X Konferencja z cyklu Diagnostyka edukacyjna
20-22 IX 2004

Bolesław NIEMIERKO
Zakład Diagnostyki Edukacyjnej, Uniwersytet Gdański

ZAŁOŻONE I UZYSKANE, SPRAWDZAJĄCE I RÓŻNICUJĄCE ZNACZENIE TREŚCIOWE EGZAMINACYJNEJ SKALI POMIAROWEJ

Celem przedstawionych tu poglądów, rozważań i analiz jest szukanie drogi do pogłębienia interpretacji wyników egzaminu zewnętrznego. Jak zróżnicowana **forma zadań** wpływa na treść pomiaru? O jakich **osiągnięciach uczniów** mówią wyniki punktowe uczniów, szkół i lokalnych sieci szkół oraz różnice między tymi wynikami? Jak można podnieść **znaczenie treściowe** wyników egzaminu? Jakie rodzaje **skal pomiarowych** najlepiej komunikują to znaczenie? To są pytania, na które będę starał się odpowiedzieć, wykorzystując dorobek teorii pomiarowej i wybrane doświadczenia systemu egzaminacyjnego w Polsce, w tym zwłaszcza wyniki sprawdzianu osiągnięć uczniów z 2004 roku.

I. POMIAR I SKALA POMIAROWA

Gdy **pomiar** określimy szeroko jako uściślone poznawanie rzeczywistości, to **diagnostyka edukacyjna**, rozumiana jako poznawanie przebiegu, wyników i uwarunkowań uczenia się, jawi się ogromnym polem działalności pomiarowej. W tym duchu sformułowana jest preambuła statutu Polskiego Towarzystwa Diagnostyki Edukacyjnej, w której czytamy, że *główną metodą* [diagnostyki edukacyjnej] *jest pomiar dydaktyczny, wzbogacony warsztatem psychologii i socjologii* (PTDE, 2003). Wszyscy przecież wolimy informację ściślejszą od informacji mniej ścisłej i chętnie taką wolę deklarujemy.

Rzecz komplikuje się, jak zwykle, w zastosowaniach. Dopóki nie udowodnimy przydatności pomiaru dydaktycznego do rozpoznawania osiągnięć uczniów w najważniejszych dziedzinach kształcenia — i to jego przydatności dla „zwykłych" nauczycieli, nie zaś tylko dla badaczy i władz oświatowych — stawać będziemy pod pręgierzem zarzutu pozorowania działań diagnostycznych.[1] Powszechne są obawy o dehumanizację edukacji, zalewanej morzem liczb bez istotnego znaczenia. Diagnostyka jest kojarzona z arytmetyką, a jej zwolennicy są podejrzewani o brak elementarnej wiedzy o kształceniu młodzieży.

[1] Taki zarzut postawił nam prof. Krzysztof Konarzewski na trzeciej Krajowej Konferencji Diagnostyki Edukacyjnej. W konkluzji swojego referatu (1999) wyraził pogląd, że diagnostyka edukacyjna potrzebna jest tylko jej własnym teoretykom.

Sens pomiaru jest zawarty w **skali pomiarowej**, stanowiącej zbiór symboli przedstawiających jego wyniki. Znaczenie symbolom nadają procedury poznawania rzeczywistości. Planowanie, konstrukcja i stosowanie narzędzi pomiaru służą temu, by jego wyniki dostatecznie ściśle przedstawiały określone właściwości obiektów diagnozy. To są elementarne zasady, które warto jednak przypomnieć, by odróżnić od pomiaru to, co w ostatnich latach nazwano „mierzeniem jakości pracy szkoły", a co wyraża się jedynie zachętą do jej pogłębionej diagnozy.[2] Dopóki nie jesteśmy w stanie zbudować skali jakości pracy szkoły, ustalić znaczenia poszczególnych pozycji tej skali i udowodnić tego znaczenia, lepiej mówić o ewaluacji lub, jak kto woli, o ocenie i samoocenie szkoły[3] niż o „mierzeniu" jej pracy.

II. SKALE OSIĄGNIĘĆ UCZNIÓW

W pomiarze dydaktycznym stosuje się wiele skal, spośród których wybieramy sześć najbardziej charakterystycznych. Ich zalety i wady przedstawia tab. 1.

Tabela 1. Najważniejsze skale pomiarowe stosowane w diagnostyce edukacyjnej

Rodzaj pomiaru	Jednostka skali	Założenia	Główne zalety	Główne ograniczenia
Sprawdzający	Procent punktów	Policzalność osiągnięć	Prosta interpretacja intuicyjna	Silna zależność od testu, fikcyjne bieguny skali
Sprawdzający	Poziom wymagań	Hierarchia wymagań	Jakościowe definiowanie poziomów	Chwiejność norm, nietypowe układy wyników
Różnicujący	Równoważnik klasy lub wieku	Jednorodność osiągnięć	Naturalna jednostka, podzielna według czasu	Złudzenie sprawdzania, pracochłonność konstrukcji
Różnicujący	Ranga centylowa	Rozkład prostokątny	Mała zależność od testu, zrównywanie testów	Nierówne przedziały skali, mylona z procentem punktów
Różnicujący	Stanin	Rozkład normalny	Mała zależność od testu, równe przedziały skali	Abstrakcyjne założenie, zależność od próby
Różnicujący	Teta	Dopasowanie modelu	Niezależność od testu, niezależność od próby	Komplikacja matematyczna, potrzeba dużych prób

Procent punktów, jako stosunek punktowego wyniku surowego do długości skali (liczby punktów możliwych do zdobycia w danym teście), jest najprostszym, ulubionym przez niespecjalistów, mocno chwiejnym wskaźnikiem poziomu osiągnięć. Można

[2] „Mierzenie – to przede wszystkim postawa poznawcza, humanistyczne i holistyczne spojrzenie na człowieka, a nie jakieś konkretne procedury. W mierzeniu dobre są wszystkie te czynności i działania, które służą dobru ucznia, jego satysfakcji z podejmowanych przez niego działań edukacyjnych w szkole" – napisał Stefan Wlazło (1999, s. 24). Jakkolwiek szlachetna i słuszna jest taka postawa, nie może ona zastąpić warsztatu pomiarowego.

[3] „Niepotrzebnie używamy *ewaluacji*, skoro mamy *ocenianie*, chyba jako synonimu lub w nazwach złożonych..." – zdecydował Krzysztof Kruszewski we wstępie do książki: J. MacBeath i in. (2003).

na przykład uznać, że 80 procent punktów to „dużo", a 30 procent punktów to „mało", ale trudniej będzie się zgodzić na ocenę wyniku 55 procent. Najbardziej zawodzą bieguny takiej skali, bo 0 procent nie musi oznaczać, że uczeń „nic nie umie", a 100 procent, że „wszystko umie". W tej procedurze nie ma kontroli trudności zadań, a wręcz przeciwnie — przyjęto milczące założenie o równorzędności zadań, potrzebne do zliczania prawidłowych rozwiązań.[4]

Zwolennicy testów sprawdzających wielostopniowych, których — spodziewam się — jest wielu wśród uczestników konferencji diagnostyki edukacyjnej, dobrze znają uzasadnienie potrzeby definiowania **poziomów wymagań programowych**. Przeciwnicy — podnosić będą arbitralność takich wymagań, niemożność ich sprecyzowania i nietypowe układy osiągnięć, a rzadziej — utopijne przekonanie o uczeniu się wszystkich uczniów na jednym poziomie. Jest faktem, że niczego lepszego w dziedzinie pomiaru sprawdzającego dotąd nie wymyślono, a pomiar wielostopniowy ma wciąż perspektywy korzystnego oddziaływania na kształcenie (Niemierko, 2000).

Różnorodnością mieni się natomiast instrumentarium skalowania wyników pomiaru różnicującego. W Polsce mało znane są jeszcze **równoważniki klasy** (np. *5;8* — średni poziom osiągnięć po ośmiu miesiącach nauki w klasie szóstej szkoły podstawowej) lub **wieku** (np. *12;8* — średni poziom osiągnięć uczniów w wieku 12 lat i ośmiu miesięcy), a to zapewne z dwu powodów: (1) materializmu dydaktycznego, nakazującego myśleć o programie każdej klasy jako odrębnym zakresie wiadomości, nie zaś jako etapie rozwoju podstawowych umiejętności i (2) pracochłonności i kosztu budowania takich norm.

Ponieważ rośnie zainteresowanie **badaniami podłużnymi** osiągnięć uczniów i podłużnymi porównaniami wyników, należy ostrzec przed niebezpieczeństwem nadinterpretacji równoważników: oddziały szkolne wyprzedzające średnią lub nie sięgające średniej o całą klasę i uczniowie z różnicami trzech klas lub trzech lat wobec średniej nie mogą być przedmiotem wniosków o zmianę zaszeregowania. Sens przesunięć byłby nie większy niż grupowanie dzieci według wzrostu lub ciężaru ciała, gdyż równoważniki są jednostronnymi wskaźnikami ich rozwoju, dotyczącymi jedynie podstawowych umiejętności, np. rozumienia czytanego tekstu lub sprawności rozwiązywania prostych zadań matematycznych, nie zaś pełnej wiedzy. Nikt jeszcze w Polsce nie nadużył norm równoważnikowych, ale — w świetle wielu doświadczeń systemu egzaminacyjnego — warto dmuchać nawet na chłodne.[5]

Zarówno **normy centylowe**, jak i — oparte na nich — **normy standardowe** w znacznym stopniu uniezależniają nas od testu (doboru zadań) kosztem zwiększonego uzależnienia od próby standaryzacyjnej (doboru uczniów). Ta ich właściwość z trudem przemawia do wyobraźni pedagogów, toteż centyle, rozdzielające procentowe grupy uczniów są nagminnie mylone z procentem punktów, a staniny, ze względu na dużą jednostkę (po pół odchylenia standardowego), są chętnie „przekładane" na stopnie szkolne. Z dwu założeń o rozkładzie osiągnięć w populacji, krzywa Gaussa wydaje się

---

[4] *Liczby wykazują wielokrotność jednakowych okoliczności* — przypominał Peter Krope (1994) w referacie wygłoszonym podczas pierwszej konferencji diagnostyki edukacyjnej —zaś *to, co określamy jako „jednakowe" i „niejednakowe" jest pochodną celu, do którego potrzebujemy liczenia.*

[5] Gdyby różnice trudności krajowych sprawdzianów absolwentów szkoły podstawowej w Polsce przedstawiać w kategoriach równoważników, to Sprawdzian 2003 byłby trudniejszy od Sprawdzianu 2003 o 10 tygodni nauki, a Sprawdzian 2004 byłby trudniejszy od Sprawdzianu 2003 aż o cały rok nauki. Autorzy tych narzędzi mogliby być więc pochopnie posądzeni o manipulację wymaganiami.

lepiej przybliżać rzeczywistość szkolną niż pole prostokąta, ale gdyby nawet można było udowodnić rozkład normalny wyników uczenia się, to i tak nie wiedzielibyśmy niczego o spełnieniu oczekiwań społecznych co do osiągnięć szkolnych. Wymagania przychodzą bowiem z zewnątrz rozkładu wyników i, jak dowodzą stosowne analizy (Hambleton, 1980), muszą być arbitralne.

Najbardziej zaawansowana matematycznie ze skal używanych w psychometrii jest **skala teta**, oparta na teorii wyniku zadania matematycznego. Ta teoria pozwala na oszacowanie prawdopodobieństwa rozwiązania zadania o określonej trudności przez ucznia o danym poziomie osiągnięć. Wymaga dużych prób standaryzacyjnych, ale jej wartości są — w zasadzie — niezależne od poziomu osiągnięć uczniów w próbie, a oszacowanie osiągnięć określonego ucznia lub grupy uczniów jest — w zasadzie — niezależne od trudności zadań (Hambleton i Swaminathan, 1985). Dla niespecjalisty jednak kontrola dopasowania modeli probabilistycznych do rozkładu wyników pomiaru jest zbyt trudna, a symbolika skali rozciągającej się od –3,00 do +3,00 jest niewygodna, bo sugeruje cofanie się osiągnięć u co drugiego ucznia.

III. SKALA POMIAROWA A TRAFNOŚĆ OCENIANIA

Dlaczego uczniowie, rodzice i nauczyciele skłonni są ignorować modele różnicowe i traktować każdy pomiar jako **sprawdzający**? Dlatego, że nawiązują do najprostszej dydaktyki, do zasady: „nauczać i sprawdzać, czy umie". Różnice osiągnięć między uczniami są dla nich drugoplanowe, a nawet niepożądane, mącące obraz. Czy pomiar dydaktyczny ma odwracać tę hierarchię ważności?

Dotychczas wydawało się „pomiarowcom" dydaktycznym, że nie potrafią obejść się bez skal różnicowych, trafności kryterialnej, współczynników rzetelności, wskaźników mocy różnicującej i zadań zamkniętych, ale rysują się już teorie lepiej dostosowane do potrzeb nauczycieli i do ogólnospołecznego rozumienia szkoły.[6] Podnoszą one znaczenie **kontekstu uczenia się**, stanowiącego „środowisko dydaktyczne" oceniania, procesu uczenia się i sprzężenia zwrotnego w tym procesie, motywacji uczenia się i komunikacji z uczniami (Brookhart, 2003). **Trafność oceniania**, będąca „po Messicku" (1989) dziedziną największych zmian pojęciowych, sięga po „interpretatywne nauki społeczne", nie wyłączając hermeneutyki i konstruktywizmu, dla wyjaśnienia sensu uzyskiwanej informacji (Moss, 2003), a **rzetelność oceniania** ma być rozumiana jako „wystarczalność informacji" (*sufficiency of information*) potrzebnej do podjęcia decyzji dydaktycznej (Smith, 2004). Wszystko to po to, by pomiar służył nauczycielom pomocą w codziennej sytuacji szkolnej.[7]

Kluczowe dla pomiaru dydaktycznego jest pytanie o jego trafność. Komplikacja treści kształcenia (tekstu) i uwarunkowań (kontekstu), których dotyczy, powoduje, że

[6] We wstępie do monograficznego numeru „Educational Measurement: Issues i Practice", zatytułowanego „Zmienianie sposobu, w jaki teoretycy pomiaru myślą o ocenianiu wewnątrzszkolnym" zostały one, dla odróżnienia od psychometrii, prowizorycznie nazwane „klasometrią" (Benson, 2003).

[7] O tej sytuacji tak napisano w czasopiśmie cytowanym w poprzednim odsyłaczu: *Realia klasy szkolnej to elementy środowiska klasy, którym nauczyciel musi stawić czoła. Te realia - takie jak utrzymanie poziomu pracy, promocje ze względów społecznych, niezadowalająca frekwencja, wykolejone i niesprzyjające środowisko domowe, złe postawy uczniów i niewłaściwe zachowanie, a także różnorodność uzdolnień uczniów – tworzą wyzwania, które często są przeciwne temu, w co nauczyciel wierzy i co chciałby czynić w ramach oceniania.* (McMillan, 2003, s. 37).

sprowadzenie wyników uczenia się do jednej skali przez zdefiniowanie odpowiedniej **zmiennej ukrytej** (hierarchii wymagań, sekwencji wyników) zawsze będzie budzić wątpliwości. Osiągnięcia uczniów są wielowymiarowe, a planowanie testów odbywa się zwykle w dwu lub trzech wymiarach. Jak potem je sprowadzić do jednego wymiaru?

Posłużmy się analogią z wydatkami kilkuosobowej rodziny. Gdy jej dochody wynoszą (A) tylko kilkaset złotych miesięcznie, to różnice (w górę i w dół) przekładają się głównie na treść posiłków; gdy wynoszą (B) kilka tysięcy złotych, mogą już wpłynąć głównie na sposób ubierania się; gdy wynoszą (C) kilkadziesiąt tysięcy złotych, wyrażą się zapewne wielkością rezydencji. Jak widać, skala (logarytmiczna) zamożności może mieć różne znaczenie w kolejnych odcinkach.

Rodziny mogą wydawać pieniądze na różne cele, niekoniecznie według potrzeb, które inni uważają za racjonalne, np. (A) na napoje alkoholowe, (B) na drogie samochody i (C) na spekulacje giełdowe. Nie chodzi nam więc zatem o to, by mieć pewność co do rodzaju wydatków, lecz by przynajmniej orientować się w ich jakości, w tym — na co je można przeznaczyć na danym poziomie.

Matematykom znane jest twierdzenie o niemożności dokonania **pomiaru łącznego** wartości kilku lub więcej zmiennych o niepełnej interkorelacji wartości (Nowakowska, 1975, s. 223–236). Wiązka takich zmiennych to wydatki rodzinne, a także osiągnięcia szkolne ucznia. To, co możemy określić jako trafność zmiennej ukrytej w polu zmienności złożonych zjawisk społecznych, to jak najwyższe prawdopodobieństwo wystąpienia typowych układów wartości.

IV. ZNACZENIE TREŚCIOWE SKALI

Nawet najszerzej rozbudowane pojęcie trafności teoretycznej pomiaru dydaktycznego ma u podstaw **trafność wewnętrzną**, polegającą na porównaniu treści pomiaru z programem kształcenia lub z inną dokumentacją wymagań wobec ucznia (Niemierko, 1999, s. 174n). O tym porównaniu pisze Samuel Messick (1989, s. 39):

Głównym zagadnieniem planowania testu, a także jego oceny ze względu na określone zastosowania, jest reprezentatywność, z jaką test obejmuje dziedzinę treści (lub zachowań lub procesów). Jednakże, by uzyskać reprezentatywność, trzeba nie tylko zaplanować granice dziedziny, lecz także logiczne lub psychologiczne podziały lub aspekty danej dziedziny zachowań lub właściwości. Potem konstruktor może systematycznie obejmować te pola według określonych reguł, takich jak ujęcie równe czyli jednolite, według częstości naturalnego występowania lub w proporcji do uznawanego znaczenia praktycznego. Faktyczna liczba zadań reprezentujących każdy aspekt lub pole planu może być proporcjonalna do częstości lub znaczenia, albo też, dla uzyskania tego samego efektu, mogą być zastosowane wagi (jeśli tylko pola planu, które mają być ważone, są mierzone zadowalająco rzetelnie).

Planowanie testu osiągnięć szkolnych nadaje jego skali **znaczenie treściowe** (*content meaning*), dzięki któremu o uczniu uzyskującym pewien wynik można zasadnie wnioskować, jakie czynności opanował, a jakich nie opanował (Niemierko, 1999, s. 247–248; Niemierko, 2002, rozdz. 15). Ta informacja jest potrzebna wszystkim „udziałowcom" edukacji, zaczynając od uczniów, a kończąc na politykach.

V. ZAŁOŻONE ZNACZENIE SKALI

To, co rozpatrywaliśmy dotychczas jako znaczenie treściowe skali wyników pomiaru i co Messick uznał za podstawę trafności teoretycznej pomiaru, oparte było na planowaniu testu i na procedurach stosowania planu. Nazwiemy to **założonym znaczeniem treściowym** (*attempted content meaning*) skali.

Możliwie jasno sformułowane założenia są oczywiście niezbędne, by zbudować test i przedstawić jego wyniki.[8] Te założenia są zawsze w pewnym stopniu arbitralne, istotnie ograniczające treść pomiaru i dyskusyjne z różnych punktów widzenia. Na przykład standardy wymagań egzaminacyjnych, jakimi się teraz posługujemy w szkole podstawowej (*Rozporządzenie...*, 2001, wraz z późniejszymi zmianami), obejmują umiejętności komunikacyjno-praktyczne ucznia (porozumiewanie się, rozwiązywanie problemów, przetwarzanie informacji, zastosowania praktyczne), a nie obejmują umiejętności poznawczo-społecznych (motywacja i organizacja uczenia się, współdziałanie w grupie, umiejętności negocjacyjne), także wymieniane w podstawach programowych. Poszerzenie standardów, dydaktycznie bardzo pożądane, pociągałoby za sobą zwiększenie liczby zadań otwartych, a także zespołową organizację części sprawdzianu (Poręba-Konopczyńska, 2001), co musimy odłożyć na przyszłe lata.

Decyzją ekspertów ustanowiono pięć obszarów umiejętności absolwentów szkoły podstawowej: czytanie, pisanie, rozumowanie, korzystanie z informacji i wykorzystywanie umiejętności w praktyce. Tym umiejętnościom nadano wagi, które szybko utrwaliły się w praktyce konstruktorskiej, podobnie jak dość wygodna w kalkulacjach długość skali wyników sprawdzianu (liczba punktów możliwa do zdobycia) wynosząca 40 punktów.

Założone znaczenie treściowe skali wyników sprawdzianu w ostatnich trzech latach przedstawia się, w uproszczeniu, jak w tab. 2.

Tabela 2. Założone znaczenie treściowe wyników sprawdzianu

Umiejętność	Waga	Liczba punktów
Czytanie	25%	10
Pisanie	30%	12
Rozumowanie	20%	8
Korzystanie z informacji	5%	2
Wykorzystywanie wiedzy w praktyce	20%	8
Razem	100%	40

Tab. 2 pokazuje zamianę słów na liczby, jakości (umiejętności) w ilość (punktów do zdobycia). Ta zamiana jest konieczna, gdy w masowym egzaminie chcemy — korzystając z dobrodziejstw systemu liczbowego — zobiektywizować i usprawnić sprawdzanie osiągnięć uczniów. Czy jednak nasze plany dadzą się dokładnie zrealizować? Domyślamy się, że nie.

[8] Tak to ujmuje Peter Krope w cytowanym już artykule (1994): *Ktokolwiek posługuje się liczbami bez wcześniejszego posłużenia się słowami, nie może wyjaśnić, o czym mówi. Najpierw słowa, potem liczby!* Jest to, jego zdaniem, jedna z fundamentalnych „zasad posługiwania się słowami i liczbami".

VI. UZYSKANE ZNACZENIE SKALI

Uzyskanym znaczeniem treściowym (*achieved content meaning*) skali nazwiemy jej znaczenie treściowe oparte na analizie wyników zadań rozwiązywanych przez uczniów.

Założone znaczenie treściowe skali niewiele wykraczało poza pojęcie **trafności fasadowej**, ograniczonej do przekonania autorów i użytkowników skali o tym, że przedstawia ona to, co uznali za ważne. Gdy dysponujemy wynikami, możemy sięgnąć do nich dla upewnienia się, że to właśnie mierzymy. Według Messicka — będziemy wnioskować o trafności pomiaru „na podstawie dowodów", a nie tylko na podstawie słów i wyobrażeń. Wzbogacimy ustalanie trafności pomiaru o istotne procedury **trafności teoretycznej**.

Można też uznać uzyskane znaczenie skali za niezbędne uzupełnienie jej założonego znaczenia. Sprawdzając znaczenie liczb zamieniamy je z powrotem na słowa, by porównać realizację z planem. To porównanie jest odmienne i do innych prowadzi wniosków w sprawdzaniu i w różnicowaniu osiągnięć.

VII. UZYSKANE ZNACZENIE SPRAWDZAJĄCE SKALI

Najprostsze pytanie o sens wyniku egzaminu, jakie można postawić po jego przeprowadzeniu, brzmi: co złożyło się na ten wynik? Aby na nie odpowiedzieć, musimy go przedstawić (a) w przypadku pojedynczego ucznia — jako sumę punktów uzyskanych przez tego ucznia za rozwiązanie poszczególnych zadań testu lub (b) w przypadku grupy uczniów — jako sumę wskaźników łatwości wszystkich zadań testu ważonych długością skal tych zadań. Przedstawia to wzór (1): gdzie x jest wynikiem testowania, aj jest wagą zadania j, pj jest łatwością zadania j.

$$x = \sum_{j=1}^{m} a_j p_j \qquad (1)$$

Gdy uczeń lub grupa uczniów uzyskuje maksymalny wynik testu, jego znaczenie treściowe odpowiada założeniom, bo proporcje punktów zgadzają się dokładnie z wagami przewidzianymi w planie testu (tab. 2). Gdy jednak wynik jest niższy od maksymalnego, proporcje treści zmieniają się, bo trudniejsze zadania szybciej tracą swój udział w sumie punktów. **Uzyskane znaczenie sprawdzające** skali to znaczenie określone zadaniami rozwiązanymi przez ucznia lub grupę uczniów.

Uzyskane znaczenie sprawdzające wyników egzaminu zewnętrznego po szkole podstawowej przeprowadzonego w 2004 roku przedstawia tab. 3.

Tabela 3. Uzyskane znaczenie sprawdzające wyników sprawdzianu

Umiejętność	Rodzaj zadań	Średnia łatwość	Długość skali	Wynik punktowy	Udział
Czytanie	zamknięte	,82	10	8,19	32%
Pisanie	otwarte	,69	12	8,25	32%
Rozumowanie	zamknięte	,48	4	3,68	14%
	otwarte	,44	4		
Korzystanie z informacji	zamknięte	,74	1	1,49	6%
	otwarte	,75	1		
Wykorzystanie wiedzy w praktyce	zamknięte	,62	5	3,95	16%
	otwarte	,29	3		
Razem	zamknięte	,70	20	14,0	55%
	otwarte	,58	20	11,6	45%
	łącznie	,64	40	25,6	100%

Porównując kolumnę „udział" w tab. 3 z kolumną „waga" w tab. 2, zauważamy zmianę proporcji znaczenia na korzyść „czytania" (przyrost 7 punktów procentowych) i na niekorzyść elementów przedmiotów ścisłych: „rozumowania" (ubytek 6 punktów procentowych) i „wykorzystania wiedzy w praktyce" (ubytek 4 punktów procentowych). Pewne znaczenie ma tu także rodzaj zadań: zadania wyboru wielokrotnego, wprowadzone jako wyłączne w „czytaniu", dostarczyły ogółem o 10 punktów procentowych większą część wyniku ogólnego niż — przy tej samej długości skali — zadania otwarte. Silną pozycję sprawdzającą ma jednak także „pisanie", którego jakość jest szacowana wyłącznie na podstawie zadań otwartych.

Taki wynik może cieszyć polonistów, których przedmiot ma tu mocniejszą pozycję niż w założeniach. Zadania matematyczne, jako na ogół trudniejsze, dostarczyły uczniom mniej punktów i mniej wpłynęły na ich dorobek punktowy niż zadania językowe. Ta prawidłowość jest tym silniejsza, im bardziej oddalamy się od wyniku maksymalnego ku wynikom niskim. To naturalne, że dla słabych uczniów każdy test pisemny jest przede wszystkim testem czytania.

Czy to znaczy, że uczniowie i szkoły różnią się między sobą najbardziej ze względu na umiejętności językowe i to właśnie te, które są mierzone zadaniami zamkniętymi? By odpowiedzieć na to pytanie musimy sięgnąć do teorii **pomiaru różnicującego** i zastosować procedury oparte na rozrzucie wyników wokół średnich.

VIII. UZYSKANE ZNACZENIE RÓŻNICUJĄCE SKALI

Różnice między wynikami testowania uczniów i szkół powstają z dwu źródeł: (1) ze zmienności wyników zadań, wyrażanej **wariancją zadań**, i (2) z interkorelacji wyników zadań, wyrażanych ich **kowariancją**. Upraszczając, możemy powiedzieć, że na te różnice składa się niejednakowy wynik poszczególnych uczniów w zadaniach testu oraz zgodność między ich wynikami w różnych zadaniach. Gdy sukcesy w jednych zadaniach często odpowiadają sukcesom w innych zadaniach, a braki w jednych zadaniach często odpowiadają brakom w innych zadaniach, to kowariancja wyników zadań, a co za tym idzie — wariancja wyników testowania, jest duża.

Miarą zgodności wyników zadania *j* z wynikami wszystkich zadań testu (nie wyłączając z niego zadania *j*) jest korelacyjny współczynnik mocy różnicującej zadania *j*. Ten współczynnik pomnożony przez odchylenie standardowe wyników zadania *j* przedstawia wkład zadania do odchylenia standardowego wyników testowania, nazywany **wskaźnikiem rzetelności** (*reliability index*) zadania. Suma wskaźników rzetelności dla testu złożonego z *m* zadań stanowi odchylenie standardowe wyników testowania. Przedstawia to wzór (2):[9] gdzie s_t jest odchyleniem standardowym wyników testowania, s_j jest odchyleniem standardowym wyników zadania *j*, r_{jt} jest korelacją wyników zadania *j* z wynikami testowania (współczynnikiem mocy różnicującej zadania *j*).

$$s_t = \sum_{j=1}^{m} s_j r_{jt} \qquad (2)$$

Analogia między wzorami (1) i (2) polega na sumowaniu wkładu zadań (lub grup zadań, np. testek): (1) w postaci ważonych łatwości zadań i (2) w postaci wskaźników rzetelności, to jest iloczynów odchylenia standardowego i właściwego (dla zadania lub testki) współczynnika mocy różnicującej. **Uzyskane znaczenie różnicujące** skali jest określone wkładem zadań do rozrzutu wyników pomiaru.

Analiza znaczenia treściowego różnicującego krajowych wyników *Sprawdzianu 2004*, dokonana z zastosowaniem wzoru (2), jest przedstawiona w tab. 4.

Z pierwszej kolumny liczb dowiadujemy się, że średnia wariancja zadań zamkniętych i otwartych jest dość wyrównana, a jedynie wyniki zadań wyboru wielokrotnego mierzących umiejętność czytania, znacznie łatwiejszych od pozostałych zadań, mają wariancję niższą. Z drugiej kolumny liczb odczytujemy, że zadania otwarte lepiej różnicują egzaminowanych uczniów niż zadania zamknięte, co nie jest oczywiście niespodzianką, bo samodzielność i czas pracy ucznia są na ogół większe w zadaniach otwartych, pozbawionych gotowych odpowiedzi. Kolejne trzy kolumny zawierają wartości dwu czynników oraz odpowiednich iloczynów, będących składnikami sumowania we wzorze (2). Procentowy udział testek mierzących pięć umiejętności objętych egzaminem oraz udział zadań zamkniętych i otwartych w różnicowaniu wyników sprawdzianu przedstawia prawa skrajna kolumna tab. 4.

Porównując udział testek w sprawdzaniu (tab. 3) i różnicowaniu (tab. 4) osiągnięć uczniów, zauważamy wzrost znaczenia „rozumowania" i „wykorzystywania wiedzy w praktyce", a więc elementów przedmiotów ścisłych, średnio o ponad połowę oraz dramatyczny spadek, niemal do połowy, znaczenia „czytania". Teraz górą wydają się być matematycy i przyrodnicy, których zadania lepiej mierzą umiejętności uczniów w sensie różnicowania osiągnięć. Tylko „pisanie" utrzymuje wysoką pozycję przy obydwu podejściach, sprawdzającym i różnicującym.

[9] Systematyczne wyprowadzenie wzoru podaje Magnusson (1981, s. 315-317). W mojej książce *Testy osiągnięć szkolnych. Podstawowe pojęcia i techniki obliczeniowe* (1975, s. 261-262) zamieściłem to uzasadnienie w skrócie. Bardziej złożone rozwiązanie, oparte na krzywoliniowych zależnościach między wynikami zadań (testek) przedstawił M.J. Cresswell (1987), wybiega ono jednak poza nasze obecne potrzeby.

Tabela 4. Uzyskane znaczenie różnicujące wyników sprawdzianu

Umiejętność	Rodzaj zadań	Średnia wariancja zadania *)	Średnia moc różnicująca zadania *)	Odchyl. standard. testki	Moc różnic. testki	Wskaźnik rzetelności testki	Udział
Czytanie	zamknięte	,12	,38	1,79	,78	1,39	18%
Pisanie	otwarte	,20	,45	2,98	,81	2,40	31%
Rozumowanie	zamknięte	,23	,44	2,16	,84	1,81	23%
	otwarte	,19	,54				
Korzystanie z informacji	zamknięte	,19	,48	0,69	,61	0,42	5%
	otwarte	,18	,48				
Wykorzystanie wiedzy w praktyce	zamknięte	,23	,40	2,19	,82	1,78	23%
	otwarte	,20	,61				
Razem	zamknięte	,17	,40	3,77	,90	3,41	44%
	otwarte	,20	,49	4,69	,94	4,42	56%
	łącznie	,18	,45	7,83	1,00	7,83	100%

*) W zadaniach otwartych każdą wyodrębnioną, osobno punktowaną „czynność" (aspekt rozwiązania) potraktowano tu jako osobne zadanie.

Porównując udział testek w sprawdzaniu (tab. 3) i różnicowaniu (tab. 4) osiągnięć uczniów, zauważamy wzrost znaczenia „rozumowania" i „wykorzystywania wiedzy w praktyce", a więc elementów przedmiotów ścisłych, średnio o ponad połowę oraz dramatyczny spadek, niemal do połowy, znaczenia „czytania". Teraz górą wydają się być matematycy i przyrodnicy, których zadania lepiej mierzą umiejętności uczniów w sensie różnicowania osiągnięć. Tylko „pisanie" utrzymuje wysoką pozycję przy obydwu podejściach, sprawdzającym i różnicującym.

Ogółem, zadania otwarte wniosły wyraźnie więcej (o 12 punktów procentowych) do różnicowania niż zadania zamknięte, przy takiej samej długości skal (20 punktów). Zwolennicy zadań otwartych zyskują więc potwierdzenie swoich, dobrze skądinąd uzasadnionych teoretycznie, przekonań.

IX. TYPOWE I NIETYPOWE OSIĄGNIĘCIA UCZNIÓW

Dane analizowane w tab. 3 i 4 dotyczyły średnich wyników bardzo dużej liczby uczniów (517 081), stanowiących krajową populację szóstoklasistów (z wyłączeniem uczniów o specjalnych potrzebach edukacyjnych). Znaczenie treściowe wyników poszczególnych uczniów może dość mocno różnić się od podanego w tych tabelach.

Nietypowe układy (*unusual patterns*) osiągnięć uczniów wyrażają się wysokimi wynikami zadań, które są trudne i niskimi wynikami zadań, które są łatwe w populacji uczniów. Najprostszym sposobem ich wykrycia jest obliczenie „osobistego współczynnika korelacji" wyników ucznia w poszczególnych zadaniach testu ze średnimi wyni-

kami (wskaźnikami łatwości — w przypadku punktowania 0–1) tych zadań. Gdy współczynnik jest niski, układ można nazwać nietypowym.[10]

Jest wiele przyczyn nietypowości układu wyników testowania uczniów. Obok swoistości systemu dydaktycznego nauczyciela i szkoły, powodującej, że egzamin zewnętrzny „nie pasuje" do grupy uczniów, pojedynczy uczeń może być jednostronnie uzdolniony, wąsko zainteresowany, niedouczony, roztargniony, niesamodzielny („ściągający" niektóre odpowiedzi), zbyt powolny, słabo umotywowany.[11] Należy jednak przestrzec przed zbyt łatwym sugerowaniem się zaobserwowanymi osobliwościami rozkładu wyników, gdyż ogromna większość z nich nie wykracza poza losowe wahania wyników nie tylko pojedynczych zadań, lecz także krótkich testek. To oczywiście nie wyklucza przypadków nietrafnej analizy osiągnięć ucznia „chodzącego własnymi drogami".

Typowe układy osiągnięć uczniów wyrażają się zgodnością wyników zadań z hierarchią łatwości zadań w grupie odniesienia (populacji). Można je znaleźć w raportach Centralnej Komisji Egzaminacyjnej, okręgowych komisji egzaminacyjnych oraz — pogłębione — w niektórych opracowaniach szczegółowych pracowników tych komisji.[12] W kolejnych latach z wolna przechodzić będziemy zapewne od zestawień typowych układów osiągnięć uczniów do układów nietypowych, badając różne przypadki indywidualnego i grupowego odstępstwa od statystycznych prawidłowości pomiaru. Ujawniając i wyjaśniając obszary nietypowości wyników, pogłębimy egzaminy zewnętrzne w ich funkcji diagnostycznej.

X. PERSPEKTYWY EGZAMINU ADAPTACYJNEGO

Nie ulega wątpliwości, że rozwój systemów egzaminacyjnych postąpi w kierunku indywidualizacji pomiaru osiągnięć poszczególnych uczniów. Będzie to możliwe dzięki stosowaniu **skomputeryzowanego testowania adaptacyjnego, STA** (*computerized adaptive testing, CAT*), to jest takiego wspomagania komputerowego pomiaru, by wybór następnego zadania (wiązki zadań, testki) lub decyzja o zakończeniu testowania zależały od oszacowania osiągnięć ucznia, opartego na jego wynikach w poprzednich zadaniach (Bunderson i in., 1989, s. 381). **Sekwencja zadań** rozwiązywanych przez poszczególnych uczniów różni się w STA treścią i długością, dzięki czemu pracują oni według własnych, indywidualnych potrzeb.

Matematyczna teoria STA rozwija się intensywnie od wczesnych lat siedemdziesiątych ubiegłego stulecia (Lord, 1970), obecnie głównie w ramach teorii wyniku zadania (Hambleton, Swaminathan, 1985, rozdz. 12; Reckase, 1989). Warunkiem jej stosowania jest nagromadzenie dużej liczby wysokiej jakości zadań o jednoznacznie sklasy-

[10] Zjawisko nietypowości układów osiągnięć poddał analizie Japończyk T. Sato (1971). Amerykanie, którzy „odkryli" jego teksty (pisane w języku ojczystym) rozwinęli zawarty w nich model „wskaźnika ostrzegawczego". Garść informacji na ten temat zamieściłem w książce „Pomiar sprawdzający w dydaktyce" (1990, s. 391-396).

[11] Pełen wykaz możliwych przyczyn rozchwiania struktury testu i schematy diagnozy stylu pracy uczniów można znaleźć na stronach wskazanych w poprzednim odsyłaczu.

[12] Raporty CKE powstają pod kierunkiem Adama Brożka (szkoły podstawowe) i Teresy Chrostowskiej (gimnazja). Rozwiniętą analizę znaczenia treściowego typowych wyników sprawdzianu przedstawia w niniejszym tomie Maria Krystyna Szmigel (2004).

fikowanej treści i o znanych parametrach łatwości i mocy różnicującej. Centralna Komisja Egzaminacyjna wydaje się być już bliska takiego **banku zadań.**

Zalety komputerowego egzaminu adaptacyjnego są liczne:

1. umożliwia wzbogacenie treści zadań testowych o dynamiczną symulację sytuacji praktycznych,
2. nadaje się szczególnie do nietypowych układów osiągnięć,
3. indywidualizuje poziom, tempo i czas trwania egzaminu,
4. skraca czas egzaminu średnio o połowę (Bunderson i in., 1989, s. 387),
5. zapobiega „ściąganiu" rozwiązań z sąsiedniego stanowiska,
6. na bieżąco informuje ucznia o jego wyniku,
7. usprawnia centralną rejestrację i analizę osiągnięć uczniów.

Problemem do rozwiązania w skomputeryzowanym testowaniu adaptacyjnym jest znaczenie treściowe krótkich sekwencji zadań (Niemierko, 1990, s. 362-363). Chodzi o ich **trafność wewnętrzną**, to jest o harmonijną reprezentację całego pola programowego egzaminu oraz o formę zadań naturalną dla danego rodzaju czynności ucznia.[13] Z tego powodu nie zadania, lecz **testki** (*testlets*), sprawdzające określone umiejętności na określonym poziomie, mogą być pozycjami indywidualnych sekwencji testowania (Wainer i Kiely, 1987; Niemierko, 1990, s. 285-289). Wydłuża to, rzecz jasna, czas testowania, lecz stwarza możliwość zróżnicowanej diagnozy osiągnięć uczniów w wybranych obszarach umiejętności.

Entuzjaści STA przewidują dwa kolejne etapy jego rozwoju: „pomiar ciągły" i „pomiar inteligentny". **Pomiar ciągły** (*continuous measurement*), będący obecnie przedmiotem coraz szerszych eksperymentów, polega na *zastosowaniu skalibrowanych miar wbudowanych w proces kształcenia, by systematycznie i bez zakłóceń jego przebiegu szacować dynamiczne zmiany w postępach uczniów* (Bunderson i in., 1989, s. 187). Pozwala to na płynną aktualizację **mapy osiągnięć** (*mastery map) ucznia*, przypominającą dobrze nam znaną analityczną ocenę opisową, która bywa jednak, jak dotychczas, w przewadze intuicyjna.

Pomiar inteligentny (*intelligent measurement*) ma polegać na wykorzystaniu „wiedzy pedagogicznej" (sztucznej inteligencji) komputera, który będzie nie tylko diagnozował osiągnięcia, lecz także budował indywidualny program kształcenia (tok uczenia się) na tej podstawie. Te prace są dopiero w zarodku. Dla nas ważny jest ich kierunek: rozszerzanie i ułatwianie diagnozy uczenia się przy postępującej integracji z procesem dydaktycznym.

XI. ZNACZENIE TREŚCIOWE ŚREDNICH WYNIKÓW SZKÓŁ

W miarę ulepszeń i krzepnięcia systemu egzaminów szkolnych w Polsce interpretacja wyniku szkoły traci powoli akcent ilościowy („Które miejsce w gminie, powiecie, województwie ma dana szkoła?") na rzecz znaczenia treściowego wyniku („Co umieją absolwenci tej szkoły?"). Powoduje to potrzebę zastosowania pojęć wprowadzonych w tym referacie.

[13] *Testy adaptacyjne stwarzają trudność utrzymania planu testu, który wymaga wyważonej reprezentacji różnych zakresów treści* – przyznali Victor Bunderson i współautorzy rozprawy *Cztery generacje skomputeryzowanego pomiaru dydaktycznego* (1989, s. 386).

Na **uzyskane znaczenie sprawdzające średniego wyniku szkoły** składają się punkty uzyskane przez wszystkich jej uczniów poddanych egzaminowaniu. Udział poszczególnych umiejętności w tej średniej może odbiegać od struktury krajowej, przedstawionych w tab. 3, i wskazywać na sukcesy i na zaniedbania w niektórych obszarach wymagań egzaminacyjnych. Ich źródła mogą stać się przedmiotem analiz dla dyrekcji i dla rady pedagogicznej danej szkoły.

Uzyskane znaczenie różnicujące średnich wyników szkół, interesujące głównie administrację i nadzór pedagogiczny edukacji, może być oszacowane przez zastosowanie wzoru (2) w zbiorze średnich wyników szkół, krajowym lub lokalnym.[14] Gdy chodzi o sprawdzian osiągnięć uczniów po szkole podstawowej, to zapewne wynik oszacowania nie odbiegnie daleko od tab. 4, ale to, co różni uczniów w szkole (w populacji) nie musi być tożsame ze źródłami różnic między szkołami.[15]

Porównywanie osiągnięć jednych szkół z innymi, silnie obwarowane znajomością kontekstu ich pracy, może być pożyteczne raczej w makroskali, gdy prowadzi do wykrycia użytecznych prawidłowości, niż w mikroskali (dwie lub niewiele więcej szkół), grożącej pochopnością sądów. Celem porównań powinno być zawsze udzielanie pomocy słabszym, nigdy zaś „wytykanie i zawstydzanie", jak to kiedyś praktykowano w „ogarniętej obsesją kontroli Anglii" (Potworowski, 2000).[16]

XII. POSTĘP W OSIĄGNIĘCIACH UCZNIA I SZKOŁY

W cieniu zaawansowanej teorii pomiaru różnicującego *(norm-referenced measurement)* oraz — młodszej i słabszej, ale dla pedagogów cenniejszej — teorii pomiaru sprawdzającego *(criterion-referenced measurement)*, kiełkuje teoria **pomiaru postępu** *(self-referenced measurement)*. Postęp w osiągnięciach uczniów jest zmianą, ustalaną przez porównanie wyników pomiaru w dwu wybranych momentach procesu uczenia się. Wynik wcześniejszy staje się punktem (układem) odniesienia dla wyniku pomiaru późniejszego.

Pomiar postępu był długo lekceważony w psychometrii, jako nieoryginalny (pochodny od dwu odrębnych pomiarów), obciążony zwiększonym błędem standardowym (dwa źródła wahań losowych) i trudny w interpretacji (Niemierko, 1990, s. 354). Zauważmy jednak, że w pokoju nauczycielskim, a tym bardziej — w typowej rodzinie ucznia, mówienie o jego „postępach w nauce" było zawsze naturalne. Obecnie, przy silnej tendencji do integracji pomiaru z nauczaniem, zainteresowanie teoretyków zmianą w osiągnięciach uczniów stale rośnie.

Gdy zaczęto traktować szkołę jako „organizację samoucząca się" (Kruszewski, 2003, s. 5), pomiar postępu szkoły stał się ważnym zadaniem. Na razie uzupełnia jedynie ocenę (ewaluację) jej stanu, ale wiele wskazuje na to, że dojrzeje do roli głównego

[14] W tym przypadku *st* będzie odchyleniem standardowym średnich wyników szkół, *sj* – odchyleniem standardowym średnich wyników zadania *j* w szkołach, *rjt* – korelacją średnich wyników zadania *j* ze średnimi wynikami szkół (współczynnikiem mocy różnicującej zadania *j* między szkołami).

[15] Istotność tych rozbieżności może być sprawdzona za pomocą analizy wariancji.

[16] Jedną z oznak takiej polityki może być nazewnictwo poziomów osiągnięć. Obok już wcześniej zalecanej skali osiągnięć uczniów: „nowicjusz" (*novice*), „praktykant" (*apprentice*), „biegły" (*proficient*), „mistrz" (*master*) i „ekspert" (*expert*), można by wprowadzić następującą skalę osiągnięć szkół: „szczególnej troski" (*special care*), „na dorobku" (*progressing*), „wiodąca" (*leading*) i „na medal" (*distinguished*). Ostatnia nazwa jest zaproponowana przez Marię Krystynę Szmigel (2004).

wskaźnika jakości pracy szkoły.[17] Jak nauczycielska ocena osiągnięć ucznia może być zdominowana zasadą „Najważniejsze, by się uczył", tak kuratoryjna ocena osiągnięć szkoły może mieć hasło „Najważniejsze, by się rozwijała".

XIII. SKALE POSTĘPU W OSIĄGNIĘCIACH SZKOLNYCH

Gdy interesuje nas postęp, a nie stan chwilowy, to inaczej patrzymy na instrumentarium skalowania osiągnięć szkolnych. Odmiennie niż w tab. 1, zbudowanej dla pojedynczych obserwacji, jednostką staje się różnica miedzy dwiema kolejnymi obserwacjami osiągnięć ucznia lub szkoły. Przy tych samych założeniach o pojedynczej obserwacji skale pomiarowe mogą stracić lub zyskać na takim podejściu.

Jak poprzednio, scharakteryzujemy sprawdzanie i różnicowanie. Główne skale tych dwu rodzajów przedstawia tab. 5.

Tabela. 5. Skale pomiarowe stosowane do pomiaru przyrostu osiągnięć

Rodzaj pomiaru	Jednostka skali	Założenia	Główne zalety	Główne ograniczenia
Sprawdzający	Różnica procentu punktów	Policzalność osiągnięć	Prosta interpretacja intuicyjna	Zależność od testu, silny efekt pułapu
Sprawdzający	Różnica poziomu wymagań	Hierarchia wymagań	Rozwój osiągnięć jako podnoszenie jakości	Bardzo duża jednostka, trudno definiowalna
Różnicujący	Różnica równoważników	Jednorodność osiągnięć	Bardzo prosta interpretacja	Przecenianie różnic, efekt wachlarzowy
Różnicujący	Różnica rang centylowych	Rozkład prostokątny	Dokładny wskaźnik zmiany pozycji w grupie	Przecenianie różnic w środkowej części skali
Różnicujący	Różnica stanin	Rozkład normalny	Właściwa miara postępu względem populacji	Za duża jednostka dla różnic grupowych
Różnicujący	Różnica wartości teta	Dopasowanie modelu	Niezależność od testu, wysoka precyzja	Złożoność matematyczna, potrzeba banku zadań

Różnica procentu punktów („W zeszłym roku nasi uczniowie uzyskali 85 procent punktów z czytania, a w tym roku — tylko 80 procent..."), najłatwiejsza do zarejestrowania, jest bardzo silnie zależna od trudności kolejnych testów, a w najwyższych przedziałach skali jest często — jako miara różnicy osiągnięć — pomniejszona ze względu na **efekt pułapu**, czyli niemożność przekroczenia 100 procent poprawności. Z kolei w dolnych częściach skali znaczenie treściowe różnicy jest notorycznie niskie, bo większość przewidzianych czynności pozostaje nieopanowana.

[17] W cytowanej już książce *Czy nasza szkoła jest dobra?* (MacBeath i in., 2003) w arkuszu „Profil samooceny szkoły" (s. 24, tab. 1.1) każdy z 12 wymiarów jakości pracy szkoły jest zaopatrzony w dwie skale: stanu i postępu („zmiana na korzyść" — „brak zmiany" — „zmiana na niekorzyść"). „Osiągnięcia w nauce" są na czele tych wymiarów, lecz ponieważ są oceniane jedynie intuicyjnie, bez zastosowania pomiaru dydaktycznego, ich postęp budzi największe kontrowersje (s. 60, tab. 4.1).

Różnica poziomu wymagań (np. awans z poziomu „praktykant" na poziom „biegły") jest dla pojedynczego ucznia dystansem ogromnym, w dodatku niepodzielnym na etapy, bo wymagania tworzą całość o złożonej strukturze. Ich rozstęp może mieć korzystne oddziaływanie motywacyjne (Brophy, 2002), ale powoduje, że wysoko trafny pomiar sprawdzający postępy uczniów w dostatecznie szerokim zakresie programowym musiałby często przynosić rozczarowujące **wyniki zerowe**, oznaczające brak zmiany. Inaczej jest z postępem grupy (klasy, szkoły), gdzie można porównywać liczby uczniów spełniających wymagania poszczególnych stopni. Tu standardy edukacyjne, gdy są czytelne, dość dokładnie opisują postęp.

Równoważniki (klasy lub wieku) wydają się znakomitą miarą postępu w osiągnięciach (o tydzień, o miesiąc, o rok nauki szkolnej lub życia), ale ostrzeżenia pod ich adresem, sformułowane na początku referatu, muszą być utrzymane, a nawet wzmocnione. Obok przeceniania wielkości różnic pojawia się tu **efekt wachlarzowy** (*fan effect*), polegający na wzroście rozrzutu osiągnięć wraz z ich wzrostem w grupie. Na skutek efektu wachlarzowego około połowy uczniów zostaje z roku na rok coraz bardziej w tyle za ogólną średnią i choć druga połowa grupy coraz bardziej wyprzedza średnią, odczucia niespecjalistów z tym związane są negatywne, bo powszechnie oczekuje się, że szkoła będzie wyrównywać szanse dzieci[18].

Centyle („Uczeń przesunął się w hierarchii osiągnięć populacji o pięć procent jej liczebności") nie nadają się na wskaźniki postępu w osiągnięciach szkolnych, bo na skutek nierealistycznego założenia prostokątnego rozkładu osiągnięć — silnie przeceniają różnice w pobliżu mediany i nie doceniają różnic na krańcach skali. Tej wady nie mają skale standardowe, w tym — **skala staninowa**, jeżeli tylko rozkład wyników testowania jest, choć w przybliżeniu, regularny (symetryczny, wypiętrzony). Z powodu zbyt dużej (pół odchylenia standardowego) jednostki skali staninowej, różnice między wynikami szkół powinny być wyrażane w dziesiątych częściach stanina („Szkoła dokonała postępu z kategorii niżej średniej, 4,3 stanina, do średniej, 4,6 stanina") do czego ta skala, jako przedziałowa, uprawnia (wyjąwszy dwa staniny skrajne, najniższy i najwyższy, stanowiące przedziały półotwarte).

Najmniej wiemy o zastosowaniach **skali teta** do pomiaru różnic w osiągnięciach uczniów. Jej główna zaleta, jaką jest oszacowanie osiągnięć niezależne od trudności testu, rokuje wysoką użyteczność dla instytucji posiadających wykwalifikowany personel i duże banki zadań. Z drugiej strony jednak pamiętać należy, że żadna ze skal różnicowych nie nada, sama przez się, postępom uczniów znaczenia treściowego. To znaczenie ma tylko trafnie zaobserwowana różnica poziomu spełnionych wymagań. W innych przypadkach potrzebne są dodatkowe analizy, do których teraz przechodzimy.

[18] Efekt wachlarzowy norm podłużnych ilustruje tabela zamieszczona w książce *Testy osiągnięć szkolnych* (Niemierko, 1975, s. 156-157). W ciągu 13 miesięcy nauki najlepsi uczniowie uzyskali w teście rozumienia czytanego tekstu postęp równy 22 miesiącom, a najsłabsi – tylko 3 miesiącom.

XIV. WARTOŚĆ DODANA WYNIKU UCZNIA I SZKOŁY

Z Anglii, która nie popisała się wobec świata publikowaniem oficjalnych „tabel ligowych" szkół na podstawie egzaminów zewnętrznych (Potworowski, 2000, s. 55), pochodzi cenne pedagogicznie pojęcie „wartości dodanej". **Wartość dodana** (*value added*) **wyniku ucznia** jest przyrostem osiągnięć w wybranym zakresie programowym w określonym czasie. Taki przyrost ma aspekt ilościowy („Jaki postęp uzyskali uczniowie przez trzy lata nauki w naszym gimnazjum w stosunku do średniej krajowej?") i jakościowy („Co uczniowie danej klasy umieją lepiej po roku?").

Szkoły angielskie na ogół dobrze przyjęły ten sposób interpretacji faktów, dający szansę wykazania się nie tylko placówkom skupiającym najzdolniejszą młodzież (Gipps, 1993, s. 44n). Czas na ostrożne próby zastosowania go w Polsce.

W określaniu wartości dodanej wyniku ucznia i grupy uczniów (szkoły) jesteśmy ograniczeni do tego, co mierzą egzaminy zewnętrzne. Dopóki nie istnieją standardy podłużne i normy podłużne naszych testów, mamy do wyboru dwie strategie:

Zastosowanie „na wyjściu" (np. na zakończenie gimnazjum) jakiejś wersji egzaminu przeprowadzonego „na wejściu" (sprawdzianu dla absolwentów szkoły podstawowej) dla ustalenia, jak dana szkoła pomogła uczniom o określonych brakach z niższego szczebla edukacji.

Zastosowanie „na wejściu" (np. na początku gimnazjum) jakiejś wersji egzaminu, który będzie prowadzony „na wyjściu" (egzaminu dla absolwentów gimnazjum) dla ustalenia, jaka część osiągnięć uczniów była zadatkiem wyniesionym z niższego szczebla edukacji, a jaka część jest dorobkiem szkoły[19].

Te dwie strategie uzupełniają się. Stosowanie wyłącznie pierwszej strategii mogłoby doprowadzić do przesadnej koncentracji na elementarnych umiejętnościach uczniów, a zatem do obniżenia poziomu pracy szkoły, jak to było w amerykańskich programach „minimum kompetencji" (Niemierko, 1990, s. 94). Stosowanie wyłącznie drugiej strategii mogłoby onieśmielać i zniechęcać słabszych uczniów. Dwie strategie łącznie dają pole do popisu każdego poziomu uczniom i szkołom.

Po stwierdzeniu, dość oczywistej, arytmetycznej różnicy wyników dwu pomiarów, przechodzimy do analizy wartości dodanej. Jej wielkość może być porównana z różnicą uzyskaną przez innych uczniów i przez inne szkoły. Jej znaczenie treściowe może być ustalone w procedurze opisanej w tab. 3, gdzie „średnią łatwość" zastępujemy „średnią różnic łatwości", a „wynik punktowy" — „różnicą punktową".

Do pomyślenia jest także przebadanie na te sposoby populacji uczniów i szkół określonego szczebla, oszacowanie korelacji różnic w poszczególnych umiejętnościach z różnicami w całym mierzonym zakresie (mocy różnicującej wartości dodanej tych umiejętności) i przedstawienie obrazu znaczenia różnicującego wartości dodanej tych umiejętności w postępie uzyskanym przez ucznia lub szkołę. To byłoby już zadanie badawcze dla instytucji naukowej.

[19] Te przykłady dotyczą wyłącznie gimnazjum, gdyż tylko ta szkoła ma obecnie egzaminy zewnętrzne „na wejściu" i „na wyjściu". Wprowadzenie „nowej matury" i, ewentualnie, egzaminu przegłądowego w klasie III szkoły podstawowej, na wzór „trzeciotescika" stosowanego przez Instytut Badań Kompetencji w Wałbrzychu (Mulawa i Sroka, 2003), rozszerzyłoby zakres diagnozy wartości dodanej wyników pomiaru osiągnięć uczniów i szkół. Możliwe jest także wykorzystanie badań dojrzałości szkolnej (poziom „zerówki" – wejście do szkoły podstawowej).

XV. PODSUMOWANIE

Najważniejsze ustalenia i propozycje, jakie zostały zawarte w tym referacie, są następujące:

Skala pomiarowa syntetyzuje i komunikuje sens pomiaru. Jej dobór ma zasadnicze znaczenie dla wartości i użyteczności wyników pomiaru.

Największą wartość dla diagnostyki edukacyjnej ma **skala wymagań programowych**. Pomocniczą rolę odgrywają skale różnicowe: równoważnikowa, staninowa i — potencjalnie — skala teta.

Na skutek złożoności i wielowymiarowości osiągnięć szkolnych przedziały skali pomiarowej wykazują nie tylko różnice ilościowe, lecz także **różnice jakościowe** w osiągnięciach uczniów. Interpretacje oparte na jednorodnej „zmiennej ukrytej" mają zatem ograniczoną wartość.

Głównym postulatem trafności teoretycznej pomiaru osiągnięć uczniów jest właściwe **znaczenie treściowe** wyników. W uproszczeniu, jest to odpowiedź na pytanie „Co uczniowie umieją?"

Plan testu przedstawia **założone znaczenie treściowe** skali wyników pomiaru. Jest zbiorem życzeń, których spełnienie powinno być udowodnione.

Uzyskane znaczenie treściowe jest oparte na analizie wyników zadań rozwiązywanych przez uczniów. Może różnić się od założonego.

Uzyskane znaczenie sprawdzające skali pomiarowej jest określone zadaniami prawidłowo rozwiązanymi przez ucznia lub grupę uczniów. Zadania łatwiejsze mają w nim większy udział od zadań trudniejszych.

Uzyskane znaczenie treściowe różnicujące skali pomiarowej jest określone wkładem zadań do rozrzutu wyników pomiaru. Zadania o większej wariancji i wyższej mocy różnicującej mają większy wkład różnicujący.

Nietypowe układy osiągnięć uczniów i szkół wyrażają się wysokimi wynikami zadań trudnych i niskimi wynikami zadań łatwych. Analiza przyczyn nietypowości osiągnięć należy do problematyki diagnostyki edukacyjnej.

Przyszłość egzaminowania zewnętrznego stanowi **skomputeryzowane testowania adaptacyjne**. Przewiduje się, że umożliwi ono integrację egzaminowania z kształceniem szkolnym.

Także średnie wyniki szkół mogą być interpretowane w kategoriach uzyskanego znaczenia **sprawdzającego** i **różnicującego**. To pierwsze jest ważne dla pojedynczej szkoły, a to drugie — dla sieci szkół.

Postęp w osiągnięciach uczniów, oszacowany przez porównanie wyników pomiaru w dwu wybranych momentach procesu uczenia się, coraz bardziej interesuje teoretyków i praktyków pomiaru dydaktycznego.

Spośród skal pomiarowych najbardziej przydatne do mierzenia postępu w osiągnięciach uczniów są **skale wymagań programowych** i **skale standardowe**. Najlepiej, gdy te dwa rodzaje skal są stosowane równolegle, jako podwójne normowanie osiągniętego postępu.

Wartość dodana wyniku ucznia jest przyrostem osiągnięć w wybranym zakresie programowym w określonym czasie. Jest kłopotliwa w ustalaniu, ale uważana za najbardziej kontekstowo niezależną miarę osiągnięć ucznia i szkoły.

Tyle pomiaru dydaktycznego, ile treściowej interpretacji jego wyników!

LITERATURA

Benson J., (2003), *Editorial*, „Educational Measurement: Issues and Practice. Special Issue *Changing the way measurement theorists think about classroom assessment*, nr 4.

Brookhart S. M., (2003), *Developing measurement theory for classroom assessment purposes and uses*, „Educational Measurement: Issues and Practice. Special Issue *Changing the way measurement theorists think about classroom assessment*, nr 4.

Brophy J., (2002), *Motywowanie uczniów do nauki*, PWN, Warszawa.

Cresswell M.J., (1987), *A more generally useful measure of the weight of examination components*, „British Journal of Mathematical and Statistical Psychology", 40, s. 61-79.

Gipps C., Stobart G., (1993), *Assessment. A teachers' guide to the issues*, Hodder, London.

Hambleton R.K., (1980), *Tests core validity and standard-setting method* (w:) R.A. Berk (red.) *Criterion-referenced measurement: The state of the art.*, The John Hopkins Univ. Press, Baltimore.

Hambleton R.K., Swaminathan H., (1985), *Item Response Theory: Principles and Applications*, Kluwer-Nijhoff, Norwell.

Konarzewski K., (1999), *Komu jest potrzebna diagnostyka oświatowa?* (w:) B. Niemierko i B. Machowska (red.), *Diagnoza edukacyjna. Oczekiwania, problemy, przykłady.* Ośrodek Diagnozy, Egzaminów Szkolnych i Informacji Pedagogicznej, Legnica.

Krope P., (1994), *Ocena opisowa w pedagogice źródłem nieporozumień* (w:) B. Niemierko (red.), *Diagnostyka edukacyjna*, Wyd. Uniwersytetu Gdańskiego, Gdańsk.

Kruszewski K., (2003), *Od tłumacza* (w:) J. MacBeath i in., *Czy nasza szkoła jest dobra?*, WSiP, Warszawa.

Lord F.M., (1970), *Some theory for tailored testing*, (w:) W. Holtzman (red.), Computer assisted instruction, testing, and guidance, Harper, New York.

MacBeath J., Schratz M., Meuret D., Jakobsen L., (2003), *Czy nasza szkoła jest dobra?*, WSiP, Warszawa.

Magnusson D., (1981), *Wprowadzenie do teorii testów*, PWN, Warszawa.

McMillan J.H., (2003), *Understanding and improving teacher's Classroom Assessment Decision Making*, „Educational Measurement: Issues and Practice. Special Issue *Changing the wa measurement theorists think about classroom assessment*, nr 4.

Messick S., (1989), *Validity*, (w:) R. L Linn (red.) *Educational Measurement. Third Edition*, American Coucil on Education, Washington.

Moss P.A., (2003), *Reconceptualizing validity for classroom assessment*, „Educational Measurement: Issues and Practice. Special Issue *Changing the way measurement theorists think about classroom assessment*, nr 4.

Mulawa J., Sroka D., (2003), *Trzy lata Trzecioteścika, czyli co można powiedzieć o kompetencjachuczniów trzecich klas szkół podstawowych w latach 2000 — 2003*, IBK, Wrocław.

Niemierko B., (1975), *Testy osiągnięć szkolnych. Podstawowe pojęcia i techniki obliczeniowe*, WSiP, Warszawa.

Niemierko B., (1990), *Pomiar sprawdzający w dydaktyce. Teoria i zastosowania*, PWN, Warszawa.

Niemierko B., (1999), *Pomiar wyników kształcenia*, WSiP, Warszawa.

Niemierko B., (2000), *Kształcenie według wymagań*, (w:) K. Kruszewski (red.), *Pedagogika w pokoju nauczycielskim*, WSiP, Warszawa.

Nowakowska M., (1975), *Psychologia ilościowa z elementami naukometrii*, PWN, Warszawa.

Poręba-Konopczyńska A., (2001), *Test zespołowy z wyposażeniem jako narzędzie oceniania*, (w:) K. Wenta (red.) *Pomiar edukacyjny jako kompetencje pedagogiczne.* Wyd. Naukowe Uniwersytetu Szczecińskiego, Szczecin.

Potworowski J., (2000), *Pierwiastek angielski w transformacji polskiej oświaty*, (w:) K. Kruszewski (red.), *Pedagogika w pokoju nauczycielskim*, WSiP, Warszawa.

PTDE, (2003), *O potrzebie diagnozy i diagnostyki edukacyjnej*, www.ptde.org.

Rozporządzenie Ministra Edukacji Narodowej z dnia 21 marca 2001 r. w sprawie warunków i sposobu oceniania, klasyfikowania i promowania uczniów i słuchaczy oraz przeprowadzania egzaminów i sprawdzianów w szkołach publicznych, (2001), Dz. U. Nr 29, poz. 323.

Reckase M.D., (1989), *Adaptive testing: The evolution of a good idea.* „Educational Measurement: Issues and Practice", nr 3.

Sato T., (1971), *Analiza danych dotyczących osiągnięć uczniów*, (w:) K. Hirata, T. Sato (red.), [*Analizowanie odpowiedzi*], Kyoiku-Kogakusha, Tokio.

Smith J.K., *Reconsidering reliability in classroom assessment and grading.* „Educational Measurement: Issues and Practice. Special Issue *Changing the way measurement theorists think about classroom assessment*, nr 4.

Szmigel M.K., (2004), *Treściowe znaczenie wyniku egzaminacyjnego ucznia i szkoły na przykładzie sprawdzianu w 2004 roku*, (w:) B. Niemierko, H. Szaleniec (red.), *Standardy wymagań i normy testowe w diagnostyce edukacyjnej*, Okręgowa Komisja Egzaminacyjna, Kraków.

Wainer H., Kiely G.L., (1987), *Item clusters and computerized adaptive testing: A case for testlets*, „Journal of Educational Measurement", 24, s. 185-201.

Wlazło S., (1999), *Mierzenie jakości pracy szkoły. Część trzecia*, MarMar, Wrocław.