**dr Maciej Jakubowski**
OECD, Directorate for Education
Uniwersytet Warszawski

# Międzynarodowe badania umiejętności uczniów a badania w Polsce[1]

Międzynarodowe badania umiejętności uczniów mają już ponad 40-letnią tradycję. Pierwsze tego rodzaju projekty pojawiły się pod koniec lat 50., a przez ostatnie kilkanaście lat mamy do czynienia ze znacznym wzrostem ich popularności. Rozwój tych badań jest ściśle związany z rozwojem metodologii pomiaru umiejętności uczniów, metodologii badań sondażowych, metodologii analizy danych pochodzących z takich badań, a także wykorzystaniem badań naukowych w kreowaniu polityki edukacyjnej. Można stwierdzić, że badania międzynarodowe miały ogromny wpływ zarówno na metodologię badań, jak i na ich wykorzystanie. Z jednej strony wyzwania stojące przed realizującymi te projekty, a także niemałe zasoby ludzkie i finansowe zaangażowane w ich urzeczywistnienie, spowodowały, że badania te od lat wyznaczają kierunki rozwoju, którymi często podążają instytucje realizujące podobne prace w pojedynczych krajach. Z drugiej strony, możliwość porównania umiejętności uczniów między krajami, odniesienia ich do kontekstu społeczno-ekonomicznego, ale także do metod nauczania czy organizacji systemu szkolnictwa, spowodował ogromne zainteresowanie wynikami tych badań wśród polityków, nauczycieli i rodziców.

Przykładem może być badanie PISA, największe obecnie badanie międzynarodowe uczniów, które w krajach, takich jak Niemcy, przyniosło prawdziwą rewolucję w myśleniu o własnym systemie edukacji. W wielu krajach wyniki pierwszej edycji PISA były w 2000 roku prawdziwym szokiem (np. wspomniane Niemcy, ale i Francja czy Norwegia), w innych wysoki poziom umiejętności uczniów był zaskoczeniem, np. w Finlandii czy Nowej Zelandii. Co więcej, okazało się, że kolejne edycje badania PISA, a także wyniki innych badań w rodzaju PIRLS czy TIMSS, potwierdziły wysoką pozycję krajów, takich jak Finlandia czy Korea Południowa, a także słabe wyniki uczniów w takich krajach jak Norwegia czy Portugalia. Badania te odnotowały także dość zaskakujące zmiany w wynikach uczniów, w tym wzrost przeciętnych osiągnięć połączony ze spadkiem ich zróżnicowania w Polsce. Wszystkie te obserwacje szeroko dyskutowane są nie tylko w świecie akademickim, ale i wśród polityków coraz częściej zainteresowanych jakością kształcenia w ich krajach.

[1] Tekst ten częściowo pokrywa się z opisem badania PISA oraz opisami baz danych PISA, PIRLS i TIMSS, przedstawionymi w książce M. Jakubowskiego i A. Pokropka, „Badając egzaminy", wydanej niedawno przez Centralną Komisję Egzaminacyjną i dostępnej w wersji elektronicznej w dziale publikacje na stronie www.ewd.edu.pl.

Można więc stwierdzić, że międzynarodowe badania umiejętności uczniów miały znaczący wpływ zarówno na metody wykorzystywane w badaniach edukacyjnych, jak i na politykę edukacyjną. W różnych krajach jednak wpływ ten miał odmienny charakter. Kraje w rodzaju USA, gdzie istnieją krajowe badania stanowiące pod względem metodologicznym wzorzec dla innych, a także dające często znacznie lepszą podstawę do analiz na potrzeby polityki edukacyjnej, są z reguły mniej zainteresowane dogłębnym wykorzystaniem międzynarodowych badań i przenoszeniem ich metodologii na własny grunt. Kraje te wykorzystują badania w rodzaju PISA czy PIRLS głównie do porównań poziomu umiejętności z innymi państwami. Jednak w wielu innych krajach badania międzynarodowe stanowią główne źródło wiedzy nie tylko o poziomie umiejętności uczniów, ale i o zależnościach między umiejętnościami a cechami rodziny ucznia, czy też metodami dydaktycznymi lub organizacją pracy szkoły. Co więcej, w krajach tych metodologia badań międzynarodowych może stanowić źródło inspiracji dla badań krajowych. Do krajów tego rodzaju niewątpliwie należy Polska, gdzie kolejne cykle badania PISA, czy zrealizowany po raz pierwszy w 2006 roku PIRLS, stanowią główne źródło wiedzy o populacji uczniów.

Niniejsza praca ma na celu omówienie podstawowych cech wyróżniających metodologię międzynarodowych badań edukacyjnych i odniesienie ich do badań prowadzących w Polsce. Zaczniemy od przybliżenia historii międzynarodowych badań umiejętności uczniów. Następnie krótko omówimy różnice między współczesnymi badaniami PISA, TIMSS oraz PIRLS. W kolejnej części skupimy się na metodologii pomiaru i skalowania umiejętności uczniów w badaniach międzynarodowych, aby na koniec odnieść ją do badań i metod pomiaru stosowanych w Polsce.

**Historia międzynarodowych badań uczniów**

Pierwsze międzynarodowe badanie zrealizowano w latach 1959-62 przez grupę badaczy, którzy stanowili podstawę założonej wkrótce organizacji IEA (International Association for the Evaluation of Educational Achievement). Badanie to stanowiło pilotaż przyszłych badań międzynarodowych. Miało na celu określenie, czy możliwa jest realizacja reprezentatywnych badań dających porównywalny międzynarodowo pomiar wiedzy uczniów z różnych dziedzin. W badaniu tym przebadano umiejętności 13-latków w 12 uczestniczących krajach, z zakresu matematyki, czytania, geografii, nauk ścisłych oraz zdolności niewerbalnych.

Pozytywne rezultaty tego pilotażowego projektu zachęciły do kolejnych badań w rodzaju First International Mathematics Study mierzącego umiejętności matematyczne w 12 krajach wśród populacji 13-latków oraz uczniów na ostatnim etapie nauczania przed studiami wyższymi. Kolejne badania poszerzały stopniowo zakres zebranej wiedzy o poziom umiejętności w innych dziedzinach (np. nauki ścisłe, czytanie, literatura, język angielski i francuski, edukacja obywatelska), a także w innych populacjach: szkoły podstawowe, szkoły średnie niższe (obecnie nasze gimnazja), szkoły średnie wyższe (obecnie nasze szkoły średnie).

Stopniowo zmieniała się metodologia badań. Pierwsze badania podawały wyniki na prostej skali sumy poprawnych odpowiedzi, w kolejnych zaczęto skalować wyniki metodami IRT. Umożliwiło to znaczne zwiększenie precyzji pomiaru, a także porównywalności międzynarodowej wyników. Zaczęto stosować metodę rotacyjnych *booklets*, która dodatkowo zwiększa porównywalność wyników dzięki temu, że zwiększa się liczba zadań, na które muszą odpowiedzieć uczniowie. Obniża to ryzyko, że któreś z zadań wypada gorzej w danym kraju ze względu na odmienny program nauczania czy różnice kulturowe, a także ogranicza błąd pomiaru. Coraz bardziej zaawansowane metody analizy statystycznej odpowiedzi na zadania dodatkowo umożliwiały wyłapywanie „dziwnie" zachowujących się zadań poprzez porównanie odpowiednich parametrów w różnych krajach. Dobór próby zastępowano bardziej złożonym losowaniem, dzięki któremu zwiększała się precyzja badania. Zaczęto stosować także odpowiednie algorytmy szacowania precyzji wyników, które obecnie biorą pod uwagę zarówno złożoność doboru prób, jak i niepewność pomiaru umiejętności uczniów.

**Różnice między badaniami PISA, PIRLS oraz TIMSS**

W ostatnich kilkunastu latach 3 międzynarodowe badania zdecydowanie zdominowały inne propozycje. Największym z nich jest badanie PISA koordynowane przez OECD (Organizacja Współpracy Gospodarczej i Rozwoju) i realizowane przez PISA Consortium, w którym liderem jest australijski ACER, a grono ekspertów zawiera przedstawicieli kilkudziesięciu krajów. W najnowszej edycji PISA 2009 uczestniczy 67 krajów, które ogółem odpowiadają za produkcję niemal 90% światowego PKB. Można więc stwierdzić, że badanie to stanowi dobrą podstawę dla porównania umiejętności 15-latków w najważniejszych gospodarkach świata. Wyniki badania zawsze odnoszone są do średniej i wariancji w krajach OECD, dzięki czemu nabierają łatwej interpretacji. Badanie PISA dotyczy umiejętności z zakresu czytania, matematyki oraz nauk ścisłych, przy czym każda z tych dziedzin mierzona jest dokładniej w danym cyklu. I tak, w 2000 główną dziedziną było czytanie, podobnie jak w 2009. W 2003 nacisk położono na matematykę, która ponownie będzie dokładniej mierzona w 2012. Cykle w 2006 oraz 2015 skoncentrowane są na pomiarze w zakresie nauk ścisłych.

Badania PIRLS oraz TIMSS realizowane są pod przewodnictwem wspomnianej już IEA. PIRLS poświęcony jest umiejętnościom czytania wśród 4-klasistów. TIMSS mierzy umiejętności matematyczne oraz z nauk ścisłych wśród 4- i 8-klasistów. W ostatniej edycji PIRLS w 2006 roku uczestniczyło ok. 40 krajów, w tym Polska, a w TIMSS 2007 wzięło udział ok. 60 krajów. Grupa krajów uczestniczących w badaniach PIRLS oraz TIMSS jest znacznie bardziej zróżnicowana, zawiera sporo krajów rozwijających się, a także poszczególne prowincje (kraje, regiony) z państw o ustroju federacyjnym, np. Kanady. Kolejne edycje PIRLS i TIMSS zaplanowane są na 2011 rok, a ciekawym przedsięwzięciem było rozszerzenie w 2008 roku badania TIMSS o opcję mierzącą umiejętności uczniów wybierających zaawansowane ścieżki nauczania, w której uczestniczyło kilkanaście krajów.

Najczęściej wskazywaną różnicą między badaniem PISA a badaniami PIRLS oraz TIMSS jest to, że mierzone umiejętności definiowane są nieco inaczej. W badaniu PISA z założenia mierzy się umiejętności potrzebne we współczesnym społeczeństwie i na nowoczesnym rynku pracy. Z tego względu tzw. *assessment framework* jest w pewnym stopniu niezależny od tego, czego uczy się w szkołach. TIMSS, a w nieco mniejszym stopniu PIRLS, podejmują próbę stworzenia międzynarodowego wzorca programu nauczania, który stanowić mógłby podstawę tworzenia testów. Inaczej mówiąc, badania te znacznie częściej odwołują się do tego, czego uczy się w szkołach. Oczywiście, w praktyce zadania testowe w każdym przypadku przynajmniej częściowo odnoszą się do kontekstu nauczania i nawiązują do treści programowych. Nie może więc dziwić, że średnie wyniki uczniów uzyskiwane w tych badaniach są ze sobą silnie skorelowane[2].

Wyznacznikiem PISA jest to, że badaną populację stanowią 15-latkowie. Dzięki temu na porównywalność wyników nie mają wpływu różnice w wieku uczniów. Badania TIMSS oraz PIRLS są reprezentatywne dla uczniów na danym poziomie nauczania (4 lub 8 klasa), a przez to zróżnicowanie wieku jest tu znacznie większe, a jego związek z wynikami nieco silniejszy. Z tego względu porównania wyników różnych krajów w PIRLS lub TIMSS są czasem silnie uzależnione od tego, w jakim wieku byli testowani uczniowie w tych krajach. Na przykład w PIRLS 2006 najlepiej wypadli uczniowie z Rosji, którzy jednak byli średnio starsi od uczniów polskich o 9 miesięcy. Z tego względu wyniki PIRLS oraz TIMSS warto porównywać między krajami, biorąc pod uwagę różnice w wieku uczniów, których nieuwzględnienie może prowadzić do całkowicie błędnych wniosków[3].

PISA różni się także od TIMSS oraz PIRLS punktem odniesienia. W PISA wyniki uczniów są skalowane tak, aby miały tę samą średnią i wariancję (odpowiednio 500 oraz 100) dla krajów OECD dla pierwszego badania danej domeny. I tak, skala wyników czytania miała średnią 500 i wariancję 100 w PISA 2000, gdzie czytanie było główną domeną, podobnie jak matematyka w PISA 2003 i nauki ścisłe w PISA 2006. W kolejnych latach wyniki są skalowane modelami IRT tak, aby możliwe było ich bezpośrednie porównywanie między cyklami, stąd jednak średnia i wariancja mogą się nieco zmieniać. W TIMSS oraz PIRLS wyniki są skalowane podobnie, jednak grupę krajów odniesienia stanowią uczestniczy pierwszego badania, a więc TIMSS 1995 lub PIRLS 2001. Stąd w TIMSS lub PIRLS bezpośrednie porównanie wartości średnich dla krajów jest dość trudne, bowiem zależy ona od tego, jakie inne kraje uczestniczyły w pierwszym badaniu. W PISA zawsze odnosimy się do średniej OECD, co ma sens o tyle, że jest to grupa najbogatszych państw świata.

[2] por. J. Micklewright, S. Schnepf, 2004. "Educational Achievement in English-Speaking Countries: Do Different Surveys Tell the Same Story?"

[3] por. M. Jakubowski, 2010, "Institutional tracking and achievement growth. Exploring difference-in-differences approach to PIRLS, TIMSS and PISA data", w: J. Dronkers (red.): "Assessing the Quality of Education and its Relationships with Inequality in European and Other Modern Societies", Springer.

W każdym z tych badań możliwe jest jednak porównanie % uczniów o danym poziomie umiejętności, co często ma także większy sens, bowiem mówi nam, ilu uczniów w każdym z krajów posiada dobrze opisany zestaw umiejętności.

Dla praktyków szkolnych i badaczy zainteresowanych mechanizmami wewnątrzszkolnymi, badania PIRLS oraz TIMSS są prawdopodobnie bardziej interesujące. Badanie te bowiem opierają się na próbie szkół oraz klas, a nie tak jak w przypadku PISA - na próbie szkół oraz losowej próbie uczniów wewnątrz szkół. Dane dla klas dostarczają zapewne więcej informacji w zakresie metod nauczania i zachowania uczniów. TIMSS i PIRLS zawierają też znaczną liczbę pytań dotyczących samego procesu nauczania. W PISA uczniowie wylosowani z danej szkoły należą najczęściej do różnych klas i mają do czynienia z innymi nauczycielami. Z tego względu obraz wewnątrzszkolnych zależności może tu być mniej wyraźny. Z drugiej strony badania TIMSS oraz PIRLS są bardzo ubogie pod względem informacji o pochodzeniu społeczno-ekonomicznym ucznia. Niektóre z edycji dla 4-klasistów nie zawierają nawet pytań o wykształcenie rodziców, przez co są właściwie nieprzydatne dla socjologów czy ekonomistów, dla których pochodzenie ucznia jest niemal zawsze kluczową zmienną w prowadzonych analizach. Z tego też względu trudne jest analizowanie i porównywanie krajów ze względu na to, jak ich systemy szkolne radzą sobie ze zmniejszaniem nierówności społecznych, czy też mówiąc inaczej, z wyrównywaniem szans edukacyjnych. Pod tym względem PISA ma zdecydowaną przewagę, gromadząc dokładne informacje nie tylko o wykształceniu i zawodzie rodziców ucznia, ale i o wyposażeniu gospodarstwa domowego w dobra kulturowe, edukacyjne oraz konsumpcyjne. Co więcej, bazy danych PISA oferują gotowe indeksy pochodzenia-społecznego ucznia i zasobności jego rodziny (np. standardowy w badaniach socjologów indeks ISEI czy też dostępny jedynie w PISA indeks ESCS - indeks ekonomicznego, społecznego i kulturowego statusu ucznia").

Badania PISA, PIRLS, TIMSS czy NAEP podają statystyki dla wyników uczniów zarówno na skali punktowej, jak i na skali porządkowej umożliwiającej jakościową interpretację wyników. Przykładowo, badanie NAEP wyróżnia 3 poziomy umiejętności i wiedzy uczniów: podstawowy, biegły, zaawansowany. PISA 2006 podaje wyniki w naukach przyrodniczych, stosując 6 opisowych poziomów wraz z kategorią uczniów, którzy nie osiągnęli nawet poziomu najniższego. Dzięki temu wyniki można odnieść do realnych umiejętności uczniów. I tak na przykład wynik na poziomie 2 w naukach przyrodniczych w badaniu PISA 2006 oznacza, że uczeń posiada wiedzę z nauk przyrodniczych umożliwiającą wyjaśnianie prostych zjawisk w znanym kontekście. Wynik na poziomie 5 oznacza już, że uczeń potrafi w sposób naukowy interpretować złożone zjawiska, stosować posiadaną wiedzę i naukowe metody poznania, oceniać wartość informacji z punktu widzenia nauk przyrodniczych, krytycznie analizować nieznane mu zjawiska i budować ich własną naukową interpretację.

Badania PISA, PIRLS oraz TIMSS są dość podobne pod względem metodologicznym. W obu stosowane są złożone schematy losowania prób, poszczególni uczniowie wypełniają jedynie część zadań z pełnej puli przygotowanej do badania, a wyniki skalowane są za pomocą populacyjnych modeli IRT i publikowane jako tzw. plausible values. W istocie, badania te w dużej mierze opierają się o wiedzę tych samych grup ekspertów, a także doświadczenia amerykańskiego badania NAEP, które od lat 60-tych ubiegłego wieku co dwa lata, a ostatnio corocznie, ocenia poziom umiejętności amerykańskich uczniów w kilku dziedzinach. Poniżej omawiamy metodologię badania PISA, podając jedynie krótkie informacje, czym różnią się metody stosowane w badaniach PIRLS oraz TIMSS.

**Metodologia badania PISA i innych badań międzynarodowych**

Podstawowym celem pomiaru w badaniach międzynarodowych nie jest określenie wyniku poszczególnego ucznia, lecz uzyskanie jak najbardziej precyzyjnej oceny wyników uczniów w danej populacji (najczęściej w danym kraju) lub też w subpopulacji (np. dziewczynki w Polsce). Można wykazać, że korzystanie z wyników stanowiących proste zsumowanie poprawnych odpowiedzi czy też rezultatów uzyskanych z prostych modeli IRT, niebiorących pod uwagę cech uczniów, daje błędne oceny średnich wyników w populacji, a tym bardziej błędne oceny zróżnicowania tych wyników (na przykład ich wariancji). Z tego wynika, że prostsze metody dają też błędne oszacowania procentu uczniów na określonym poziomie umiejętności (np. procentu uczniów na poziomie określonym jako „podstawowy"), różnic w średnich wynikach między grupami (na przykład chłopców i dziewczynek, uczniów ze wsi i z miasta) czy też zależności między wynikami a interesującymi nas zmiennymi (na przykład wynikami a pochodzeniem społecznym ucznia, czyli zmiennymi wykorzystywanymi w niemal każdym badaniu edukacyjnym). Z tego względu stosuje się korektę wyników, wykorzystując dla całej populacji model opierający się na bogatym zestawie cech ucznia i jego rodziny, wynikach w testach z różnych dziedzin, a także odpowiedziach na zadania dotyczące na przykład strategii uczenia się czy stosunku do nauki. Cechy te zbierane są poprzez dodatkowy kwestionariusz, który wypełniany jest po teście wiadomości i umiejętności, a większość z nich stanowi dobre predykatory wyników ucznia.

Ze względu na to, że wiedza, jaką badają PISA, PIRLS i TIMSS, jest niezwykle rozległa, a ograniczenia budżetowe są zawsze dość rygorystyczne, opracowano metody, w których uczniowie odpowiadają na różne zestawy zadań, tzw. *booklets* (dosłownie „książeczki" czy „broszury"). Każdy uczeń dostaje więc tylko część zadań przygotowanych z danej dziedziny, na przykład jedynie zadania z *booklets* numer 1 i 3. Następnie metodami IRT szacowane są parametry statystyczne zadań z każdej z *booklets*. Ponieważ testy są losowo przypisywane uczniom, to parametry te nie powinny się różnić między zestawami zadań. Jeżeli jednak różnice wystąpią, to są korygowane. Jest to możliwe, ponieważ uczniowie odpowiadają na te same pytania w różnych konfiguracjach *booklets*.

Dodatkowo sprawdza się, czy te same zadania mają podobne parametry w różnych krajach. Jeśli mimo długiego procesu przygotowywania zadań (pilotażowego badania ich właściwości w różnych krajach, wielokrotnego sprawdzania tłumaczeń, usuwania wątków odczytywanych różnie w zależności od kultury i historii danego kraju etc.) stało się tak, że jakieś zadanie okazuje się znacznie bardziej trudne lub łatwe w tylko jednym z krajów, to jest ono usuwane, zanim skalowane są wyniki uczniów tego kraju. Przykładowo, jeśli w procesie przygotowania zadań „przepuszczono" zadanie z czytania przywołujące postać świętego Mikołaja, to zadanie to zapewne zostanie usunięte w krajach arabskich, bowiem najprostsze statystyki pokażą, że nawet najlepiej czytający uczniowie w tych krajach nie są w stanie go rozwiązać, podczas gdy nie sprawia ono trudności uczniom polskim, którym zarówno święty Mikołaj, jak i zima są zapewne dużo bliższe. W końcu uzyskujemy więc w pełni porównywalne wyniki testów, które można poddać dalszej „obróbce". W praktyce proces przygotowania zadań jest tak drobiazgowy, wspierany także badaniami próbnymi w większości krajów, że w każdej edycji badania PISA usuwane jest na koniec dosłownie tylko kilka zadań. Pozostałe zadania okazują się posiadać niemal identyczne właściwości psychometryczne we wszystkich krajach.

Posiadając odpowiedzi uczniów na zadania testowe, nawet w pełni porównywalne międzynarodowo, stoimy wciąż przed dwoma problemami. Po pierwsze, każdy uczeń odpowiedział jedynie na część z całej baterii zadań. Po drugie, wynik każdego ucznia oddany jest na skali z zaokrągleniem do pełnej liczby, obarczony zapewne sporym błędem pomiaru. Oba problemy rozwiązuje model tzw. *plausible values* (dosłownie „wiarygodnych wartości"), gdzie uczniom przypisywane jest kilka równie prawdopodobnych wyników, które biorą pod uwagę nie tylko uzyskany przez ucznia „surowy" wynik z przedłożonych mu zadań z różnych dziedzin, ale i związki odpowiedzi na zadania testowe z cechami uczniów w całej populacji. Ten dość skomplikowany model statystyczny, przypisujący 5 wyników każdemu uczniowi, ma na celu przede wszystkim zwiększenie precyzji pomiaru dla całej populacji uczniów i jej podgrup, a także odtworzenie prawdziwych zależności między cechami ucznia, jego rodziny i szkoły a jego osiągnięciami.

Wykazano, że posługując się pięcioma *plausible values*, można odtworzyć rozkład prawdziwych umiejętności w całej populacji[4]. Tak więc nie tylko średnie wyniki w całej populacji, ale i wyniki dla podgrup, wariancja tych wyników oraz dalsze analizy odnoszące osiągnięcia uczniów do ich cech, programów edukacyjnych itp. ukazują wartości bliskie prawdziwym. Co więcej, model *plausible values* przewiduje wyniki dla uczniów, którzy rozwiązywali różne zadania testowe. Jest to możliwe dzięki uwzględnieniu cech uczniów oraz ich odpowiedzi na wszystkie pytania testowe, a także dzięki temu, że każdy *booklet* posiada zadania wspólne z innymi *booklets*, a model statystyczny równocześnie bierze pod uwagę wszystkich uczniów i wszystkie zadania testowe[5].

Wokół modeli *plausible values* narosło sporo nieporozumień. Niektórzy uważają, że w ten sposób zniekształca się „prawdziwe" wyniki. Zarówno matematyczne dowody, jak i symulacje pokazują jednak, że to posługując się *plausible values* można odtworzyć prawdziwe wyniki i zależności między innymi zmiennymi a wynikami w populacji. Co więcej, zmniejszając błąd pomiaru, zapobiega się błędnym wynikom uzyskiwanym np. metodami regresyjnymi. Zalety *plausible values* można wyjaśnić na prostym przykładzie. Załóżmy, że na 10 zadań testowych punktowanych jako 0 albo 1 uczeń poprawnie rozwiązał połowę z nich, uzyskując 5 punktów. Załóżmy, że testowi poddaliśmy tysiąc uczniów i 200 z nich uzyskało podobny wynik. Trudno oczekiwać, że na tak „krótkiej" skali wyników możliwe jest rozróżnienie rzeczywistych umiejętności uczniów. Z pewnością 200 uczniów, którzy uzyskali 5 punktów, są mocno zróżnicowani pod względem prawdziwego poziomu umiejętności. Mamy więc do czynienia ze sporym błędem pomiaru. Jeśli połowa tych uczniów to chłopcy, a druga połowa - dziewczynki, to porównanie między nimi może być mocno zniekształcone, jeśli różnią się rzeczywistym poziomem wiedzy, o ile w ogóle dokonanie rozróżnienia będzie możliwe przy tak krótkiej skali. Załóżmy jednak, że znamy pochodzenie społeczne uczniów, a także ich wynik w innym teście. Jak wiadomo, uwzględnienie

[4] Można tego dowodzić analitycznie, posługując się matematycznymi modelami, jak i poprzez symulację, gdzie zmienną ukrytą (określającą prawdziwy poziom wiedzy ucznia) odzwierciedlają obserwowalne odpowiedzi na zadania testowe, a związki między odpowiedziami i cechami ucznia mają z góry założone charakterystyki. W ten sposób badacz zna w pełni strukturę danych i sprawdza, czy model potrafi odtworzyć wartości zmiennej ukrytej (znanej nam w symulacji, ale nieuwzględnianej bezpośrednio w analizie statystycznej, tak jak to jest w rzeczywistości, gdzie zmiennej tej nie obserwujemy). Okazuje się, że model *plausible values* jest w stanie w pełni odtworzyć nie tylko średnie wartości zmiennej ukrytej, ale i cały jej rozkład, co nie jest możliwe w wypadku prostszych modeli IRT (por. Margaret Wu, 2005, "The role of plausible values in large-scale surveys", Studies In Educational Evaluation, Vol. 31, No. 2-3, str. 114-128; por. także przykłady w: OECD, 2009, "PISA Data Analysis Manual", dostępne za darmo na stronach www.pisa.oecd.org). Możliwe jest także wykorzystanie większej liczby plausible values, np. 10. W tym przypadku jednak korzyści ze zwiększonej precyzji są już znikome, a wymogi obliczeniowe są znacznie wyższe. Praktyczne różnice między ocenami uzyskanymi z modeli regresji opierającymi się na różnie skalowanych wynikach uczniów omawiano w: M. Jakubowski, „Impact of IRT scaling on the secondary analysis" (w publikacji; wyniki zaprezentowano na konferencji ESRA 2009).

[5] Dokładniej, model równocześnie szacuje wyniki w kilku dziedzinach (por. OECD, 2009, „PISA Technical Report", rozdział 9).

kilku pomiarów znacząco obniża błąd pomiaru. Wiedząc też jednak, że np. uczniowie rodziców z wyższym wykształceniem przeciętnie uzyskują wyższe wyniki niż uczniowie rodziców z wykształceniem podstawowym, warto wykorzystać tę informację, szacując prawdziwe umiejętności uczniów. W ten sposób uczniowie z rodzin o wyższym wykształceniu lub o wyższym wyniku drugiego testu, będą częściej otrzymywać wyższy wynik przypisany przez model *plausible values*. Nie ma tutaj jednak żadnego nieuzasadnionego podciągania lub zaniżania wyników. Po prostu wykorzystujemy informację o zależnościach w populacji między odpowiedziami na test a innymi zmiennymi celem oszacowania najbardziej prawdopodobnego wyniku. W ten sposób może okazać się, że dwie grupy uczniów, które uzyskały średnio po 5 punktów w rzeczywistości mogą różnić się osiągnięciami, nawet dość znacznie. Co prawda dla głównych domen badania międzynarodowe wykorzystują ok. 100 zadań do oceny umiejętności uczniów, to już jednak dla kolejnych domen, czy np. podskal umiejętności, liczba zadań może być znaczne mniejsza.

W praktyce, przy szacowaniu *plausible values* wykorzystuje się wszelką informację, jaką posiadamy o uczniu, a często i o szkole. W badaniu PISA brane pod uwagę są wszystkie odpowiedzi uzyskane w kwestionariuszu ucznia zawierającym kilkadziesiąt pytań zakodowanych jako ponad 250 zmiennych, a także zmienne identyfikujące każdą szkołę (tzw. „school dummies"), dzięki czemu wszelkie wspólne cechy uczniów każdej placówki są także uwzględnione przy szacowaniu wyników. Stosowane są dwa modele, jeden do równoczesnego szacowania wyniku głównego w 3 podstawowych domenach badania PISA: czytaniu, matematyce oraz naukach ścisłych, a drugi model do szacowania podskal ze szczegółowymi wynikami w tzw. głównej domenie, czyli np. w naukach ścisłych w PISA 2006. Model *plausible values* tak naprawdę przewiduje cały prawdziwy rozkład wyników dla każdego ucznia, a nie jedynie jedną lub kilka liczb oznaczających jego wynik. Następnie z tego rozkładu losowo brane jest 5 wyników, które publikowane są w bazie danych PISA i wykorzystywane do obliczania wszelkich statystyk w oficjalnych publikacjach OECD. Oczywiście, można by z już oszacowanych rozkładów wylosować większą liczbę wyników, np. 10 czy 50, jednak odpowiednie obliczenia jak i symulacje pokazują, że korzyść byłaby tu niewielka, a znacznie zwiększyłyby się wymagania obliczeniowe, które i tak są już spore.

Wymagania obliczeniowe są tak duże, ponieważ w badaniu PISA zastosowano także specjalną metodę liczenia błędów standardowych dla dowolnego rodzaju statystyk, tzw. metodę *BRR weights*. Dzięki tej metodzie możliwe jest szacowanie precyzji dowolnej statystyki (średniej w populacji, percentyla wyników dla danej podgrupy, czy też współczynnika równania regresji) uwzględniające złożony system losowania szkół oraz informacje dostępne jedynie losującym, a nieudostępniane w bazie ze względu na poufność informacji (dokładniej, ze względu na ryzyko zidentyfikowania szkoły). Metoda ta jest stosunkowo prosta, lecz wymagająca obliczeniowo. W bazie PISA dostarczane jest 80 wag, a obliczenie błędów standardowych polega na policzeniu statystyki dla głównej wagi i porównanie jej z podobnym wynikiem obliczonym dla każdej z 80 wag.

Tak więc dla jednej *plausible value* należy policzyć statystykę 81 razy, przy czym obliczenia trzeba powtórzyć dla każdej *plausible value*, a więc 5 razy, co daje 405 powtórzeń w każdym przypadku. Obliczenie zwykłych średnich zajmuje na dobrym komputerze kilka lub nawet kilkanaście godzin, a oszacowanie bardziej złożonych modeli, np. *quantile regression*, może zająć kilka tygodni nawet na dedykowanym temu serwerze obliczeniowym (czego doświadczył autor niniejszego artykułu). Wykorzystanie 80 wag BRR zapewnia odpowiednie określenie błędu związanego z wnioskowaniem o populacji na podstawie próby uczniów, podczas gdy powtórzenie obliczeń dla każdej *plausible value* poprawnie bierze pod uwagę błąd wprowadzany przez metodę ich szacowania. Nieskorzystanie z wag BRR lub innej metody odpowiedniej dla złożonego losowania z populacji, spowoduje zazwyczaj bardzo poważne zaniżenie błędów standardowych, a przez to wyciąganie zbyt pochopnych wniosków z wyników badania. Powtarzanie obliczeń dla każdej plausible value ma zazwyczaj mniejsze znaczenie, chyba że liczymy wyniki dla niewielkich lub odmiennych z uwagi na ważne cechy podgrup (np. dla imigrantów czy dla dolnego decyla uczniów). Można więc obliczyć wstępne wyniki z jedną z *plausible values,* jednak końcowe obliczenia najlepiej wykonać w pełni poprawnie, z pięcioma. Znacznym błędem jest wyciąganie średniej z *plausible values* i traktowanie tej liczby jako wyniku ucznia. Niestety, jest to wciąż popularna praktyka, lecz często prowadząca do błędnych wniosków.

Badania PIRLS oraz TIMSS wykorzystują bardzo podobne metody. Różnica polega na tym, że podstawowym modelem nie jest jednoparametryczny model IRT (model Rascha) stosowany w PISA, lecz trzyparametryczny model IRT biorący pod uwagę moc dyskryminacyjną oraz łatwość zgadywania dla każdego zadania. Błędy standardowe szacowane są natomiast metodą *jackknife* oddającą złożone losowanie szkół, a następnie klas. Poza tym także stosowany jest model *plausible values* podobny do wykorzystywanego w PISA.

**Wykorzystanie wyników badań międzynarodowych**

Oczywiście, głównym sposobem wykorzystania wyników badań międzynarodowych jest porównanie średniego poziomu umiejętności między krajami. Jak już wspomniano we wstępie, porównania te spotykają się z olbrzymim zainteresowaniem ze strony polityków, nauczycieli a nawet rodziców. Wyniki PISA trafiają na czołówki gazet na całym świecie, podobnie jak wyniki TIMSS oraz PIRLS. Wszyscy już wiemy, że Finowie mają najlepiej wykształconych 15-latków, a ich japońscy i koreańscy rówieśnicy potrafią niewiele mniej. Wiemy też, że np. USA, mimo olbrzymich nakładów na edukację, nie osiąga wyników zbliżonych do wymienionych powyżej liderów.

Wyniki PISA interpretowane są także w wymiarze zróżnicowania wyników uczniów. Cała metodologia badania, jak i towarzyszące kognitywnym testom kwestionariusze, nastawione są na pomiar zróżnicowania, a także zbieranie informacji i późniejszą prezentację wyników w odniesieniu do kontekstu społeczno-ekonomicznego. I tak wiemy, że Finowie osiągają nie tylko najwyższe przeciętne wyniki, ale i posiadają niezwykle niskie zróżnicowanie osiągnięć.

Japończycy posiadają system szkolny niezwykle silnie segregujący uczniów między szkołami ze względu na wyniki. Podobnie jest w krajach, gdzie utrzymywany jest system dzielący uczniów na różne typy szkół, np. w Niemczech. Duże zróżnicowanie wyników, a także niski wynik średni zaobserwowano także w Polsce w PISA 2000. PISA 2003 pokazała jednak, że w nowo utworzonych gimnazjach wyniki uzyskiwane przez 15-latków są nie tylko wyższe, ale i mniej zróżnicowane.

Te podstawowe wyniki, które znaleźć można w każdej edycji głównego raportu PISA wraz z bardziej szczegółowymi danymi i wstępnymi analizami, są już same w sobie niezwykle interesujące. Powinny być one jednak jedynie przyczynkiem do pogłębionych studiów, które są o tyle łatwe do zrealizowania, że zarówno dane z badań międzynarodowych, jak i przystępna dokumentacja, są dostępne za darmo, do ściągnięcia przez każdego w Internecie. OECD wraz z konsorcjum PISA publikują nie tylko bazę ze wszystkimi zebranymi danymi dla uczniów i szkół, ale i „PISA Technical Report" zawierający szczegółowy opis metodologii badania oraz „PISA Data Analysis Manual". Ta ostatnia publikacja zawiera przykłady analiz wykorzystujące specjalnie napisane „makra", które umożliwiają wykonanie podstawowych analiz w bardzo prosty sposób w pakietach SAS oraz SPSS. Nic więc nie stoi na przeszkodzie, żeby samemu rozpocząć analizy w oparciu o te zbiory danych. Podobnie zresztą jest w przypadku TIMSS oraz PIRLS, które co prawda nie udostępniają pakietu makr wraz z dokumentacją, ale publikowane za każdym razem podręczniki użytkownika do baz danych zawierają pomocne przykłady analiz.

Mogłoby się więc wydawać, że analizy w oparciu o zbiory PISA, TIMSS oraz PIRLS powinny być niezwykle popularne, jednak okazuje się, że nie jest to związane z łatwością dostępu do danych i przystępnością dokumentacji technicznej. Sporą liczbę badań przeprowadzono w Niemczech, gdzie PISA odbiła się dość głośnym echem i spowodowała wzrost zainteresowania wśród badaczy oraz grantodawców. Ciekawe badania przeprowadzono we Włoszech, gdzie w ostatnim cyklu PISA zwiększono kilkukrotnie próbę uczniów, aby opublikować wyniki w podziale na prowincje. Kraje, takie jak: Kanada, Australia, Szwajcaria, a częściowo także Czechy, wykorzystały badanie PISA do stworzenia własnego projektu badania panelowego (inaczej: podłużnego), czyli śledzącego losy 15-latków w latach późniejszych. Dostarcza to nowych wyników, niezwykle interesujących ze względu na możliwość spojrzenia na relację między osiągnięciami szkolnymi, późniejszymi decyzjami edukacyjnymi uczniów, a w końcu ich sukcesem na rynku pracy. Na wyniki dotyczące rynku pracy trzeba jeszcze poczekać, ale przygotowywane w tym roku raporty z Kanady i Szwajcarii pokazują niezwykle interesujące analizy ścieżek edukacyjnych 15-latków badanych w PISA. Nie należy też dziwić się brakiem szerszego wykorzystania danych PISA w krajach, takich jak: USA, Wielka Brytania, czy Holandia, posiadających znacznie dokładniejsze badania krajowe.

W Polsce dane PISA są niestety rzadko wykorzystywane. Być może złożoność analiz odstrasza potencjalnych chętnych, jednak powinno zachęcać bogactwo danych, jakie PISA oferuje. Oprócz wstępnych raportów opierających się na

analizach zespołu IFiS PAN realizującego badanie PISA w Polsce, nie powstało zbyt wiele prac badawczych. Należy mieć jednak nadzieję, że to się wkrótce zmieni. Na ostatniej konferencji poświęconej PISA (PISA Research Conference 2009, Kiel, Niemcy) badania poświęcone danym z Polski były łatwo zauważalne. Przedstawiono ciekawą analizę porównującą egzaminy zewnętrzne w Polsce z wynikami PISA (M. Grzęda, B. Ostrowska), szczegółową analizę wzrostu umiejętności uczniów w Polsce od 2000 roku (M. Jakubowski, H. Patrinos, E. Porta, J. Wiśniewski), spojrzenie na zmiany w osiągnięciach przez analizę odpowiedzi na pojedyncze pytania (E. Bartnik, M. Federowicz), a także dekompozycję wpływu zasobów ekonomicznych, kulturowych i społecznych rodziny ucznia na jego wyniki i przyrost umiejętności między gimnazjum a szkołą średnią (M. Jakubowski, A. Pokropek). Prezentacje tych badań, jak i innych przedstawionych na konferencji dostępne są w Internecie, pod adresem: http://www.pisaresconf09.org/index.php?id=2-18.

Mimo że liczba badań nad wynikami PISA jest w Polsce niewielka, to powyższe przykłady pokazują, że ulega to powoli zmianie. Trzeba też podkreślić, że PISA miała znaczny wpływ na inne badania przeprowadzone w Polsce. Badanie CKE realizowane przez firmę Pentor było częściowo wzorowane na PISA. Można też powiedzieć, że dyskusja nad systemem oświaty w świetle reprezentatywnych badań uczniów została zapoczątkowana przez PISA 2000. Miejmy nadzieję, że wkrótce pojawi się więcej prac badawczych wykorzystujących dane PISA, podobnie jak dane PIRLS, czy w przyszłości TIMSS.

**Badania międzynarodowe a krajowe systemy oceny umiejętności uczniów**

Można wskazać kilka sfer, gdzie krajowe systemy oceny umiejętności uczniów mogą czerpać z doświadczeń badań międzynarodowych:

1. tworzenie podstawy definiującej zakres umiejętności i wiedzy mierzony testami;
2. tworzenie zasobów z zadaniami, sprawdzanie ich właściwości psychometrycznych i konstruowanie końcowych testów;
3. skalowanie wyników testów;
4. sposoby publikacji i interpretacji wyników;
5. konstruowanie pochodnych miar określających jakość lub efektywność nauczania;
6. tworzenie zasobów do wykorzystania w pogłębionych analizach wyników uczniów.

Poniżej krótko omawiam każdy z tych punktów.

1) Tworzenie podstawy definiującej zakres umiejętności i wiedzy mierzony testami

Proces budowania *assessment framework* dla badań międzynarodowych jest przykładem, jak powinna być tworzona podstawa definiująca zakres ocenianych umiejętności i wiedzy uczniów. Podstawa ta powinna też pomóc w interpretacji wyników tak, aby wspomóc kształtowanie procesu nauczania. Umożliwia to szczegółowe zdefiniowanie umiejętności i wiedzy, jakie mają posiąść uczniowie

w poddziedzinach, które potem można wyodrębnić z ogólnego wyniku testu. W ten sposób test daje możliwość oceny wielowymiarowej, tak ważnej w procesie nauczania. Tworzenie podstawy dla badania PISA trwa za każdym razem kilka lat. Grupa międzynarodowych ekspertów, także z Polski, dyskutuje, jakie umiejętności i wiedza decydują o poziomie zaawansowania ucznia w danej dziedzinie, określa konstrukty stojące za tymi umiejętnościami i kierunki, w jakich powinny być rozwijane testy je mierzące. Jest to każdorazowo niezwykle ważny dokument, który tworzy podstawę dla konstrukcji testów, ich skalowania, a następnie raportowania.

Publikacje dokumentujące tworzenie *assessment framework* są ogólnie dostępne na stronach OECD. Warto do nich sięgnąć, zastanawiając się nad podstawami pomiaru w naszym kraju. Wiele krajów uczestniczących w badaniach międzynarodowych korzysta z tej dokumentacji i pracy ekspertów, tworząc własne systemy oceny uczniów bezpośrednio odnoszące się do pomiaru w badaniu międzynarodowym. Przykładem może być TIMSS, który jako bliski amerykańskiego programowi nauczania został wykorzystany w kilku stanach USA do porównania umiejętności uczniów w kontekście międzynarodowym. Ostatnio coraz częściej mówi się o podobnym wykorzystaniu PISA w USA, a niektóre inne kraje, w tym Polska, już stosują wiele z idei rozwiniętych przy tworzeniu *assessment framework* w trakcie rozwijania własnych systemów oceny uczniów. Z pewnością skorzystanie z wiedzy eksperckiej zgromadzonej w pracy nad PISA czy TIMSS, a przy tym możliwość skonfrontowania własnych poglądów z tym, co uważane jest za istotne w innych krajach, mogą w znacznym stopniu przyczynić się do udoskonalenia i poszerzenia krajowego systemu oceny uczniów.

2) Tworzenie zasobów z zadaniami, sprawdzanie ich właściwości psychometrycznych i konstruowanie końcowych testów

Skrupulatny proces tworzenia zbiorów potencjalnych zadań testowych i drobiazgowe sprawdzanie ich właściwości pomiarowych w badaniach międzynarodowych są świetnym przykładem, jak powinno się tworzyć testy, które nie tylko mierzą konstrukty zdefiniowane w podstawie pomiaru, ale i same w sobie mają wystarczająco dobre właściwości psychometryczne. Procesy te opisane są dość dokładnie w dokumentacji PISA i innych badań międzynarodowych. Można też wykorzystać ekspertów w nie zaangażowanych. Nawet powierzchowne spojrzenie na opisane działania pokazuje, jak wiele z tych rozwiązań można by zastosować w Polsce, szczególnie w procesie tworzenia egzaminów zewnętrznych. Tworzenie zadań w oparciu o szerokie grupy eksperckie starające się oddać to, co zostało zdefiniowanie w podstawie pomiaru, drobiazgowe badania pilotażowe połączone z analizami psychometrycznymi, konstruowanie testów z uwzględnieniem właściwości psychometrycznych poszczególnych pytań; wszystkie te etapy są świetnie opisane w dokumentacji badań międzynarodowych i słabo rozwinięte w polskim systemie egzaminacyjnym. Poprawiając te elementy w polskim systemie, warto by skorzystać z doświadczeń badań międzynarodowych.

3) Skalowanie wyników testów

Pomiar umiejętności i wiedzy uczniów ma zazwyczaj dwa różne cele. Bardzo często za pomocą testów chcemy ocenić umiejętności i wiedzę każdego ucznia. Dokładniej mówiąc, naszym celem jest prezentacja osobnego wyniku dla każdego ucznia, który jak najbardziej precyzyjnie oddaje jego prawdziwy poziom wiadomości i umiejętności. Wyniki na poziomie szkół czy całej populacji są tutaj sprawą drugorzędną. Nieco inne podejście powinno być stosowane, gdy naszym celem jest określenie przeciętnego poziomu umiejętności i wiedzy dla grup uczniów czy też całych populacji. W takim przypadku wynik indywidualny ucznia nie jest prezentowany, stanowiąc jedynie pośrednią informację pomagającą określić wynik dla szerszej grupy uczniów.

Oczywiście, system egzaminów zewnętrznych ma na celu przede wszystkim uzyskanie wyniku dla poszczególnego ucznia. Z tego względu skalowanie wyników powinno być nastawione na uzyskanie pojedynczego, możliwie najbardziej precyzyjnego wyniku dla każdego ucznia, być może w rozbiciu na poszczególne zakresy umiejętności i wiedzy. Można powiedzieć, że średnie wyniki dla szkół czy miary w rodzaju EWD, to już raczej pochodne systemu, być może ważne, ale nie najważniejsze. Dla miar tego typu można rozważyć zastosowanie osobnych modeli skalowania, o czym w punkcie (5).

Na ile więc metody stosowane w badaniach międzynarodowych, gdzie wynik pojedynczego ucznia nie ma znaczenia, mogą być przydatne dla krajowych systemów oceny uczniów? Przede wszystkim podstawowe metody są tu najczęściej podobne. Stosuje się proste modele IRT dla oceny własności psychometrycznych zadań, a następnie dla szacowania dla każdego ucznia wyniku, który stanowi podstawę dla modelu *plausible values.* Oczywiście, tworzenie *plausible values* dla wyników egzaminacyjnych nie ma większego sensu. Proszę sobie wyobrazić ucznia, który dowiaduje się, że ma 5 równie prawdopodobnych wyników z egzaminu z matematyki, przy czym jego koleżanka, która identycznie rozwiązała test, ma nieco inny wynik ze względu na różnice w pochodzeniu społecznym rodzin. Jednak skalowanie wyników prostymi modelami IRT ma olbrzymi sens, i to nie tylko na etapie końcowym, ale i na kolejnych etapach tworzenia testu.

Polskie wyniki egzaminacyjne nie są w żaden sposób skalowane. Choć prowadzono analizy metodami IRT już opublikowanych wyników, to uczniowie wciąż otrzymują prostą sumę poprawnych odpowiedzi, a szkołom przypisuje się ich średnią. Takie rozwiązanie ma niewątpliwie jedną zaletę: przejrzystość. Dla każdego jest jasne, że wynik egzaminu odpowiada sumie poprawnych odpowiedzi. Jest to jednak chyba jedyna zaleta, która często jest też poważną wadą. Przede wszystkim, wszelkie pomyłki w tworzeniu lub drukowaniu testu, czy też w trakcie egzaminowania, nie mogą być naprawione. Jeśli wyniki byłyby skalowane, to po pierwsze w procesie skalowania łatwo byłoby określić zadania, na które część uczniów nie odpowiedziała, choć powinna (np. ponieważ zostały błędnie wydrukowane w określonej liczbie testów). Co więcej, jeszcze przed udostępnieniem wyników można by zastanowić się, co z tymi zadaniami zrobić. Czy brać je pod uwagę, licząc wyniki egzaminu, czy też opuścić?

Czy jeśli bierzemy je pod uwagę, to tylko dla uczniów, gdzie test wydrukowany był prawidłowo, a dla pozostałych uwzględniamy tylko odpowiedzi na zadania wydrukowane poprawnie? Władze edukacyjne mogą tutaj podjąć decyzję, która jest znaczne łatwiejsza i mniej kontrowersyjna niż późniejsze decydowanie ad hoc co zrobić z uczniami, którzy otrzymali błędnie wydrukowany test. Oczywiście, po skalowaniu wynik prezentowany jest na standaryzowanej skali, więc wszyscy uczniowie mogą uzyskać tę samą maksymalną liczbę punktów, nawet jeśli dla części z nich kilka zadań nie zostało uwzględnione.

Inną zaletą skalowania, zapewne podstawową pod względem prawidłowości pomiaru, jest to, że odpowiedzi na zadania są różnie ważone, zależnie od ich trudności, a jeśli zastosujemy bardziej skomplikowane modele IRT, także od mocy dyskryminacyjnej czy też możliwości „strzelania" (kluczowe przy testach wielokrotnego wyboru). Co więcej, odpowiednie skalowanie umożliwia tworzenie wyników w podskalach, które rzeczywiście odpowiadają konstruktom opisanym w podstawie egzaminowania. Bez zastosowania modeli IRT, nasze klasyfikacje pytań na poszczególne grupy umiejętności mają charakter uznaniowy. Widać to przy analizach polskich wyników egzaminów zewnętrznych, gdzie zdarza się, że pytania z części humanistycznej są silniej powiązane z pytaniami mającymi mierzyć umiejętności matematyczne.

Wszystkie powyżej wspomniane modele stosowane są w badaniach międzynarodowych na wielu etapach, począwszy od sprawdzania zadań testowych, jak i przy skalowaniu końcowych wyników. Można czerpać z tych doświadczeń, jak i sporej literatury opisującej ich zastosowania w systemach krajowych. Trudno doprawdy znaleźć argumenty na rzecz niestosowania tych metod przy konstruowaniu i skalowaniu wyników polskich egzaminów zewnętrznych.

4) Sposoby publikacji i interpretacji wyników

Ta kwestia ściśle łączy się z poprzednimi, jest też kluczowa dla odbioru wyników egzaminacyjnych. Poświęca się jej często mniej uwagi, jednak jest to podejście błędne. To, jak przekazywane są wyniki, jest równie ważne jak to, jak tworzone i skalowane są testy. Jednak bez dobrze opisanej i przemyślanej podstawy egzaminowania, a także bez odpowiedniego skalowania wyników, publikowanie wyników testów w sposób przydatny i zrozumiały jest trudne. Żeby dobrze opisać, co oznacza 30 punktów egzaminacyjnych, trzeba by móc odnieść ten wynik do skali opisowej określającej, jaki zestaw umiejętności i wiedzy posiada uczeń o takim wyniku. Podobnie z wynikami w podskalach, np. umiejętności czytania ze zrozumieniem w teście humanistycznym. Muszą być one dobrze zdefiniowane, analizowane metodami IRT, aby wynik miał wiarygodną i poprawną interpretację. Punkty powinny być także publikowane w odpowiednio przygotowanej skali. Czy różnica między 15 a 17 jest taka sama w skali polskich egzaminów jak różnica między 18 a 20? Czy wynik równy 30 w 2005 roku oznacza tyle samo co 30 w 2008 roku? Oczywiście, w przypadku polskich egzaminów zarówno różnica, jak i średni wynik mogą odzwierciedlać zupełnie inne przedziały umiejętności. Trudno więc, o ile wyniki nie zostaną przełożone na jedną skalę o tej samej

jednostce pomiaru, średniej i wariancji, publikować rezultaty egzaminów w sposób jasny, zapobiegający nieuprawnionym, ale bardzo dziś popularnym porównaniom.

Badania międzynarodowe mogą stanowić wzór, w jaki sposób publikowane powinny być wyniki testów, przynajmniej jeśli chodzi o informacje dla grup uczniów. Wyniki prezentowane są zawsze w tej samej skali (500/100, porównywalnej między kolejnymi cyklami) i przekładane na kategorie opisowe (np. w PISA jest od 5 do 6 poziomów umiejętności i wiedzy). Warto skorzystać z tych wzorów, które stosowane są już na co dzień w wielu krajach o dłuższej tradycji testowania wiedzy uczniów (USA czy Wielka Brytania). Pod tym względem polski system egzaminów ma wiele do nadrobienia, jednak jak wspomniano na początku, bez odpowiednich podstaw trudno tutaj wiele zmienić.

5) Konstruowanie pochodnych miar określających jakość lub efektywność nauczania

System egzaminów zewnętrznych ma przede wszystkim na celu określenie wyniku dla każdego ucznia, jednak często jest też wykorzystywany do określenia średniego wyniku szkoły, gminy, powiatu czy też nawet porównywania regionów (wyników między rejonami poszczególnych OKE czy też województwami) i subpopulacji uczniów całego kraju (np. dziewczęta i chłopcy, uczniowie ze wsi a uczniowie z miast). Co więcej, wyniki egzaminacyjne wykorzystywane są do tworzenia bardziej złożonych miar mających np. określić jakość nauczania w danej szkole. Taką miarą jest – przykładowo – wskaźnik EWD (edukacyjnej wartości dodanej) gimnazjum, który uwzględnia poziom wiadomości i umiejętności ucznia pod koniec nauki w szkole podstawowej (wynik sprawdzianu), oceniając, ile zyskał on podczas nauki w gimnazjum.

Jak już wspomniano, oceny dla większych grup zyskują na precyzji, gdy pod uwagę brany jest model wyników dla całej populacji uczniów. Załóżmy, że uczeń uzyskał 30 punktów na sprawdzianie szóstoklasistów. Jak wiemy, wynik ten jest miarą prawdziwych umiejętności ucznia obarczoną błędem pomiaru, czyli niepewnością związaną z zestawem rozwiązywanych zadań, dyspozycją ucznia danego dnia etc. Co więcej, 30 punktów może być dość szeroką kategorią zawierającą uczniów o bardzo różnym poziomie wiadomości i umiejętności, których nie możemy rozróżnić ze względu na ograniczoną liczbę zadań w teście (np. 30 punktów na sprawdzianie z 2007 roku uzyskało ponad 20 tysięcy szóstoklasistów).

Ten sam uczeń poddany identycznemu testowi uzyskałby zapewne nieco inny wynik. Najczęściej bardzo bliski poprzedniemu, jednak z pewnym prawdopodobieństwem nieco inny. Jeśli moglibyśmy podawać wyniki z dokładnością do kilku cyfr po przecinku, to na pewno kolejne testy pokazywałyby inne wartości, nawet gdyby po zaokrągleniu uczeń otrzymywał zawsze 30 punktów. Co więcej, moglibyśmy też rozróżnić wyniki poszczególnych uczniów i tak na przykład, zamiast 20 tysięcy szóstoklasistów z tym samym wynikiem równym 30 punktów, mielibyśmy tysiąc szóstoklasistów z wynikiem równym 30,27 punktów. W zależności od tego, jak wykorzystujemy ten wynik, takie rozróżnienie, związane

z większą precyzją pomiaru, może mieć spore znaczenie lub też być zupełnie zbędne z praktycznego punktu widzenia. Dla indywidualnego wyniku sprawdzianu udostępnianego uczniowi i nauczycielom taka dokładność nie jest konieczna. Wiemy przecież, że błąd pomiaru wynosi kilka punktów, więc w porównaniu z nim ułamki są naprawdę sprawą pomijalną. Jeśli jednak wynik stanowi podstawę kolejnych obliczeń, np. średniej czy EWD szkoły, to już obliczenie wyniku ucznia z jak największą dokładnością zwiększy istotnie precyzję oszacowań tych wskaźników, dając dodatkowo znacznie lepszą podstawę do określenia zakresu, w jakim znajduje się prawdziwy wynik (przedziału ufności). Intuicyjnie, o ile korzyści płynące z dużej dokładności są niewielkie przy pojedynczym wyniku, to nabierają znaczenia, gdy korzystamy z kilkudziesięciu wyników obliczonych z większą precyzją.

Można więc zastosować modele populacyjne opracowane w badaniach międzynarodowych. Czerpać z ich doświadczeń i metodologii przy tworzeniu własnych miar. Jest to możliwe w oparciu o wyniki egzaminacyjne w Polsce, choć wymagałoby zbierania dodatkowych informacji o uczniach lub szkołach. Informacje te, w rodzaju poziomu wykształcenia i zawodu rodziców, liczby dzieci uzyskujących pomoc społeczną itp. powinny być zbierane, o ile zależy nam na odpowiedniej interpretacji wyników, przede wszystkim w odniesieniu ich do kontekstu szkoły. Tutaj badania w rodzaju PISA mają wiele do zaoferowania, bowiem stworzyły cały zestaw narzędzi pomiarowych służących analizowaniu wyników ucznia po uwzględnieniu jego pochodzenia społeczno-ekonomicznego. Podobne informacje zbierane są w systemach egzaminacyjnych wielu krajów, np. w Wielkiej Brytanii, i wykorzystywane nie tyle przy ocenianiu poszczególnych uczniów, co przy analizowaniu wyników szkół i grup uczniów. Dla prowadzenia efektywnej polityki edukacyjnej takie informacje mają charakter kluczowy.

6) Tworzenie zasobów do wykorzystania w pogłębionych analizach wyników uczniów

Na koniec warto podkreślić jeszcze raz, że wyniki badań międzynarodowych służą nie tylko tworzeniu końcowych raportów przedstawiających ich podstawowe rezultaty, ale i są szeroko wykorzystywane do pogłębionych analiz odnoszących osiągnięcia i postawy uczniów do cech ich rodzin, szkół oraz całych systemów szkolnych. Bazy danych ze wszystkich badań są dostępne za darmo w Internecie wraz z drobiazgową i dość przystępną dokumentacją. Podobnie wykorzystywane mogłyby być wyniki egzaminów zewnętrznych. Przykładem może być badanie zrealizowane na zlecenie CKE przez firmę Pentor, gdzie połączono wyniki egzaminów z danymi zebranymi w ankietach wzorowanych na badaniu PISA. Została opracowana dokumentacja dla tej bazy i miejmy nadzieję, że wkrótce będzie ona dostępna dla wszystkich chętnych. Podobnie można postępować z wynikami kolejnych fal egzaminów zewnętrznych. W ten sposób zyskujemy możliwość prowadzenia szczegółowych analiz, które mogą być pomocne w kształtowaniu polityki edukacyjnej, ale i poprawianiu systemu egzaminów zewnętrznych. Przykłady, jak tworzyć i udostępniać takie zbiory danych, dają badania międzynarodowe. Warto z nich skorzystać.