Aleksandra Jasińska
Michał Modzelewski
Instytut Badań Edukacyjnych

Można inaczej.
Wykorzystanie IRT do konstrukcji testów osiągnięć szkolnych

Skonstruowanie dobrych testów osiągnięć szkolnych jest poważnym wyzwaniem teoretycznym i logistycznym. Opracowanie trafnych i rzetelnych narzędzi wymaga przede wszystkim starannego ich zaplanowania oraz systematyczności i konsekwencji w realizacji przyjętej koncepcji. W Polsce najważniejsze z punktu widzenia różnych funkcji - selekcyjnej, diagnostycznej, ewaluacyjnej, formatywnej - testy osiągnięć szkolnych są tworzone od 10 lat przez krajowy system egzaminów zewnętrznych. Mimo wieloletniego doświadczenia w tworzeniu testów jakość egzaminów zewnętrznych pozostawia wiele do życzenia. Ich wady coraz częściej wypunktowuje opinia publiczna, która zazwyczaj podaje w wątpliwość trafność fasadową zadań oraz całych egzaminów. Negatywne emocje budzi praktycznie każda edycja egzaminów państwowych. Odczucia opinii publicznej, sugerujące, że coś z polskimi egzaminami jest nie tak, zdają się potwierdzać specjalistyczne analizy prowadzone przez środowiska naukowe. Warto nadmienić, że badacze edukacyjni wciąż w zbyt małym stopniu poddają gruntownej analizie jakość egzaminów zewnętrznych. Nieliczne dostępne analizy skupiają się na wypunktowywaniu słabości psychometrycznych egzaminów - zwraca się uwagę na niską rzetelność testów oraz problem porównywalności wyników między wersjami egzaminu (Pokropek, 2011), problemem jest także brak jasnego sprecyzowania funkcji poszczególnych egzaminów (Dolata i in., 2004), monitorowanie jakości egzaminów i szkolnictwa utrudnia zaś brak systemowego rozwiązania, pozwalającego na porównywanie egzaminów między latami (Szaleniec i in., w druku, 2012)[1].

Czasami zwraca się także uwagę na brak ustalonych standardów konstrukcji testów oraz procedur, które gwarantowałyby wysoką jakość narzędzi egzaminacyjnych (Dolata i in., 2004). Wydaje się bowiem, że problemy egzaminów, na które wskazuje intuicyjnie opinia publiczna, jak również bardziej specjalistyczne ich mankamenty dostrzegane przez badaczy, wynikają właśnie z przyjętej metody konstrukcji testów. Można argumentować, że to właśnie niewystarczające badania pilotażowe i związana z tym słabość analizy właściwości psychometrycznych zadań i testów, jeszcze zanim staną się narzędziami egzaminacyjnymi, przesądzają o niedostatecznej jakości ostatecznych narzędzi. Przy braku precyzyjnie określonych standardów i procedur, stworzenie dobrego psychometrycznie narzędzia służącego do pomiaru osiągnięć szkolnych jest rzeczywiście bardzo mało prawdopodobne. Istnieją jednak sposoby

[1] Pracownia Analiz Osiągnięć Uczniów Instytutu Badań Edukacyjnych prowadzi badania mające na celu zrównanie wyników egzaminów z ubiegłych lat, jednak nie jest to rozwiązanie systemowe. Nie zapewnia ono możliwości przedstawienia na jednej skali wyników z kolejnych egzaminów, bez ponawiania badań.

pozwalające na utrzymanie kontroli nad jakością konstruowanego narzędzia. Standardy tworzenia i ewaluacji narzędzi do pomiaru dydaktycznego, opisane w *Standards for Educational and Psychological Testing* (AERA, APA, NCME; 1999), stosowane są przez najbardziej renomowane i doświadczone instytucje konstruujące testy kompetencji i osiągnięć szkolnych. Kluczowym etapem, pozwalającym na konstrukcję dobrego psychometrycznie narzędzia, jest badanie pilotażowe, które stanowi punkt centralny całego procesu.

Instytut Badań Edukacyjnych w ramach *Badania szkolnych uwarunkowań efektywności kształcenia* w 2010 roku stanął przed możliwością zbliżenia się do światowych standardów w konstrukcji testów osiągnięć. W artykule tym przedstawione zostaną najważniejsze założenia i etapy konstrukcji narzędzi pomiarowych przyjęte przez zespół badawczy pracujący nad testami oraz rekomendacje płynące ze zdobytego doświadczenia. Ewaluacji zostanie poddane także zrealizowane badanie pilotażowe, aby pokazać znaczenie zastosowanych rozwiązań dla konstrukcji testów osiągnięć. Wnioski płynące z tego artykułu mogą dostarczyć argumentów dla myślenia o zmianach w polskim systemie egzaminacyjnym, a także mogą być inspirujące dla osób tworzących testy osiągnięć zarówno do celów badawczych, jak i dla potrzeb szkolnych.

Badanie szkolnych uwarunkowań efektywności kształcenia (SUEK)[2] to studium podłużne, które ma na celu identyfikację kluczowych czynników szkolnych warunkujących efektywność kształcenia w szkołach podstawowych. Badanie rozpoczęło się w roku szkolnym 2010/11 i objęło kohortę rozpoczynającą naukę w III klasach szkół podstawowych. Będzie ono kontynuowane co najmniej do czasu, w którym badani uczniowie ukończą szkołę podstawową. Badanie wymagało przygotowania narzędzi do pomiaru kluczowej zmiennej zależnej, czyli poznawczych wyników kształcenia. Pierwszym krokiem było opracowanie testów osiągnięć szkolnych podsumowujących nauczanie na I etapie kształcenia.

Etapy tworzenia testów osiągnięć szkolnych na potrzeby I pomiaru w badaniu SUEK

Proces przygotowania testów osiągnięć szkolnych na potrzeby badania SUEK można podzielić na 12 etapów (por. Downing, 2006a). Pierwszym etapem było określenie ogólnych założeń dotyczących pomiaru. Zdefiniowano populację, która ma zostać poddana badaniu[3]. Przyjęto teorię pomiaru, która rzutowała na sposób konstrukcji zadań testowych i schematów oceniania, wybór zadań po badaniu pilotażowym oraz skalowanie ostatecznych wyników z testów. Zdecydowano się na teorię odpowiedzi na zadanie testowe (*item response theory*, IRT) z uwagi na to, że jest to bardziej

[2] Realizowane przez Pracownię Szkolnych Uwarunkowań Efektywności Kształcenia w Instytucie Badań Edukacyjnych.

[3] Kohorta uczniów rozpoczynająca w roku szkolnym 2010/11 naukę w III klasie szkoły podstawowej. Są to uczniowie nieobjęci jeszcze reformą programową wprowadzoną rozporządzeniem Ministra Edukacji Narodowej z dnia 23 grudnia 2008 r.

adekwatny model analizowania właściwości zadań i testów, niezwykle pomocny w konstruowaniu testów dostosowanych do potrzeb, a także umożliwiający precyzyjne porównanie wyników różnych wersji testów[4] oraz tworzenie skal pionowych[5]. Założono, że wykorzystany zostanie najbardziej restrykcyjny model IRT, czyli model Rascha, który zakłada, że wszystkie zadania mają tak samo dobre właściwości pomiarowe (de Ayala, 2009, s. 33-34). Stworzenie testów z zadań spełniających założenie modelu Rascha jest niezwykle istotne z punktu widzenia komunikowania wyników w postaci informacji o osiągnięciu pewnego poziomu wiadomości i umiejętności[6], a także tworzenia tablic przeliczeniowych wyniku surowego na wynik przeskalowany[7], co umożliwia wykorzystywanie testów w diagnozie nauczycielskiej i porównanie otrzymanych wyników z określonymi normami.

Charakter badania podłużnego SUEK wymagał tego, by skonstruowane na jego potrzeby testy były testami szerokiego zasięgu, czyli takimi, które pozwalają na zdobycie jak największej ilości informacji o całej badanej grupie. Testy osiągnięć szkolnych musiały być więc tak przygotowane, by umożliwiały precyzyjny pomiar umiejętności wszystkich uczniów, niezależnie od poziomu ich wiedzy. Z uwagi na to, że poziomem analizy, który jest najbardziej interesujący z punktu widzenia badania SUEK, jest poziom klas[8], priorytetem jest rzetelne określenie efektów kształcenia na poziomie badanych klas. Niemniej ważny jest także jak najdokładniejszy pomiar wyników indywidualnych z uwagi na to, że dostrzega się możliwość, by po zakończeniu projektu powstałe testy służyły celom diagnostycznym. Tak postawione cele pociągały za sobą konieczność określenia warunków w stosunku do kształtu i położenia krzywej informacyjnej testu oraz krzywej charakterystycznej testu[9]. Spełnienie tych założeń wymagało dobrania do testu zadań z całego zakresu skali.

Określając ogólne założenia pomiaru, sprecyzowano także wytyczne dotyczące praktycznej strony konstrukcji testów takie jak: forma arkuszy (papierowa, do samodzielnego wypełniania), czas przewidziany na rozwiązanie testu, szacunkową liczbę zadań, harmonogram projektu. Zdecydowano się także na stworzenie dwóch quasi-równoległych wersji testu, z pulą zadań wspólnych (kotwiczących), aby wydłużyć test, co miało na celu zwiększenie trafności pomiaru przez

[4] Co miało znaczenie dla przygotowywanego pomiaru z uwagi na planowane skonstruowanie dwóch zakotwiczonych wersji testów.

[5] W badaniu SUEK kolejne pomiary osiągnięć planuje się przedstawiać na zrównanych skalach pionowych.

[6] Jest to możliwe dzięki temu, że w modelu Rascha uporządkowanie pytań ze względu na trudność jest takie samo dla każdego ucznia (każdego poziomu umiejętności), ponieważ krzywe charakterystyczne zadań się nie przecinają, a także umiejętności uczniów i trudności pytań są wyrażone na tej samej skali.

[7] Surowa liczba punktów jest statystyką dostateczną (de Ayala, 2009, s. 25).

[8] Efektywność nauczania przypisywana jest bowiem przede wszystkim klasie (to w klasie odbywa się nauczanie; wiele gromadzonych danych będą stanowiły zmienne z poziomu klasy), a w dalszej kolejności szkołom.

[9] Środek krzywej charakterystycznej, wyznaczonej dla testu dla danego poziomu nauczania (w tym przypadku klas III) powinien znajdować się w punkcie oznaczającym średni poziom umiejętności badanych uczniów z tego etapu nauczania. Maksimum funkcji informacyjnej także powinno lokować się w tym punkcie, a ponadto pole pod krzywą informacyjną powinno być jak największe, a wartości jej funkcji jak najwyższe w jak najszerszym spektrum skali.

lepsze pokrycie zadaniami szerokiego spektrum treści i umiejętności nauczanych na danym etapie szkoły podstawowej, czyli lepsze oszacowanie wyników nauczania na poziomie klasy, a także podniesienie wiarygodności wyników dzięki utrudnieniu ściągania wśród uczniów podczas badania testowego.

Na drugim etapie prac nad testami zdefiniowano obszary treści i umiejętności, które będą objęte pomiarem. Prace te opierały się na analizie podstawy programowej kształcenia ogólnego, krajowych, międzynarodowych i zagranicznych badań umiejętności oraz analizie wniosków płynących z projektu poświęconego nowej formule sprawdzianu dla klasy VI[10]. Etap ten pozwolił na wyłonienie najważniejszych z punktu widzenia kształcenia w szkole podstawowej obszarów, w których zakresie będzie można wnioskować o wynikach nauczania, i zdefiniowanie skal pomiarowych. Opracowana koncepcja skal pomiarowych podlegała konsultacjom merytorycznym z ekspertami i została poddana zewnętrznej recenzji. Zebrane uwagi pozwoliły na dopracowanie ostatecznego kształtu koncepcji (Jasińska, 2012). Pomiar osiągnięć szkolnych w badaniu SUEK objął umiejętność czytania, świadomość językową i umiejętności matematyczne. Opis struktury każdego z trzech testów oraz wynikających z niej planów testów miał na celu zagwarantowanie różnorodności i reprezentatywności mierzonych treści i umiejętności, a tym samym zapewnienie trafności treściowej testu.

Trzecim etapem było opracowanie planów testów dla każdej skali pomiarowej. Plany testów precyzują, ile zadań, mierzących jakie szczegółowe umiejętności powinno znaleźć się w teście. Proporcje zadań były ustalane na podstawie analizy podstawy programowej kształcenia ogólnego i programów nauczania.

Kolejne dwa etapy obejmowały tworzenie zadań testowych, zgodnych z założeniami pomiaru i planami testów. Opracowano zasady konstrukcji zadań testowych i schematów oceniania oraz wskazówki dla autorów zadań, które były pomocne przy układaniu zadań spełniających założone kryteria. W tworzeniu tych wskazówek niezwykle pomocne były doświadczenia amerykańskie (Haladyna i in., 2002; Downing, 2006b). Równolegle ogłoszono konkurs na autorów zadań. Do współpracy zaproszono osoby, których nadesłane próbki autorskich zadań zostały najwyżej ocenione. Praca nad pozyskiwaniem zadań polegała na zamawianiu u autorów zadań mierzących konkretne umiejętności tak, by odpowiednio wypełnić plan testu, recenzowaniu ich na bieżąco i, jeśli to było wskazane, odsyłaniu do poprawy. Pula ostatecznie przyjętych zadań podlegała dodatkowej ocenie i korekcie podczas warsztatów 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.

Na szósty etap prac nad testami składało się zaplanowanie i realizacja badania pilotażowego 823 przygotowanych zadań składających się na trzy skale pomiarowe. W przygotowaniu badania pilotażowego najważniejsze było odpowiednie podzielenie zadań na zeszyty testowe oraz opracowanie takiego planu

[10] *Nowa formuła sprawdzianu w klasie VI,* projekt realizowany w latach 2007-2010 przez CKE, koordynator: Anna Pregler.

testowania (określającego, którzy uczniowie mają rozwiązywać które zeszyty testowe), który zapewni zrównoważone próbkowanie macierzowe, umożliwiające wspólne skalibrowanie zadań z jednego testu oraz jak najdokładniejsze oszacowanie parametrów psychometrycznych zadań i każdego z trzech testów. Plan testowania musiał uwzględniać także fakt, iż badanie pilotażowe było realizowane z udziałem uczniów klas III i V[11]. Aby podnieść trafność fasadową testu pilotażowego, zdecydowano, że uczniowie klas III nie będą rozwiązywali zadań najtrudniejszych, a uczniowie klasy V - zadań najłatwiejszych. Biorąc pod uwagę wszystkie założenia oraz liczbę przygotowanych zadań do każdego z trzech testów, przygotowano do badania pilotażowego 44 zeszyty testowe. Plan testowania został ostatecznie przygotowany tak, by każdy zeszyt testowy występował z każdym innym dla pewnej grupy uczniów. Równoważył także prawdopodobieństwo rozwiązywania w próbie każdego zeszytu testowego. Ponadto uwzględniał różną kolejność rozwiązywania zeszytów testowych w poszczególnych dniach testowania. Każdy uczeń rozwiązywał 4 zeszyty z jednego testu. Realizację badania powierzono wykonawcy zewnętrznemu wyłonionemu w drodze przetargu publicznego.

Kolejnym etapem była analiza danych z badania pilotażowego mająca na celu wybór zadań o najlepszych właściwościach pomiarowych. Analizy wykonano przy pomocy oprogramowania ConQuest 2.0[12]. Wykorzystano model Rascha dla zadań punktowanych 0-1 oraz Partial Credit Model dla zadań wielopunktowych. Wszystkie zadania w ramach każdego testu były skalowane jednocześnie. Analizy polegały na wielokrotnym powtarzaniu następujących kroków: łączne wyskalowanie zadań z danego testu, usunięcie z analiz kilku zadań najsłabiej dopasowanych do założonego modelu lub skrócenie skali punktowej dla zadań źle dopasowanych, ponowne wyskalowanie zadań z danego testu[13]. Podejmując decyzję o usunięciu zadań, brano pod uwagę: dopasowanie zadania do modelu (miary dopasowania: INFIT - weighted fit, OUTFIT - unweighted fit; kształt empirycznych krzywych charakterystycznych zadań w stosunku do krzywych wynikających z modelu), trudność zadania, wnioski wyprowadzane z analizy dystraktorów, wnioski wyprowadzane z analiz efektów zróżnicowanego funkcjonowania zadania DIF (*differential item functioning*). Analizy doprowadziły do utworzenia banku zadań o wystarczająco

[11] 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. Harmonogram pracy nad testami niestety nie pozwolił na realizację badania pilotażowego w czasie najbardziej wskazanym - na dokładnie rok przed planowanym pomiarem zasadniczym (czyli na początku roku szkolnego 2010/11). Dlatego zdecydowano się objąć badaniem uczniów trochę młodszych oraz trochę starszych, ze świadomością tego, że nie jest to rozwiązanie idealne. Wybór klasy V był podyktowany tym, że część zadań przygotowywanych do pierwszego pomiaru osiągnięć miała być z założenia zadaniami kotwiczącymi dla kolejnych pomiarów obejmujących uczniów po zakończeniu nauki w klasie IV. Istniała więc konieczność przetestowania ich na populacji uczniów starszych. Takie zdefiniowanie populacji uczniów objętych badaniem pilotażowym było jednak korzystne z powodu możliwości przetestowania zadań wśród uczniów o bardzo zróżnicowanych umiejętnościach.

[12] Margaret L. Wu i in., *ACER ConQuest version 2.0: Generalised Item Response Modelling Software*, ACER Press, Australian Council for Educational Research, 2007.

[13] Usunięcie każdego zadania powoduje zmiany w oszacowaniu parametrów pozostałych zadań (zmiana definicji mierzonej cechy, szacowanie umiejętności uczniów i parametrów zadań na podstawie trochę innego zestawu danych). Dlatego ważne jest to, by usuwanie zadań z analiz było czynione w małych krokach.

dobrych właściwościach pomiarowych, dopasowanych do założonego modelu analiz, który stanowił podstawę konstrukcji testów do badania zasadniczego.

Tworzenie ostatecznych wersji testów, czyli ósmy etap prac, polegał na takim wyborze zadań o najlepszych właściwościach psychometrycznych, by mierzyły one założone przez plany testów wiadomości i umiejętności, a także by były jak najlepiej dostosowane do prognozowanego rozkładu umiejętności uczniów w populacji docelowej oraz miały odpowiednio zróżnicowaną trudność[14]. Każda propozycja testu była skalowana, aby móc ocenić i porównać ją z innymi. Po ustaleniu ostatecznej puli zadań tworzących test osiągnięć, podzielono zadania na dwie quasi-równoległe wersje z pulą zadań kotwiczących. Podziału dokonywano w taki sposób, by zarówno zadania kotwiczące, jak i zadania w każdej wersji testu stanowiły reprezentatywną próbką planu testu, a także by obie wersje miały taką samą trudność. W ten sposób powstały trzy testy zawierające od 43 do 51 zadań (od 29 do 33 zadań przypadających na wersję). Zadania z każdego testu rozłożono na dwa zeszyty testowe dla każdej wersji.

Harmonogram projektu uniemożliwiał przeprowadzenie badania standaryzacyjnego dla powstałych testów. Testy przygotowane w opisany sposób zostały wykorzystane w badaniu zasadniczym, które można potraktować także jako badanie ewaluacyjne przyjętej procedury ich konstrukcji. Był to dziewiąty etap prac nad testami. Badanie to zostało przeprowadzone na ogólnopolskiej, losowej próbie 174 szkół podstawowych w listopadzie 2011 roku. Badaniem zostało objętych 5156 uczniów rozpoczynających naukę w klasie IV. Ostatnie trzy etapy konstrukcji testów osiągnięć zawierały analizę danych z badania zasadniczego, mającą na celu oszacowanie ostatecznych parametrów zadań, wyskalowanie testów i obliczenie wyników dla uczniów biorących udział w badaniu oraz opracowanie raportu technicznego wraz z normami wykonania podsumowującego proces powstawania i właściwości stworzonych testów (Jarnutowska i in., w przygotowaniu).

Znaczenie badania pilotażowego dla konstrukcji testów osiągnięć szkolnych

Jak wynika z powyższego opisu kolejnych etapów prac, narzędzia, jakimi są testy osiągnięć szkolnych stworzone na potrzeby *Badania szkolnych uwarunkowań efektywności kształcenia,* są narzędziami zaprojektowanymi. Z zadań przetestowanych na etapie badania pilotażowego skonstruowano docelowe testy, co do których sformułowano określone hipotezy dotyczące ich jakości i właściwości psychometrycznych. Przyjrzyjmy się dwóm z tych najważniejszych, a dotyczących zakresu stosowalności i dopasowaniu testów do modelu.

Po ponownym wyskalowaniu zadań wchodzących w skład zasadniczych testów, bazując już na wynikach badania na populacji docelowej, można było ocenić, na ile udało się te założenia udowodnić. W pilotażu na każde zadanie odpowiedziało około 300 uczniów, pochodzili oni jednak z dwóch różnych populacji, które stanowiły niejako warunki brzegowe dla zamierzonej populacji.

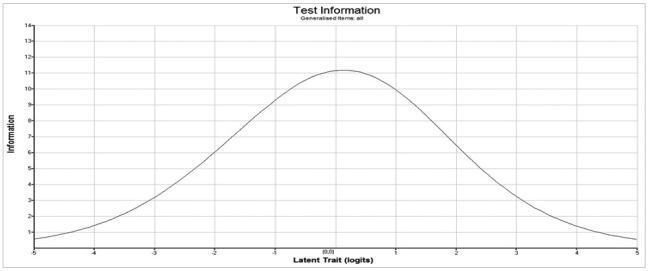
[14] Aby krzywa informacyjna testu była jak najwyższa w jak najszerszym obszarze poziomu umiejętności uczniów.

Przy konstrukcji testów przyjęto założenie, że docelowe narzędzia powinny dostarczać najwięcej informacji w takim przedziale umiejętności, który zdefiniować można jako średni poziom badanej umiejętności - uczniowie, którzy bowiem zakończyli edukację na pierwszym etapie kształcenia, znajdują się w kontekście poziomu swoich umiejętności gdzieś pomiędzy uczniami na początku klasy trzeciej a uczniami na początku klasy piątej szkoły podstawowej. Czy udało się zatem odtworzyć rozkład trudności zadań w badaniu zasadniczym? Czy skonstruowane testy rzeczywiście mierzą założony obszar umiejętności? IRT dostarcza narzędzia, które pozwala odpowiedzieć na te pytania.

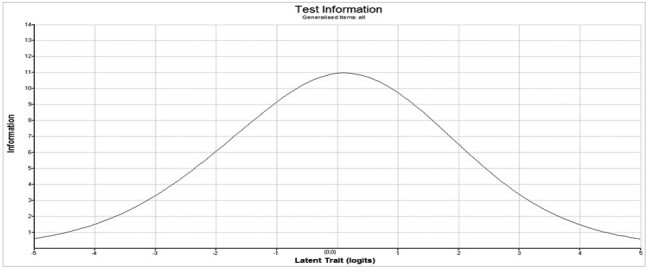
Krzywa informacyjna testu jest przydatnym narzędziem pozwalającym graficznie przedstawić zdolność testu do szacowania poziomu umiejętności w różnych jej zakresach (de Ayala, 2009, s. 31-33). Im większa wartość funkcji informacyjnej, tym większa precyzja pomiaru badanej cechy. W zależności od uwzględnionych w teście zadań, przyjętego modelu estymacji, krzywa informacyjna dla testu może przybierać różne kształty. Narzędzie to pozwala więc dosłownie na projektowanie własności pomiarowych testu, w zależności od potrzeb - wystarczy określić pożądany kształt krzywej. Krzywa informacyjna dla idealnego testu osiągałaby wysoką wartość maksimum dla średniego wyniku w populacji i utrzymywałaby wysoki wynik aż po skraje skali. Znaczyłoby to - przy zachowaniu założeń modelu - że test pozwala na bardzo dokładny pomiar umiejętności wśród wszystkich członków badanej populacji, niezależnie od ich poziomu umiejętności. Narzędzie takie musiałoby składać się z bardzo wielu zadań, o bardzo zróżnicowanej trudności[15].

Algorytm używany przez program ConQuest 2.0 przedstawia wyniki estymacji parametrów zadań na skali logitowej, przyjmując jako punkt *zero* średnią trudność zadań w teście. Po oszacowaniu parametrów zadań program estymuje średni poziom umiejętności w populacji na podstawie wyników uczniów w próbie. Zakłada się, że rozkład umiejętności w populacji jest normalny. Jaki kształt krzywej informacyjnej testu jest więc pożądany w przypadku testów dla badania SUEK? Chęć zbliżenia się do ideału testu nakreślonego powyżej, i założenie jako celu pomiaru jak najdokładniejszego zmierzenia poziomu umiejętności uczniów w całej populacji, prowadzą nas do wniosku, że najlepszą krzywą jest krzywa w kształcie rozciągniętego wszerz dzwonu, która osiąga maksimum dla średniego wyniku w populacji. Test o takiej krzywej pozwala na najdokładniejsze rozróżnienie poziomu umiejętności uczniów w obszarze, który jest najgęściej „zaludniony", dostarcza więc najwięcej informacji dla największej części populacji. Rysunki 1. i 2. przedstawiają porównanie krzywych informacyjnych dla testu z matematyki z etapu pilotażu i badania zasadniczego. W tabeli 1. umieszczono zaś zestawienie średnich wyników uczniów dla każdego z testów z obu tych etapów.

[15] Należy jednak pamiętać, że dokładność pomiaru, wraz z czasem potrzebnym na jego przeprowadzenie, z zasady znajdują się w relacji monotonicznie rosnącej: im więcej zadań w teście, tym dokładniejszy pomiar. Jednym ze sposobów na zwiększenie precyzji pomiaru, przy zachowaniu rozsądnego czasu badania, stanowi testowanie adaptatywne: poziom umiejętności badanego szacowany jest na bieżąco, a każde kolejne zadanie dobierane jest ze względu poprzednie odpowiedzi badanego. W ten sposób narzędzie zatacza coraz węższe okrążenia wokół najpewniejszego oszacowania poziomu umiejętności badanego, aż do osiągnięcia pożądanego stopnia precyzji.



Rysunek 1. Krzywa informacyjna dla testu z matematyki z etapu pilotażu



Rysunek 2. Krzywa informacyjna dla testu z matematyki dla badania zasadniczego

Tabela 1. Porównanie średnich wyników uczniów w testach SUEK z etapu pilotażu i badania zasadniczego

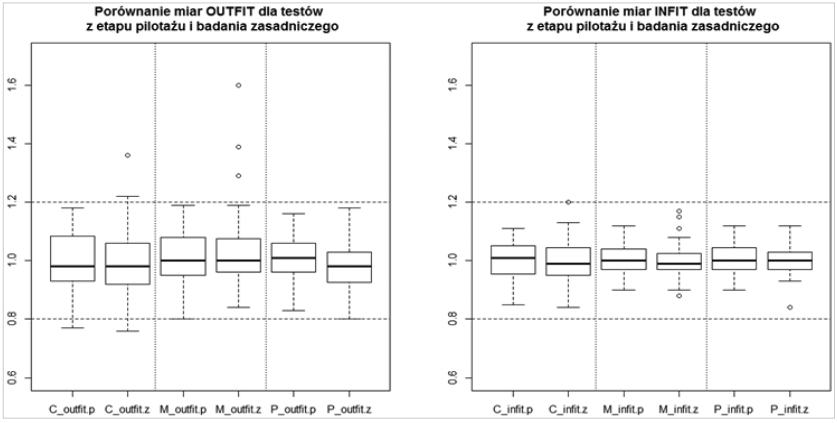
Etap / Rodzaj testu	Test umiejętności matematycznych	Test czytania	Test świadomości językowej
Pilotaż	0,182	0,745	-0,041
Badanie zasadnicze	0,263	0,512	-0,117

Krzywe informacyjne dla pozostałych testów wyglądają bardzo podobnie do krzywych dla testu matematycznego - oznacza to, że rozkład względnej trudności zadań oszacowany na podstawie badania pilotażowego potwierdził się w badaniu zasadniczym. Należy przy tym przypomnieć, że badanie pilotażowe obejmowało o wiele więcej zadań i że każdy uczeń rozwiązywał na tym etapie tylko niewielką próbkę z nich. Pomimo tego, dzięki zastosowaniu modelu Rascha, zadania testowane często na różnych podpróbach uczniów, można było połączyć w jedno narzędzie, a ich parametry nie uległy wielkiej zmianie. W związku z tym, narzędzie, które zostało zaprojektowane z najlepszych psychometrycznie zadań, nietestowane w całości na żadnej podróbce uczniów (żaden z uczniów w pilotażu nie zetknął się z całą wiązką zadań, które weszły do zasadniczego narzędzia), potwierdziło w badaniu zasadniczym swoje dobre właściwości pomiarowe.

Można jednak zauważyć, że przewidywane średnie umiejętności uczniów nie wypadają ściśle w punkcie *zero*, zarówno w pilotażu, jak i w badaniu zasadniczym. Pierwszą informacją, jaką należy z tego porównania średnich wywnioskować, to fakt, że przewidywana na etapie pilotażu średnia umiejętność uczniów w docelowej populacji, nie odbiega znacząco - z wyłączeniem

testu czytania - od tej uzyskanej na podstawie danych z badania zasadniczego. Przyjęte założenie o średniej umiejętności uczniów w docelowej populacji okazało się w dużej mierze słuszne. Choć testy skonstruowane po analizie danych z pilotażu nie były idealnie dopasowane do zakładanej (a nie badanej, dodajmy) populacji, to odchylenia te były z zasady niewielkie i wynikały z nałożenia na siebie kilku ograniczeń - dobrego dopasowania zadań do restrykcyjnego modelu, ścisłych planów testów oraz dostępności zadań o zróżnicowanej trudności. To właśnie brak dobrych psychometrycznie zadań o wysokiej trudności w teście czytania uniemożliwił stworzenie narzędzia lepiej dopasowanego do badanej populacji. Jak się jednak okazało, w przypadku testu czytania niedopasowanie to zmniejszyło się po ostatecznym oszacowaniu średniej w interesującej nas populacji.

Kolejnym problemem, którym się zajmiemy, to dopasowanie zadań do modelu. Tylko zadania, które wykazywały wysoki poziom zgodności z przewidywaniami modelu Rascha mogły zostać włączone do ostatecznego narzędzia. Czy zadania, które „zachowywały się" dobrze na etapie pilotażu, nie straciły tej własności w docelowej populacji? Miary dopasowań INFIT i OUTFIT, które były brane pod uwagę przy włączaniu zadań do testu, zdają sprawę ze zgodności przewidywań modelu i danych empirycznych. Gdy miary te osiągają jeden, przyjmuje się, że zadanie jest dobrze dopasowane do modelu. W praktyce dopuszcza się jednak pewne rozchwianie wartości tych statystyk (de Ayala, 2009, s. 55-57). Na etapie badania pilotażowego zadania, których wartości dla miar dopasowania wahały się od ok. 0,8 do ok. 1,2, uważano za odpowiednio dobrze dopasowane. Rysunek 3. przedstawia wykresy skrzynkowe dla miar dopasowania z etapu pilotażu i badania zasadniczego.



Rysunek 3. Wykres skrzynkowy dla miar dopasowania dla testów (prefiks „C" - czytanie, „M" - matematyka, „P" - test świadomości językowej) dla dwóch etapów (sufiks „p" - pilotaż, „z" - badanie zasadnicze)

Wykres ten przedstawia rozkład statystyk dopasowania: dolna granica każdej skrzynki odpowiada wartości dla pierwszego kwartyla, gruba linia wewnątrz skrzynki przedstawia medianę, górny bok skrzynki to trzeci kwartyl. Wąsy rozciągają się do wartości maksymalnej i minimalnej dla każdej zmiennej

lub maksymalnie do wartości wyznaczonej przez następujący przedział: <Q1 - 1,5*IQR; Q3 + 1,5*IQR>, gdzie Q1, Q3 to odpowiednio wartości dla pierwszego i trzeciego kwartyla, a IQR to wartość rozstępu ćwiartkowego. Wartości odstające oznaczone są kółkami.

Widzimy, że miary dopasowań dla zadań w większości utrzymują się w założonych granicach (oznaczonych na wykresie poziomymi przerywanymi liniami). Oznacza to, że przewidywania dla dopasowania zadań testowych do modelu w bardzo dużym stopniu potwierdziły się na etapie badania zasadniczego. Na etapie tym jednakże, kilka zadań w teście matematycznym i jedno zadanie w teście czytania wykroczyło poza wyznaczone limity. Część zadań, które nie wykazują się dobrym dopasowaniem do modelu na etapie badania zasadniczego, zachowano ze względu na potrzebę respektowania planu testu. Zadania rażąco odbiegające od przewidywań modelu nie zostały zaś ostatecznie wzięte pod uwagę przy szacowaniu poziomu umiejętności uczniów z docelowej populacji (dotyczy to po jednym zadaniu z testu matematycznego i testu czytania).

Należy jeszcze raz podkreślić, że w związku z brakiem badania standaryzacyjnego, ostateczne oszacowania parametrów dla zadań i ocena jakości narzędzia zostały przeprowadzone na podstawie wyników badania zasadniczego. Dopiero na tym etapie, zadania wybrane na podstawie danych z pilotażu zostały faktycznie ułożone obok siebie w formie papierowych zeszytów. Dopiero na tym etapie, na każde zadanie kotwiczące odpowiedziało ponad 5000 uczniów z docelowej populacji, a na każde zadanie specyficzne dla wersji około połowa tej liczby. Jednak pomimo tak wielu różnic w okolicznościach przeprowadzania badania pilotażowego i zasadniczego, zaprojektowane na podstawie wyników pilotażu narzędzie okazało się spełniać w bardzo wysokim stopniu pokładane w nim nadzieje - zadania wykazują dużą zgodność z przewidywaniami modelu, testy mierzą umiejętności z dużą dokładnością w szerokich zakresach. Owocem długiego, przemyślanego i złożonego procesu tworzenia testów osiągnięć szkolnych są narzędzia z pewnością nie doskonałe, ale będące silnym świadectwem, że gdy chodzi o możliwość stworzenia dobrego narzędzia do pomiaru efektów kształcenia, to istnieją metody, które zbliżają nas do jej zrealizowania.

Rekomendacje

Na zakończenie warto podzielić się kilkoma spostrzeżeniami dotyczącymi tworzenia testów osiągnięć szkolnych. Na pierwszy plan wysuwa się kwestia założeń dotyczących pomiaru: dokładna wizja końcowego narzędzia, jego formy, zakresu stosowalności i celu pomiaru. Jasna koncepcja testu pozwala wytyczyć drogę prowadzącą do wytworzenia narzędzia o pożądanych własnościach. Drugim, równie ważnym etapem jest pozyskanie dobrych zadań. W Polsce wciąż cierpimy na brak profesjonalnych autorów zadań do testów osiągnięć. Na szczęście nie jest tak, jak wielu osobom może się wydawać, że dobrym autorem zadań trzeba się urodzić - tworzenia dobrych zadań można się nauczyć. Niestety problematyka pomiaru osiągnięć szkolnych na uniwersytetach w Polsce nie ma silnego zaplecza naukowego. Nie organizuje się studiów, które kształciłyby studentów w tym kierunku. Po drugie, należy zwrócić uwagę na

empiryczną stronę oceny zadań - dobre zadanie to takie, które *faktycznie* dobrze mierzy założony konstrukt. Oczywiście nie znaczy to, że ocena jakościowa jest zbędna - tylko z pomocą oka eksperta jesteśmy w stanie określić, czy zadanie jest trafne, czy jest w ogóle w stanie mierzyć to, co chcemy zmierzyć - jednakże ocena, na *ile* dobrze to robi, jest już kwestią, którą można rozstrzygnąć tylko na drodze empirii, czyli sprawdzając je w badaniu pilotażowym.

Warto w tym miejscu wspomnieć o potrzebie jak najlepszego dopracowywania zadań jeszcze na etapie przed pilotażem. W związku z tym, że badanie pilotażowe - zorganizowane w opisany powyżej sposób - pozwala na przetestowanie wielu zadań, nie warto dopuszczać do niego zadań ewidentnie słabych konstrukcyjnie, na zasadzie: „jeśli jest słabe, to nie wejdzie do ostatecznej puli zadań". Przewidywania te bardzo często się sprawdzają, jednakże, kierując się taką zasadą, zmarnowaliśmy właśnie miejsce w puli zadań pilotażowych na zadanie, którego nie mieliśmy zamiaru używać, tylko dlatego, że nie dopracowaliśmy się lepszego na jego miejsce. Bardzo dobrym pomysłem, wydłużającym niestety czas potrzebny na przygotowanie narzędzia i zwiększającym koszty, jest wprowadzenie dwuetapowego pilotażu, gdzie - pomiędzy kolejnymi edycjami pilotażu - na bazie faktycznych odpowiedzi uczniów jesteśmy w stanie ulepszyć zadania lub klucze kodowe, służące ocenie zadań otwartych.

Pracownia SUEK wykorzystała do przygotowania narzędzia szczególny przypadek modelu IRT - model Rascha. Model ten posiada bardzo cenne zalety pozwalające na konstrukcję narzędzia w zależności od potrzeb - niektóre zostały zaprezentowane w tekście. Nie jest jednak tak, że autorzy są przekonani, że metoda ta jest jedyną słuszną metodą pozwalającą na konstruowanie dobrych narzędzi psychometrycznych. Główną tezą postawioną w tekście jest przekonanie, że dobre narzędzie jest owocem przemyślanego procesu, w którym kolejne decyzje dotyczące tworzenia testu nie są podejmowane arbitralnie, lecz wynikają z metody, dostarczającej obiektywnych, empirycznych przesłanek dla rzetelnej oceny postępów i jednoznacznych procedur zapewniania jakości wytwarzanemu produktowi. Pracownia SUEK podjęła się konstrukcji testów osiągnięć od podstaw nie tylko dlatego, że posiadanie dobrego i rzetelnego narzędzia służącego pomiarowi osiągnięć szkolnych jest konieczne dla powodzenia *Badania szkolnych uwarunkowań efektywności kształcenia*, ale też dlatego, żeby pokazać na gruncie polskim, że można stworzyć testy osiągnięć we w pełni zaplanowany i przemyślany sposób i że to się opłaca. Słowem, że można inaczej.

Bibliografia:

1. AERA, APA, NCME (1999), *Standards for Educational and Psychological Testing*.
2. de Ayala R. J. (2009), *The Theory and Practice of Item Response Theory*, New York London, The Guilford Press.
3. Dolata R., Putkiewicz E., Wiłkomirska A. (2004), *Reforma egzaminu maturalnego*

- *oceny i rekomendacje*, Warszawa, Wydawnictwo Instytutu Spraw Publicznych.
4. Downing S. M. (2006a), *Twelve Steps for Effective Test Develpopment* [w:] *Handbook of Test Development*, S. M. Downing, T. M. Haladyna (red.), Lawrence Erlbaum Associates Publishers.
5. Downing S. M. (2006b), *Selected-Response Item Formats in Test Development* [w:] *Handbook of Test Development*, S. M. Downing, T. M. Haladyna (red.), Lawrence Erlbaum Associates Publishers.
6. Haladyna T., Downing S., Rodriguez M. (2002), *A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment Applied Measurement in Education*, „Applied Measurement in Education", 15(3), 309-334.
7. Jarnutowska E., Jasińska A., Modzelewski M. [w przygotowaniu], *Testy osiągnięć szkolnych dla uczniów po pierwszym etapie kształcenia. Badanie szkolnych uwarunkowań efektywności kształcenia*, *Raport techniczny,* dostęp na stronie: http://www.eduentuzjasci.edu.pl/pl/suek.
8. Jasińska A. (2012), *Koncepcja testów osiągnięć szkolnych uczniów szkół podstawowych. Badanie szkolnych uwarunkowań efektywności kształcenia*, dostęp na stronie: http://www.eduentuzjasci.edu.pl/pl/suek.
9. Pokropek A. (2011), *Matura z języka polskiego. Wybrane problemy psychometryczne [w:] Ewaluacja w edukacji: koncepcje, metody, perspektywy. Materiały XVII Konferencji Diagnostyki Edukacyjnej*, Niemierko B., Szmigel M.K. (red.), Kraków, Polskie Towarzystwo Diagnostyki Edukacyjnej.
10. Szaleniec H., Grudniewska M., Kondratek B., Kulon F., Pokropek A. [w druku, 2012], *Zrównanie egzaminu gimnazjalnego dla lat 2002-2010*, „Edukacja" nr 3 (119).