej wersji testu	Liczba zadań kotwiczących
Umiejętność czytania	51	33	15
Świadomość językowa	45	30 i 31	16
Umiejętności matematyczne	53	34 i 35	16

obszarów: umiejętności czytania, świadomości językowej i umiejętności matematycznych. Każdemu obszarowi odpowiada jeden test osiągnięć. Ze względu na ograniczoną ich długość, nie zakładano tworzenia dla wyszczególnionych skal podskal przedstawiających wyniki uczniów w ramach bardziej szczegółowych umiejętności.

Poniżej opisano strukturę każdego z trzech testów. Na podstawie założonej struktury opracowane zostały plany testów dla każdej skali pomiarowej, które precyzowały, ile zadań, mierzących jakie szczegółowe umiejętności, powinno znaleźć się w teście. Opis struktury oraz wynikających z niej planów testów miał na celu zagwarantowanie różnorodności i reprezentatywności mierzonych treści i umiejętności, a tym samym zapewnienie trafności treściowej testu. Plany testów zostały przedstawione w Aneksie do artykułu.

Test umiejętności czytania

Test umiejętności czytania mierzy poziom rozumienia znaczenia czytanych samodzielnie tekstów różnego typu: literackich (prozatorskich i poetyckich), popularnonaukowych i użytkowych (ogłoszenie, regulamin, ulotka).

Pytania do każdego tekstu sprawdzają różne kompetencje. Osiemnaście zadań wymaga od uczniów wyszukania informacji zawartej w tekście, podjęcia decyzji, które informacje są ważne, a które nie, ze względu na ich związek z tematem lub pytaniem, a także ustalenia kolejności wydarzeń. Siedemnaście zadań mierzy umiejętność interpretacji tekstu, tj. wydobycia i wyjaśnienia jego sensu, określenia tematu i głównej myśli utworu, porównywania informacji zawartych w tekście, dostrzegania i wyjaśniania przyczyn i skutków opisanych zdarzeń, sytuacji, zjawisk, podania przypuszczalnych motywów działania, zachowania lub postawy bohaterów. Szesnaście zadań odwołuje się do umiejętności dokonania refleksji nad tekstem i jego oceny. Sprawdzają one to, czy uczeń potrafi odnieść tekst do własnego doświadczenia i wiedzy o świecie, czy umie dokonać oceny zdarzeń, postaci i poglądów w kontekście własnego doświadczenia czytelniczego i pozaszkolnego, jak również ocenić kompletność i spójność tekstu.

Test świadomości językowej

Zadania wchodzące w skład testu świadomości językowej można podzielić na trzy grupy. W pierwszej znajdują się zadania mające na celu pomiar bogactwa słownikowego uczniów (18 zadań). Są to: zadania polegające na utworzeniu (lub wybraniu spośród podanych) wyrazu o podobnym lub przeciwstawnym znaczeniu, zadania na dobranie poprawnej definicji podanego wyrazu lub wpisanie odpowiedniego wyrazu do podanej definicji, zadania na tworzenie lub rozpoznawanie powszechnie występujących porównań i wyjaśnianie znaczenia związków frazeologicznych, a także zadania na rozpoznanie niepoprawnego użycia słowa w zdaniu ze względu na jego znaczenie.

W drugiej grupie znajdują się zadania sprawdzające elementy wiedzy o języku (19 zadań). Wśród nich wyróżnić można zadania mierzące umiejętność tworzenia i uzupełniania zdań zgodnie z zasadami składni, rozpoznawania w tekście i tworzenia zdań oznajmujących, pytających, rozkazujących, a także zadania z zakresu ortografii i interpunkcji oraz poprawności językowej.

Wśród zadań mierzących umiejętności związane z pisaniem tekstów (8 zadań) znajdują się zadania sprawdzające umiejętność redagowania tekstu, rozpoznawania i tworzenia czytelnej struktury tekstu, rozpoznawania i nadawania poprawnego stylu wypowiedzi oraz umiejętności argumentowania.

W teście świadomości językowej nie ma natomiast zadań wymagających napisania

tekstu na zadany temat. Decyzja o ich niewprowadzaniu została podjęta z dwóch powodów. Po pierwsze, badania wskazują na istnienie znaczących różnic w ocenianiu uczniowskich wypowiedzi pisemnych między egzaminatorami (tzw. efekt egzaminatora), niezależnie od szczegółowości kryteriów oceniania i jakości szkolenia egzaminatorów (Dolata, Putkiewicz i Wiłkomirska, 2004). Po drugie, wykorzystanie w teście jednego zadania wymagającego napisania dłuższej wypowiedzi daje mniejszą (w sensie statystycznym) liczbę informacji przekładającą się na precyzję pomiaru, niż wykorzystanie kilku lub kilkunastu krótszych zadań, które uczeń może rozwiązać w tym samym czasie. Brak w teście świadomości językowej zadań na napisanie tekstu na zadany temat jest częściowo rekompensowany obecnością zadań z trzeciej grupy.

Test umiejętności matematycznych

Każde zadanie testu umiejętności matematycznych można opisać za pomocą dwóch kategorii: umiejętności, którymi należy się posłużyć, by je rozwiązać, oraz treści matematycznych, do których się odwołuje.

Zadania mierzą trzy umiejętności: odtwarzania wiadomości i dobrze wyćwiczonych schematów (7 zadań), którą zdefiniowano jako umiejętność rozwiązywania zadań typowych, wymagających użycia wyćwiczonych, prostych technik i posłużenia się dobrze znanymi obiektami. Innymi słowy – przywołania z pamięci znanych pojęć lub algorytmów. Kolejne 33 zadania wymagają odwołania się do umiejętności powiązania różnych wiadomości i dobrze wyćwiczonych schematów na potrzeby rozwiązania zadań mniej rutynowych, ale niezbyt odległych od zadań typowych. Uczeń musi zwykle wykonać większą liczbę kroków, aby rozwiązać zadanie z tej grupy. Musi wybrać pojęcia (modele, wzory, procedury) matematyczne odpowiednie dla rozwiązania danego problemu. Od ucznia oczekuje się, że będzie potrafił wykorzystać posiadane wiadomości do rozwiązania zadań, z których nie wynika wprost, jakie pojęcia czy procedury należy zastosować. Umiejętność przeprowadzenia prostego rozumowania matematycznego składającego się z kilku kroków sprawdzana jest przez 13 zadań. Od ucznia rozwiązującego je oczekuje się, że będzie umiał ustalić kolejność czynności prowadzących do rozwiązania problemu (sytuacji nowej, nieoczywistej dla osoby rozwiązującej test), że będzie potrafił wyciągnąć wnioski z kilku informacji podanych w różnej postaci.

Zadania mierzące każdą z tych umiejętności odwołują się do różnych treści matematycznych, które podzielono na trzy grupy: (a) ilość, (b) przestrzeń i kształt oraz (c) zmiana, związki, zależności.

Obszar „ilość", reprezentowany przez 30 zadań, odnosi się do rozumienia przez uczniów pojęcia liczby, rozumienia i odkrywania relacji między liczbami, umiejętności wykonywania obliczeń oraz rozumienia znaczenia tych operacji, a także umiejętności wykorzystania opisanych kompetencji w sytuacjach praktycznych. W obszarze tym mieszczą się także zagadnienia związane z pomiarem właściwości fizycznych przedmiotów. Odwołują się one do rozumienia problematyki długości, ciężaru, objętości, temperatury i czasu.

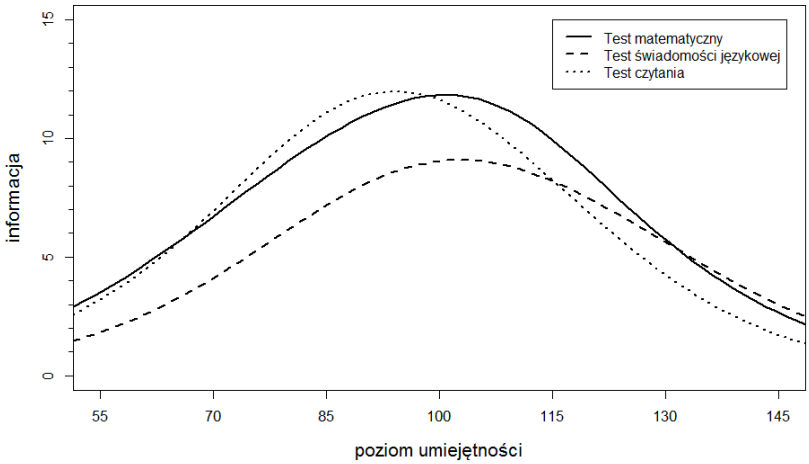
Na obszar „przestrzeń i kształt" (15 zadań) składają się zadania dotyczące problemów geometrycznych oraz związków przestrzennych między obiektami. Obejmuje on umiejętność rozpoznawania i rysowania figur geometrycznych, dostrzegania symetrii i regularności oraz wymaga wykorzystania wyobraźni przestrzennej (zadania te nie wymagają zastosowania wiedzy formalnej).

Ostatni obszar: „zmiana, związki, zależności" (8 zadań) obejmuje umiejętność dostrzegania przez ucznia związków i zależności, reprezentowanych w sposób graficzny, słowny, tabelaryczny lub symboliczny.

Rzetelność skal pomiarowych

Jakość skal pomiarowych możemy ocenić, przyglądając się precyzji, z jaką test pozwala oszacować wynik uczniów o różnym poziomie umiejętności. Warunkiem wstępnym takiej analizy w ramach modelu Rascha jest jednak odpowiednie dopasowanie zadań do modelu – warunek ten udało się spełnić w obu badaniach, w których wykorzystano TOS3. W ramach IRT, inaczej niż dla KTT, precyzja pomiaru (rzetelność) nie jest globalną własnością testu, ale stanowi złożenie precyzji pomiaru dla poszczególnych zadań wchodzących w skład narzędzia (Wright, 1990). Jednym ze sposobów pozwalających ocenić własności pomiarowe narzędzia jest krzywa informacyjna testu (Ayala, 2009), bazująca na informacji Fishera (Wright, 1990). Im większa wartość funkcji informacyjnej, tym większa precyzja pomiaru badanej cechy. W zależności od uwzględnionych w teście zadań przyjętego modelu IRT, krzywa informacyjna dla testu może przybierać różne kształty. Ulokowanie jej nad skalą umiejętności pozwala określić, czy dany test dobrze „wpasował się" w zakres umiejętności uczniów. Na Rysunku 2 przedstawione zostały krzywe informacyjne dla trzech testów TOS3 wykorzystanych w ramach badania EWD. Do ich wyrysowania uwzględniono wszystkie zadania wchodzące w skład testów.

Analizując wykres, można powiedzieć, że testy są dobrze dopasowane do rozkładu umiejętności uczniów w populacji – w zakresie ±2 odchylenia standardowe (w którym mieści się ok. 95% populacji) skonstruowane testy osiągnięć szkolnych pozwalają na pomiar umiejętności z zadowalającym stopniem precyzji. Stopień dopasowania testów waha się jednak dla poszczególnych umiejętności. Test czytania okazał się trochę za łatwy w stosunku do poziomu uczniów badanych. Wskazuje to na problemy związane z pozyskaniem zadań o wyższej trudności na etapie konstrukcji testu. Test świadomości językowej z kolei posiada zdecydowanie mniej zadań niż matematyczny czy wspomniany test czytania – stąd maksimum jego funkcji informacyjnej jest niższe niż dwóch pozostałych narzędzi, jednakże



Rysunek 2. Porównanie krzywych informacyjnych dla trzech testów TOS3 w badaniu EWD. Krzywa dla każdego testu odniesiona jest do wystandaryzowanego poziomu umiejętności (o średniej 100 i odchyleniu standardowym 15).

charakteryzuje się porównywalną do testu z matematyki precyzją pomiaru uczniów zdolnych.

Błąd pomiaru jest powiązany funkcyjnie z informacją Fishera (Wright, 1990) – im więcej informacji dostarcza test dla danego zakresu skali, z tym mniejszym błędem szacowane są wyniki dla tego zakresu (Ayala, 2009). Należy przy tym pamiętać, że na błąd oszacowania wyniku ucznia ma wpływ liczba rozwiązywanych przez niego zadań (do im większej liczby zadań uczeń podszedł, tym mniejszy błąd), a także ich trudność (im lepiej dopasowany zakres trudności zadań do poziomu umiejętności ucznia, tym mniejszy błąd).

W związku z faktem, że testy TOS3 składają się z dwóch wersji, a te z kolei z kilku zeszytów testowych, istnieje możliwość wyrysowywania wielu krzywych informacyjnych. Potencjalnie można taką krzywą wyrysować dla każdego z możliwych sposobów rozwiązania testu przez ucznia (od pojedynczego zeszytu po wszystkie zeszyty w wersji). Precyzja pomiaru będzie wtedy różna, zgodnie z regułami opisanymi wyżej. Choć kwestia indywidualnego błędu pomiaru ma kluczowe znaczenie przy pomiarze diagnostycznym, to ze względu na cele badań SUEK i EWD, poziomami analizy będącymi w centrum zainteresowania są oddział i szkoła, a w związku z tym wystarczy scharakteryzowanie pomiaru na ogólniejszym poziomie.

Choć z wykresu z krzywymi informacyjnymi wyraźnie widać, że poziom dokładności pomiaru jest zróżnicowany dla różnych poziomów umiejętności, to w ramach zastosowanej metody estymacji wyników uczniów (MML) – istnieje możliwość oszacowania jednego, ogólnego parametru podsumowującego stopień precyzji pomiaru, tzw. współczynnika rzetelności EAP/PV (Adams, 2005).

Współczynnik ten jest stosunkiem wariancji dwóch rodzajów estymatorów wyników uczniów. Oba związane są z tzw. rozkładem *a posteriori* wyniku ucznia, czyli przemnożenia funkcji wiarygodności dla wektora odpowiedzi ucznia przez funkcję gęstości dla rozkładu *a priori*, czyli rozkładu umiejętności w populacji, który na potrzeby wyliczania wyników uczniów w badaniach SUEK i EWD został opisany przez rozkład normalny o estymowanych parametrach (średniej i odchyleniu standardowym). Rozkład *a posteriori* dla ucznia jest tym bardziej zbliżony do rozkładu w populacji, im mniejsza jest precyzja pomiaru. Jego średnia jest estymatorem *expected a posteriori* (EAP). *Plausible values* (PV), to z kolei losowa próbka wartości („prawdopodobnych") z tego rozkładu. Znany jest efekt zaniżania oszacowania wariancji w populacji z wykorzystaniem estymatora EAP (w przeciwieństwie do PV; Wu, 2005), który wynika z błędu pomiaru – średnia rozkładu *a posteriori* ściągana jest do średniej populacyjnej tym bardziej, im mniej informacji o poziomie umiejętności ucznia dostarcza test. Z drugiej strony, im więcej tej informacji, tym węższy rozkład *a posteriori* (mniejsza niepewność związana z oszacowaniem wyniku ucznia). Innymi słowy, im mniejszy błąd pomiaru, tym mniejsze zaniżenie wariancji z wykorzystaniem estymatora EAP (mniejsze „ściąganie" ku średniej populacyjnej). W sytuacji doskonale rzetelnego pomiaru rozkłady *a posteriori* wyników uczniów byłyby punktowe, a więc oba estymatory poziomu umiejętności dla danego ucznia byłyby sobie równe. Wariancje tych estymatorów w próbie również byłyby równe, a współczynnik rzetelności przyjmowałby wartość 1. Współczynnik rzetelności EAP/PV zdaje zatem sprawę z wielkości efektu zaniżenia oszacowania wariancji estymatora EAP, a tym samym mówi nam, z jaką precyzją dokonano pomiaru umiejętności uczniów. W Tabeli 3 zamieszczono wartości współczynników rzetelności EAP/PV dla testów TOS3, uzyskane w badaniu EWD.

Tabela 3
Porównanie wartości współczynnika rzetelności EAP/PV dla testów TOS3 w badaniu EWD

Test	Wartość współczynnika rzetelności EAP/PV
Umiejętność czytania	0,858
Świadomość językowa	0,849
Umiejętności matematyczne	0,876

Wartości współczynników są wysokie dla wszystkich trzech testów, co potwierdza ogólnie wysoki poziom precyzji pomiaru obserwowany na wykresach z krzywymi informacyjnymi. Widzimy także, że większa dokładność pomiaru dla testu matematycznego dla całego zakresu skali jest odwzorowana w relatywnie wyższej wartości współczynnika jego rzetelności. Należy jednakże pamiętać, że współczynnik ten stanowi uśrednienie precyzji pomiaru dla wszystkich uczniów, którzy rozwiązywali testy TOS3 w danym badaniu, niezależnie od tego, do ilu zadań podeszli – informacja o indywidualnych błędach oszacowań wyników uczniów zostaje utracona za cenę wygody posługiwania się jedną wartością liczbową dla całego pomiaru.

Podsumowując rozważania o rzetelności testów osiągnięć szkolnych, można powiedzieć, że narzędzia charakteryzują się dobrymi własnościami psychometrycznymi. Pomiar umiejętności przeprowadzony za pomocą tych narzędzi zapewnia wyniki obarczone niepewnością o rozsądnych rozmiarach dla całego zakresu skali.

Trafność pomiaru osiągnięć szkolnych

Precyzja pomiaru nie jest wystarczającym kryterium jakości testów. Narzędzie mogłoby bowiem rzetelnie mierzyć cechy uczniów odległe od tych, o których chcemy wnioskować. TOS3 miały na celu pomiar wyników nauczania po I etapie kształcenia. Dlatego należy sprawdzić, czy mierzą one to, co było zamierzeniem ich twórców. W psychologicznym ujęciu problematyki trafności wiele miejsca poświęca się na wyszczególnianie i opisywanie różnych rodzajów trafności, konsekwencji ich braku oraz sposobów ich badania (Anastasi i Urbina, 1999). W kontekście pomiaru dydaktycznego najważniejsze wydają się pytania o to, czy test obejmuje reprezentatywną próbę mierzonych wiadomości i umiejętności oraz czy wywołuje u uczniów pożądane procesy myślowe i pozwala na zarejestrowanie dowodów, że takie procesy zaistniały (Jakubowski i Pokropek, 2009; Kane, 2006).

Trafność treściowa narzędzia pomiarowego dotyczy tego, czy pozycje testowe stanowią reprezentatywną próbę dziedziny, która ma być przedmiotem pomiaru. Ocena testu osiągnięć szkolnych pod względem tego kryterium zasadniczo polega na analizie jego treści w celu stwierdzenia, czy wchodzące w jego skład zadania mierzą wszystkie ważne z punktu widzenia badanego konstruktu wiadomości i umiejętności, oraz czy zadania odwołujące się do szczegółowych umiejętności znajdują się w teście we właściwych proporcjach. Trafność treściowa testu jest dużym stopniu wynikiem przyjętych procedur konstrukcji narzędzia, a nie oceny post factum. Aby ustrzec się przed przypadkowością pozycji składających się na test lub też nadreprezentacją zadań mierzących takie umiejętności, dla których łatwiej ułożyć dobre psychometrycznie zadanie, należy badaną dziedzinę z góry dobrze opisać a także opracować dokładną specyfikację

zawartości testu. Ważna jest także ekspercka ocena zadań oraz weryfikacja empiryczna, które mogą ustrzec przed włączeniem do testu zadania, które tylko z pozoru mierzą założoną umiejętność.

Na procedury, które miały zapewnić trafność treściową TOS3, składały się na etapie planowania narzędzia: zdefiniowanie i opisanie skal pomiarowych, analiza podstawy programowej i obowiązujących programów nauczania oraz opracowanie na ich podstawie szczegółowych planów testów precyzujących, ile zadań mierzących jakie umiejętności powinno się znaleźć w teście. Następnie koncepcja skal pomiarowych, plany testów oraz opracowane zgodnie z nimi zadania testowe zostały poddane recenzji zewnętrznej i drobiazgowej ocenie eksperckiej. Uwagi przekazane przez recenzentów pozwoliły na udoskonalenie części zadań, które tego wymagały i wykluczeniu zadań najbardziej problematycznych. Tak wyłoniona grupa zadań została poddana weryfikacji empirycznej w badaniu pilotażowym, a do testów zasadniczych zostały wybrane zadania nie tylko najlepsze pomiarowo, ale także zgodne z przyjętymi planami testów. W ten sposób udało się zapewnić różnorodność i reprezentatywność treściową opracowanych testów.

Skoro testy zostały skonstruowane w taki sposób, że zadania wchodzące w ich skład odwołują się do różnych treści i umiejętności szczegółowych, można postawić pytanie, czy jednowymiarowy model Rascha wykorzystany do ustalenia skali umiejętności, jest w tym przypadku modelem trafnym. Model ten zakłada, po pierwsze, że odpowiedzi na zadania są wskaźnikami umiejętności dającej się opisać za pomocą jednego wymiaru, a po drugie, że wszystkie pozycje testu są w takim samym stopniu związane z wynikiem ogólnym (Ayala, 2009). Procedura konstrukcji testów została tak zaplanowana i zrealizowana, aby dane z pomiaru tymi narzędziami można było z powodzeniem dopasować do jednowymiarowego modelu Rascha. Do testów ostatecznych z pokaźnego banku zadań zostały włączone tylko te pozycje, które były dopasowane do takiego modelu. Dzięki temu z jednej strony w teście znalazły się tylko zadania, które istotnie korelują z wynikiem ogólnym, a ponadto wykluczona została możliwość włączenia do testu zadań zbyt silnie determinujących wynik. Zostało to potwierdzone w badaniu zasadniczym, w którym także udało się dopasować wspomniany model. Oznacza to, że wszystkie pozycje wchodzące w skład testu są tak samo dobrymi wskaźnikami ogólnej umiejętności, która kryje się za obserwowaną reakcją ucznia na zadanie, czyli jego poprawnym lub błędnym rozwiązaniem. Umiejętność tę z powodzeniem możemy opisać, wykorzystując model jednowymiarowy.

Ostatnim elementem badania trafności będzie konfrontacja wyników testów z zewnętrznymi miarami podobnych konstruktów oraz zmiennymi opisującymi czynniki, co do których możemy oczekiwać, że są powiązane z umiejętnościami, które z założenia mają mierzyć testy TOS3. W analizach odniesiemy się do trzech zewnętrznych kryteriów: poziomu inteligencji, nauczycielskich ocen poziomu umiejętności uczniów oraz wyników innych testów osiągnięć szkolnych. Wykorzystamy wyniki badania SUEK, jako że dają one większe możliwości w zakresie konfrontacji z kryteriami zewnętrznymi.

Poziom inteligencji uczniów był mierzony w III klasie szkoły podstawowej testem matryc Ravena w wersji standard, formie klasycznej (Jaworowska i Szustrowa, 1991). Wyniki tego testu zostały wyskalowane dwuparametrycznym modelem IRT. Oceny nauczycielskie zebrano pod koniec klasy III. Nauczycieli nauczania zintegrowanego badanych klas poproszono, by ocenili wszystkich swoich uczniów na czterostopniowej skali opisowej osobno dla umiejętności językowych, osobno dla matematycznych. Trzecim kryterium były wyniki uczniów z testu z języka

polskiego i matematyki uzyskane w badaniu OBUT, w którym uczestniczyli badani uczniowie (Pregler i Wiatrak, 2011)[11]. W prezentowanych analizach wykorzystano sumę punktów.

W Tabeli 4 przedstawiono siłę związków wymienionych kryteriów z wynikami trzech testów osiągnięć (wykorzystano estymatory EAP). Dla zmiennych ciągłych (mierzonych na skali interwałowej) podano współczynnik korelacji liniowej i jego kwadrat (współczynnik determinacji r^2) mówiący o tym, jaką część zmienności wyników egzaminacyjnych możemy wyjaśnić przez dane kryterium. Dla zmiennych porządkowych (nauczycielskie oceny) siłę związku wyrażono współczynnikiem η^2, mówiącym o tym, jaką część wariancji zmiennej zależnej możemy wyjaśnić przez przynależność do poszczególnych kategorii zmiennej niezależnej (jest to miara

[11] Wykorzystując wyniki z tego pomiaru, należy zwrócić uwagę na kilka jego cech, które z punktu widzenia celów prezentowanych analiz, są jego mankamentami. Diagnoza OBUT jest przeprowadzana i oceniana przez nauczycieli uczących uczniów wypełniających testy. Szkoła otrzymuje opracowane przez zespół badawczy testy oraz instrukcje przeprowadzenia badania, jednak nie ma pewności, czy procedury te są przestrzegane. Narzędzia pomiarowe są krótkie, co sprawia, że wyniki pomiaru są mało dokładne (skala umiejętności matematycznych ma 17 rozróżnialnych kategorii wyników surowych a skala testu języka polskiego 24), a w rozkładach wyników dostrzega się silny efekt sufitowy – testy są za łatwe, by dobrze różnicować uczniów najzdolniejszych (skośność wynosi odpowiednio: -0,964 dla testu z języka polskiego i -0,384 dla testu z matematyki). Mankamenty te mogą powodować zaniżenie korelacji między wynikami testów. Ponadto wyniki z badania OBUT udało się przyłączyć tylko dla ok. 64% uczniów objętych badaniem testowym SUEK.

Tabela 4
Siła związku między wynikami testów osiągnięć a zewnętrznymi kryteriami[*]

Kryterium	Współczynnik	Test umiejętności czytania		Test świadomości językowej		Test umiejętności matematycznych	
Wynik testu matryc Ravena	r^2	0,319	(0,014)	0,387	(0,014)	0,478	(0,015)
	r	0,565	(0,013)	0,622	(0,011)	0,691	(0,011)
Oceny – język polski	η^2	0,502	(0,016)	0,548	(0,015)	0,483	(0,015)
	$b1$	85,74	(2,32)	76,27	(2,28)	91,27	(2,49)
	$b2$	8,49	(0,638)	9,51	(0,573)	8,47	(0,636)
	$b3$	8,25	(0,453)	8,45	(0,402)	8,01	(0,482)
	$b4$	10,35	(0,429)	10,36	(0,382)	10,46	(0,443)
Oceny – matematyka	η^2	0,485	(0,016)	0,517	(0,014)	0,543	(0,016)
	$b1$	88,55	(2,23)	81,13	(2,29)	80,44	(2,43)
	$b2$	7,836	(0,568)	9,59	(0,587)	9,07	(0,613)
	$b3$	8,32	(0,472)	8,36	(0,470)	8,27	(0,480)
	$b4$	10,64	(0,470)	10,08	(0,418)	11,64	(0,436)
Wynik OBUT – język polski	r^2	0,488	(0,019)	0,537	(0,017)	0,417	(0,016)
	r	0,698	(0,013)	0,733	(0,012)	0,646	(0,013)
Wynik OBUT – matematyka	r^2	0,406	(0,019)	0,411	(0,020)	0,528	(0,022)
	r	0,637	(0,015)	0,641	(0,016)	0,727	(0,015)

[*] W nawiasach podano błąd standardowy oszacowań.

o analogicznej interpretacji jak r^2). Dodatkowo podano wartości współczynników regresji pozwalające określić, jaka byłaby różnica średnich wyników testów osiągnięć uczniów o określonej ocenie w porównaniu do uczniów, którzy uzyskali ocenę o jeden niższą (oceny nauczycielskie zostały zrekodowane na zmienną pomocniczą w taki sposób, że np. współczynnik regresji $b3$ oznaczał różnicę w średnim wyniku testu między grupą uczniów, która uzyskała ocenę 3, a tymi, którzy uzyskali ocenę 2, stała regresji została oznaczona przez $b1$ i odpowiadała ona średniemu wynikowi uczniów, którzy uzyskali ocenę 1). W nawiasach podano błędy standardowe oszacowań. Wszystkie podane w tabeli współczynniki są istotne statystycznie na poziomie $p < 0{,}001$. Prezentowane parametry zostały wyliczone za pomocą modeli regresji estymowanych w programie Mplus 7 metodą największej wiarygodności, z uwzględnieniem trójstopniowego schematu doboru próby i nierównych prawdopodobieństw doboru.

Wyniki analiz są zgodne z oczekiwaniami. Potwierdzono związek wyników testów z inteligencją – konstruktem, który zgodnie z teorią i wynikami licznych badań jest powiązany z osiągnięciami szkolnymi. Siła tego związku jest zbliżona do obserwowanej w badaniach, a zgodnie z doniesieniami innych prac, jest trochę silniejsza dla przedmiotów matematyczno-przyrodniczych niż humanistycznych (Ferrer i McArdle, 2004; Sternberg, Grigorenko i Bundy, 2001; Teo, Carlson, Mathieu, Egeland i Sroufe, 1996, por. też przegląd polskich badań w: Dolata, 2008, s. 23–39).

Oceny nauczycielskie są często wykorzystywane jako zewnętrzne kryterium w badaniu trafności testów osiągnięć szkolnych (Anastasi i Urbina, 1999). Ze statystycznego punktu widzenia jest to miara mało rzetelna, zasadniczo nieporównywalna między grupami uczniów ocenianymi przez różnych nauczycieli (Jasińska, 2010). Mimo to ich wykorzystanie w badaniu trafności jest uzasadnione o tyle, że obejmują one szerokie spektrum umiejętności uczniów, które mają okazję ujawnić się w różnych okolicznościach (nie tylko podczas badania testowego). Dodatkowo są one w mniejszym stopniu uzależnione od chwilowej dyspozycji ucznia. Dlatego uznaliśmy za uzasadnione wykorzystanie ich w analizach.

Przedstawione wyniki przemawiają na rzecz trafności testów osiągnięć. Zaobserwowano umiarkowanie silne, pozytywne związki między nauczycielskimi ocenami a wynikami testów z danych przedmiotów (ok. 50–55% wariancji wyników testów można wyjaśnić za pomocą informacji o ocenie udzielonej przez nauczyciela z danego obszaru wiedzy). Dodatkowym potwierdzeniem jest silniejsza relacja dla odpowiadających sobie przedmiotów (np. między wynikiem testu z matematyki a nauczycielską oceną umiejętności matematycznych ucznia), niż dla różnych przedmiotów. Oczywiście krzyżowe relacje także występują z uwagi na to, że osiągnięcia szkolne z języka polskiego i matematyki są ze sobą powiązane.

Ostatnim kryterium są wyniki uczniów z testu z języka polskiego i matematyki uzyskane w badaniu OBUT. Otrzymany wzorzec korelacji jest zgodny z założeniami. Obserwujemy silniejszy związek wyników testu z języka polskiego z badania OBUT z wynikami testu umiejętności czytania i świadomości językowej wykorzystanymi w badaniu SUEK (ok. 0,7), niż z wynikami z testu umiejętności matematycznych (ok. 0,6). Test z matematyki z badania OBUT także silniej koreluje ze skalą umiejętności matematycznych z testów z badania SUEK (ok. 0,7) niż ze skalami umiejętności językowych (ok. 0,6). Dodatkowo korelacje między wynikami testów z odpowiadających sobie zakresów umiejętności są wyższe niż korelacja między wynikiem z testu

OBUT z języka polskiego a testem OBUT z matematyki, która wynosi 0,632[12]. Korelacje między testami SUEK są wyższe i wynoszą odpowiednio: między testem matematycznym a testem umiejętności czytania: 0,782, między testem matematycznym a testem świadomości językowej: 0,789 i między testem świadomości językowej a testem umiejętności czytania: 0,848. Nie można jednak tych wartości wprost porównać z korelacjami między testami SUEK a OBUT, gdyż te drugie, ze względu na ponad dwukrotnie krótszą długość, są testami o mniejszej rzetelności. Stąd korelacje pomiędzy testami bardziej rzetelnymi są wyższe. Ponadto w analizach tych za wynik z testów OBUT uznano sumę punktów (nie dysponowano pełnym rekordem odpowiedzi uczniów na zadania, co uniemożliwiło wyskalowanie wyników). Mogło się to przyczynić się do zaniżenia korelacji między wynikami z testów OBUT i SUEK ze względu na niespełnienie założenia o liniowym charakterze zależności.

Mając świadomość ograniczeń przytoczonych tu wyników, można mimo wszystko stwierdzić, że uzyskany wzorzec korelacji przemawia za trafnością testów TOS3.

Dyskusja

Głównym celem artykułu było opisanie zestawu trzech testów osiągnięć szkolnych TOS3 – procedury ich tworzenia, formatu narzędzia, mierzonych umiejętności oraz wyników analizy rzetelności i trafności. Nie był to jednak cel jedyny.

Celem pobocznym było zwiększenie świadomości twórców testów osiągnięć szkolnych – działających w ramach systemu egzaminacyjnego, sektora prywatnego, czy też badaczy zajmujących się pomiarem osiągnięć szkolnych – co do potrzeby szczegółowego dokumentowania procesu wytwarzania narzędzi, a także ich jakości. Brak takiej dokumentacji jest niechlubną cechą większości testów osiągnięć szkolnych wykorzystywanych w Polsce. Naszym zdaniem nie pozostaje ona bez konsekwencji.

W ciągu ostatnich kilku lat obserwujemy dwa sprzeczne ze sobą trendy związane z pomiarem osiągnięć szkolnych. Z jednej strony sytuacja jest dobra – widoczny jest wzrost zainteresowania obiektywnym testowaniem, zarówno po stronie badaczy czy administracji centralnej, ale – co być może najważniejsze – także samych szkół. Międzynarodowe i krajowe badania oraz system egzaminów zewnętrznych dostarczają nieocenionej wręcz wiedzy na temat systemu szkolnictwa w naszym kraju. Różnego rodzaju badania i programy diagnostyczne nakierowane na wspieranie szkół cieszą się bardzo dużą popularnością. Z drugiej zaś strony, coraz wyraźniej słychać głos krytyki wymierzony w kierunku testowego sprawdzania wiedzy. Krytyka ta skupia się m.in. na miarodajności otrzymywanych wyników, czyli trafności narzędzi wykorzystanych do ich otrzymania. Tych krytycznych głosów nie można lekceważyć – pobrzmiewa w nich bowiem uzasadniony niepokój.

Brak szczegółowej i publicznie dostępnej dokumentacji dotyczącej podjętych podczas konstrukcji testów działań, wpływających na ich jakość, podważa zaufanie do dostarczanych przez te narzędzia wyników. Uniemożliwia też rozwój metodologii związanej z konstrukcją testów osiągnięć szkolnych na gruncie polskim. Rozwój ten jest uzależniony od możliwości wymiany doświadczeń oraz istnienia merytorycznej krytyki stosowanych rozwiązań. Pierwszym krokiem w tym kierunku jest odrzucenie przekonania, że „testy mówią same za siebie" i podjęcie wyzwania pełniejszego opisu tworzonych narzędzi. Wydaje się, że tylko w ten sposób może funkcjonować rzetelny dyskurs dotyczący

[12] Ten i wymieniane dalej współczynniki korelacji policzono analogicznymi modelami. Wyniki przytoczono tylko w tekście artykułu, by zachować czytelność Tabeli 4.

pomiaru edukacyjnego, który zwiększy zaufanie opinii publicznej do wykorzystywanych testów osiągnięć, a także zrozumienie dla potrzeby ich stosowania. Sprawi też, że przyszłe narzędzia do pomiaru umiejętności uczniów będą lepsze, ich wyniki bardziej miarodajne, a decyzje podejmowane na ich podstawie bardziej trafne.

Literatura

Adams, R. J. (2005). Reliability as a measurement design effect. *Studies in Educational Evaluation*, *31*(2–3), 162–172. doi:10.1016/j.stueduc.2005.05.008

Anastasi, A. i Urbina, A. (1999). *Testy psychologiczne*. Warszawa: Pracownia Testów Psychologicznych Polskiego Towarzystwa Psychologicznego.

Aubrey, C., Godfrey, R. i Dahl, S. (2006). Early mathematics development and later achievement: Further evidence. *Mathematics Education Research Journal*, *18*(1), 27–46. doi:10.1007/BF03217428

Ayala, R. J. de (2009). *The theory and practice of item response theory*. New York: Guilford Press.

Baghaei, P. (2008). Local dependency and Rasch measures. *Rasch Measurement Transactions*, *21*(3), 1105–1106.

Boland, T. (1993). The importance of being literate: reading development in primary school and its consequences for the school career in secondary education. *European Journal of Psychology of Education*, *8*(3), 289–305. doi:10.1007/BF03174083

Dolata, R. (2008). *Szkoła, segregacje, nierównosci*. Warszawa: Wydawnictwa Uniwersytetu Warszawskiego.

Dolata, R., Putkiewicz, E i Wiłkomirska, A. (2004). *Reforma egzaminu maturalnego – oceny i rekomendacje: raport z badań*. Warszawa: Instytut Spraw Publicznych.

Downing, S. M. (2006a). Selected-response item formats in test development. W: *Handbook of Test Development* (s. 131–153). New York, NY: Lawrence Erlbaum.

Downing, S. M. (2006b). Twelve steps for effective test development. W: S. M. Downing i T. M. Haladyna (red.), *Handbook of Test Development* (s. 3–25). New York, NY: Lawrence Erlbaum.

Ferrer, E. i McArdle, J. J. (2004). An experimental analysis of dynamic hypotheses about cognitive abilities and achievement from childhood to early adulthood. *Developmental Psychology*, *40*(6), 935–952.

Haladyna, T. M., Downing, S. M. i Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment applied measurement in education. *Applied Measurement in Education*, *15*(3), 309–334.

Jakubowski, M. i Pokropek, A. (2009). *Badając egzaminy: podejście ilościowe w badaniach edukacyjnych*. Warszawa: Centralna Komisja Egzaminacyjna.

Jasińska, A. (2010). Pomiar gotowości szkolnej za pomocą skali quasi-obserwacyjnej. W: B. Niemierko i M. K. Szmigel (red.), *Teraźniejszość i przyszłość oceniania szkolnego* (s. 415–424). Kraków: Polskie Towarzystwo Diagnostyki Edukacyjnej.

Jasińska, A. i Modzelewski, M. (2012). Można inaczej. Wykorzystanie IRT do konstrukcji testów osiągnięć szkolnych. Referat wygłoszony na XVIII Krajowej Konferencji Diagnostyki Edukacyjnej W: B. Niemierko i M. K. Szmigel (red.), *Regionalne i lokalne diagnozy edukacyjne.* (s. 157–168). Kraków: Polskie Towarzystwo Diagnostyki Edukacyjnej.

Jaworowska, A. i Szustrowa, T. (1991). *Podręcznik do testu matryc Ravena. Wersja standard (1956). Polska standaryzacja 1989 (5;11–15;11)*. Warszawa: Pracownia Testów Psychologicznych Polskiego Towarzystwa Psychologicznego.

Kane, M. (2006). Content-related evidence in test development. W: *Handbook of Test Development* (s. 131–153). New York, NY: Lawrence Erlbaum.

Karwowski, M. i Dziedziewicz, D. (2012). *Test umiejętności na starcie szkolnym. Podręcznik*. Warszawa: Instytut Badań Edukacyjnych.

Konarzewski, K. (2012). *TIMSS i PIRLS 2011. Osiągnięcia szkolne polskich trzecioklasistów w perspektywie międzynarodowej*. Warszawa: Centralna Komisja Egzaminacyjna.

Kondratek, B. i Grudniewska, M. (2013). Test Mantel-Haenshel oraz modelowanie IRT jako narzędzia służące do wykrywania DIF oraz opisu jego wielkości na przykładzie zadań ocenianych dychotomicznie. *Edukacja*, *122*(2), 34–55.

Kondratek, B. i Pokropek, A. (2013). IRT i pomiar edukacyjny. *Edukacja*, *124*(4), 42–66.

Martin, M. O. i Mullis, I. V. S. (red.). (2013). *Methods and procedures in TIMSS and PIRLS 2011*. Chestnut Hill, MA: TIMSS i PIRLS International Study Center.

Masters, G. N. i Wright, B. D. (1997). The partial credit model. W: *Handbook of modern item response theory* (s. 101–122). New York, NY: Springer.

Pregler, A. (red.). (2013). *Ogólnopolskie badanie umiejętności trzecioklasistów. Raport OBUT 2013*. Warszawa: Instytut Badań Edukacyjnych.

Pregler, A. i Wiatrak, E. (red.). (2011). *Ogólnopolskie badanie umiejętności trzecioklasistów. Raport OBUT 2011*. Warszawa: Centralna Komisja Egzaminacyjna.

Pregler, A. i Wiatrak, E. (red.). (2012). *Ogólnopolskie badanie umiejętności trzecioklasistów. Raport OBUT 2012*. Warszawa: Centralna Komisja Egzaminacyjna.

Rasch, G., (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Rozporządzenie Ministra Edukacji Narodowej z dnia 23 grudnia 2008 r. w sprawie podstawy programowej wychowania przedszkolnego oraz kształcenia ogólnego w poszczególnych typach szkół (Dz. U. z dnia 15 stycznia 2009 r. Nr 4, poz. 17 z późn. zm.)

Slavin, R. E., Karweit, N. L. i Wasik, B. A. (1992). Preventing early school failure: what works? *Educational Leadership*, *50*(4), 10–18.

Sternberg, R. J., Grigorenko, E. L. i Bundy, D. A. (2001). The predictive value of IQ. *Merrill-Palmer Quarterly*, *47*(1), 1–41.

Teo, A., Carlson, E., Mathieu, P. J., Egeland, B. i Sroufe, L. A. (1996). A prospective longitudinal study of psychosocial predictors of achievement. *Journal of School Psychology*, *34*(3), 285–306.

Wright, B. D. (1990). What is information? *Rasch Measurement Transactions*, *4*(2), 109.

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation*, *31*(2–3), 114–128. doi:10.1016/j.stueduc.2005.05.005

Wu, M., Adams, R. J., Wilson, M. R. i Haldane, S. A. (2007). *ACER ConQuest version 2.0: Generalised Item Response Modelling Software*. Melbourne: ACER Press.

Aneks

Poniżej zamieszczono plany testów przedstawiające, ile zadań odwołujących się do jakich treści i umiejętności weszło w skład poszczególnych testów. Dane te zostały przedstawione w podziale na dwie wersje testu (A i B). Dodatkowo podano liczbę zadań kotwiczących, które wchodziły w skład każdej wersji testu.

Tabela A1
Plan testu świadomości językowej

Umiejętność ogólna	Umiejętność szczegółowa	Liczba zadań		
		Wersja A	Wersja B	Zadania kotwiczące
Bogactwo słownikowe	Rozpoznawanie niepoprawnego, ze względu na znaczenie, użycia słowa	2		1
	Tworzenie i rozpoznawanie wyrażeń porównawczych	2		1
	Tworzenie lub rozpoznawanie synonimów i antonimów	2	1	2
	Wyjaśnianie znaczenia słowa		2	1
	Wyjaśnianie znaczenia związków frazeologicznych	1	2	1
Elementy wiedzy o języku	Ortografia i interpunkcja	3	2	3
	Poprawność gramatyczna wypowiedzi	3	5	3
Umiejętności związane z pisaniem tekstów	Argumentacja		1	1
	Redagowanie tekstu	2		1
	Porządkowanie struktury wypowiedzi		1	1
	Rozpoznawanie i nadawanie poprawnego stylu wypowiedzi			1

Tabela A2
Plan testu czytania

Tekst	Umiejętność	Liczba zadań		
		Wersja A	Wersja B	Zadania kotwiczące
1: Literacki (proza)	Wyszukiwanie informacji	2		
	Interpretacja	2		
	Refleksja i ocena	1		
2: Literacki (proza)	Wyszukiwanie informacji		4	
	Interpretacja		2	
	Refleksja i ocena		5	
3: Literacki (poezja)	Wyszukiwanie informacji	1		1
	Interpretacja			2
	Refleksja i ocena	1	1	1
4: Popularnonaukowy	Wyszukiwanie informacji			4
	Interpretacja			1
	Refleksja i ocena	1		1
5: Użytkowy (ogłoszenie)	Wyszukiwanie informacji		1	
	Interpretacja		2	
	Refleksja i ocena		2	
6: Użytkowy (ulotka)	Wyszukiwanie informacji			2
	Interpretacja	1	1	3
7: Użytkowy (regulamin)	Wyszukiwanie informacji	3		
	Interpretacja	3		
	Refleksja i ocena	3		

Tabela A3
Plan testu umiejętności matematycznych

Grupa treści	Treści	Umiejętność	Liczba zadań		
			Wersja A	Wersja B	Zadania kotwiczące
Ilość	Rozumienie pojęcia liczby	Powiązania			3
	Dodawanie i odejmowanie na liczbach naturalnych	Powiązania	2	2	
		Rozumowania			1
	Obliczenia pieniężne	Powiązania	1		1
	Szacowanie wyników działań	Rozumowania	1		
	Rozwiązywanie zadań tekstowych	Odtwarzania		1	1
		Powiązania	2	2	
		Rozumowania	1		1
	Układanie zadań tekstowych	Powiązania		1	1
	Pomiar (długości, czasu, temperatury, masy lub objętości*)	Odtwarzania	1	2	1
		Powiązania	1	1	
		Rozumowania	1	1	1
Przestrzeń i kształt	Położenie obiektów względem siebie	Powiązania	1		
		Rozumowania			1
	Symetria i regularności	Powiązania	1	2	1
	Figury płaskie	Odtwarzania		1	
		Powiązania	2		1
		Rozumowania		1	
	Wyobraźnia przestrzenna	Powiązania	1	2	
		Rozumowania			1
Zmiana, związki, zależności	Dostrzeganie związków i prawidłowości	Powiązania	2	2	1
		Rozumowania	1	1	1

* Bez odwołania do wiedzy formalnej czy wzorów.