Bolesław Niemierko
Uniwersytet Gdański

Jaki pomiar dydaktyczny jest nam potrzebny?

Wstęp

W artykule zamierzam pokazać, jak rozszerza się w pedagogice pojęcie „trafności pomiaru". Te zmiany osłabiają fundamenty metody testowej i dlatego towarzyszy im opór. Jednak mimo pewnego rozmycia założeń, rośnie edukacyjna użyteczność pomiaru, a egzaminy doniosłe stają się polem fascynujących doświadczeń.

Zmiany, które przedstawię, obejmują cztery dziedziny:

1. coraz większe zainteresowanie kontekstem społecznym egzaminowania;
2. pomnażanie teorii i technik analizy wyników pomiaru;
3. rozwój problematyki komunikowania wyników egzaminu;
4. powstawanie krajowych strategii diagnostyki edukacyjnej.

Zakres zmian jest tak wielki, że można mówić o „nowym otwarciu" lub o „drugiej młodości" pomiaru dydaktycznego na świecie i – bez wątpienia – w Polsce. „Trafność teoretyczna pomiaru" w ujęciu Samuela Messicka jest w tym procesie widomym znakiem czasu i motorem postępu.

Przeczytałem kiedyś, że **dobry styl** to powaga w sprawach drobnych i swoboda w sprawach wielkich. Zauważyłem, że podobne odwrócenie ekspresji pomaga w objaśnianiu pomiaru dydaktycznego. Trzeba pisać **emocjonalnie** o kwestiach technicznych i **racjonalnie** o emocjach, jakie pomiar wywołuje. Spróbuję jednego i drugiego, a zacznę od pedagogiki.

Pedagogika jako nauka praktyczna

Rozróżnienie nauk „ścisłych" i „stosowanych", „teoretycznych" i „praktycznych", „podstawowych" i „skierowanych" jest dzisiaj niemodne, bo te pierwsze coraz szybciej znajdują zastosowanie, a te drugie, znacznie sowiciej finansowane, nie chcą dłużej pozostawać w cieniu. Mimo to warto sięgnąć do dawnych rozróżnień, by zrozumieć tendencje **diagnostyki edukacyjnej**, jako nauki o warunkach, przebiegu i wynikach zorganizowanego uczenia się.

Odpowiada mi stanowisko Heliodora Muszyńskiego, który swoistość pedagogiki dostrzegł w jej „innowacyjnym charakterze" (Muszyński 1970: 53–55). By ulepszyć kształcenie (według Muszyńskiego – „wychowanie"), pedagogika posługuje się badawczą **strategią teleologiczną**, polegającą na szukaniu najbardziej skutecznej drogi do osiągnięcia „z góry założonych celów". W przeciwieństwie do badań **obiektywistycznych**, dążących do wyjaśnienia określonych dziedzin rzeczywistości, strategia teleologiczna obejmuje wybór właściwych **wartości**, uzależniając się przez to od poglądów grup ludzi (od ich „ideologii").

Z dwóch problemów badawczych:

1. Co powoduje różnice w skuteczności kształcenia?
2. Jak w danych warunkach zapewnić zadowalającą skuteczność kształcenia określonej grupy uczniów w wybranym zakresie programowym?

wolę ten drugi, choć zdaję sobie sprawę, że rozwiązanie pierwszego, choćby średnio dokładne, ułatwiłoby uzyskanie odpowiedzi na nieskończoną liczbę problemów drugiego typu, natomiast rozwiązanie drugiego jest **incydentalne** i raczej nie nadaje się do uogólnienia.

Badania stosowane

Dotychczasowe badania osiągnięć uczniów w skali światowej, oparte na starannie dobranych próbach reprezentatywnych, nie pozwoliły zbudować dydaktyki naukowej. Najściślejsze z tych badań, dotyczące kształcenia matematycznego, a zorganizowane przez Międzynarodowe Stowarzyszenie Badań Osiągnięć Pedagogicznych (IEA), zostały zamknięte takim oto stwierdzeniem (Robitaille, Garden, 1989: 240):

> Obraz, jaki ukazał się dotychczas, jest **złożony i wieloaspektowy**. Nauczyciele i uczniowie z różnych środowisk społecznych i kulturalnych mają wiele wspólnych poglądów i postaw na temat matematyki i nauczycieli matematyki. Zarazem różnią się bardzo między sobą co do innych, równie ważnych zmiennych. **Nie udało się w tym sprawozdaniu wznieść w analizach tych podobieństw i różnic ponad odnotowanie ich istnienia** i, w paru przypadkach, próbę badania możliwych związków z osiągnięciami uczniów. Ogromna większość dalszej pracy musi być wykonana **wewnątrz systemów dydaktycznych**, a także międzynarodowo, jeżeli pełny pożytek z udziału w tych badaniach ma być uzyskany [wszystkie podkreślenia moje – B. N.].

Szanse na rozwiązanie pilnych problemów dydaktycznych przez rozwijanie **badań podstawowych**, ukierunkowanych na poszukiwanie ogólnych prawidłowości, są w perspektywie bliskiej i średniookresowej niewielkie. Rośnie więc potrzeba **badań stosowanych**, ukierunkowanych na lokalne usprawnienia. Te dwa rodzaje badań różni wiele właściwości:

Tabela 1.
Porównanie badań podstawowych i badań stosowanych nad kształceniem szkolnym

Właściwość	Badania podstawowe	Badania stosowane
Główny cel	tworzenie teorii dydaktycznej	rozwiązywanie problemów oświatowych
Strategia badawcza	znajdowanie przyczyn	sprawdzanie skutków
Oczekiwany wynik	statystycznie istotne zależności	praktycznie znaczący wzrost skuteczności
Przedmiot badań	teoria dydaktyczna	programy, warunki, wyniki działania
Inicjator badań	instytut lub badacz naukowy	administracja, ośrodek metodyczny
Tematyka badań	według zainteresowań badaczy	według potrzeb oświaty
Typ badacza	ścisła specjalizacja naukowa	wielostronna kompetencja w dziedzinie
Orientacja badacza	na trwały wynik naukowy	na bieżące potrzeby oświaty
Metody badań	najlepsza dostępna metoda	układ wielu różnorodnych metod
Warunki przebiegu	ściśle kontrolowane	naturalne, niezakłócone badaniami
Czas trwania	regulowany według przebiegu	na ogół krótki, z góry wyznaczony
Koszt badań	szacowany z etapu na etap	ograniczony, z góry określony

Trzeba zaznaczyć, że dopiero zbiór kilku właściwości spośród ujętych w tabeli 1 powoduje, że pewien projekt badawczy można określić jako „podstawowy" lub „stosowany", a i to jego klasyfikacja może być dyskusyjna. Spotyka się też stanowisko pogardliwe wobec badań stosowanych oraz wobec dyscyplin naukowych je uprawiających. Z pozycji pedagogiki „czystej" – badania stosowane są wtórne, nieprecyzyjne i sprzedajne, nieuczciwie pozorujące naukowość w dziedzinach nie objętych jeszcze właściwą teorią.

Emocjonalny składnik diagnostyki

Co najbardziej oburza strażników „czystości" nauk społecznych, w tym głównie – psychologicznych, to przenikanie **czynnika emocjonalnego** do pedagogicznych badań stosowanych. Prześledźmy następującą tabelę:

Tabela 2.
Porównanie czynności kształcenia z naukową pracą badawczą

Właściwość	Kształcenie	Praca badawcza
Główne walory działania	twórczość, elastyczność	przewidywanie, planowość
Wykorzystanie środków	różnorodność, bogactwo	maksymalna oszczędność
Postawa intelektualna wykonawcy	pełne zaangażowanie	obiektywizm, krytycyzm
Postawa emocjonalna wykonawcy	entuzjazm, wiara	chłodny sceptycyzm
Dominująca organizacja działania	praca indywidualna	praca zespołowa
Warunki komunikacji	zaufanie, chęć współpracy	wysoki poziom kompetencji
Główne walory wyniku	oryginalność, niepowtarzalność	poprawność, sprawdzalność
Styl komunikacji	obrazowość, ekspresja	ścisłość, zwięzłość

Kształcenie jest niewątpliwie **zdominowane emocjonalnie** w przeciwieństwie do pracy naukowej, która osobiste przekonania badacza musi trzymać na wodzy. Cóż jednak sądzić o uczonych demonstrujących dystans, a nawet niechęć do nauczycieli? Czy można to usprawiedliwić niedostosowaniem, w jakie nieuchronnie wpada zdolny uczeń (przyszły uczony) w naszej typowej szkole (Niemierko, 2002a: 101 i nast.)?

Pomijając przypadki skrajne, zauważymy, że niełatwo jest nauczycielom zaakceptować pomiarowe **działania zewnętrzne** w ich własnym, gorącym polu pracy edukacyjnej. Tym bardziej, gdy są dokonywane w imię beznamiętnej precyzji, bez świadomości drogi, jaką przebywa uczeń i bez przyjmowania odpowiedzialności za skutki społeczne. Demonstrowanie braku emocji jest w edukacji odbierane jako wywyższanie się i pogarda dla otoczenia.

Pedagog a psycholog

Kolejna tabela pokaże nam różnice między pracą **nauczyciela** a pracą **psychologa-terapeuty** nad regulacją procesów uczenia się. Przy właściwym podziale zadań i zrozumieniu granic skutecznego oddziaływania, obydwaj mogą być niezbędni w szkole.

Psycholog pracuje „ciszej", w warunkach dużej **dyskrecji**, a pomiar, jakiego dokonuje, nie bywa sprawą publiczną, chyba że zachodzi tu jakaś wątpliwość etyczna. Ogłaszanie „ilorazów inteligencji" uczniów, a tym bardziej nauczycieli, nie przyniosłoby żadnego pożytku ani słabym pod tym względem, ani silnym. Podobnie jak lekarz i spowiednik, psycholog znaczną część danych zachowuje dla siebie, a stara się przekazać na zewnątrz budujące wnioski.

Tabela 3.
Porównanie czynności psychologa i nauczyciela w zakresie diagnostyki edukacyjnej

Właściwość	Nauczyciel	Psycholog
Cele działania	kierowanie uczeniem się w szkole	wspomaganie uczenia się
Cele diagnozy	dostarczanie informacji zwrotnej	pokonywanie trudności w uczeniu się
Typowy obiekt diagnozy	grupa uczniów	pojedynczy uczeń
Typowe warunki diagnozy	lekcja, izba szkolna	gabinet diagnostyczno-terapeutyczny
Narzędzia pomiarowe	testy osiągnięć, kwestionariusze	inwentarze osobowości, testy uzdolnień
Organizacja pomiaru	zbiorowa	indywidualna
Dostępność wyniku	wysoka lub pełna	obwarowana kompetencją odbiorcy
Główny odbiorca wyniku	uczeń, rodzice ucznia	rodzice ucznia, nauczyciel
Wtórny odbiorca wyniku	nadzór pedagogiczny, społeczeństwo	dyrekcja szkoły
Wpływ pomiaru na nauczanie	bezpośredni, dominujący	pośredni, ograniczony
Wpływ na wychowanie	pośredni, długofalowy	bezpośredni, głęboki
Znaczenie polityczne pomiaru	duże lub bardzo duże	niewielkie lub żadne

Inaczej jest z pomiarem dydaktycznym. Tu pracuje się „przy otwartej kurtynie". Już dawno uznano, że zatajanie stopni jest pozbawianiem uczniów potrzebnej im **samosterowności**, a społeczeństwo domaga się **pełnej informacji** o wynikach doniosłych egzaminów szkolnych. Dziennikarze wyrywają tabele wyników z rąk egzaminatorów i – niepomni ostrzeżeń – interpretują je po swojemu.

Nauczyciel nie jest w stanie zastąpić psychologa w diagnozie osobowości, a psycholog nie zastąpi nauczyciela w pomiarze wyników kształcenia. Ta świadomość powoduje stopniowe wyodrębnianie się teorii pomiaru dydaktycznego i diagnostyki edukacyjnej.

Dydaktyczne rewolucje pomiarowe

Mimo że pedagogika rozwijała się już w starożytności, a dzieje psychologii jako nauki mają swój początek dopiero w połowie XIX w., to psychologom zawdzięczamy **klasyczną teorię pomiaru** stosowaną w naukach społecznych.

Psychologiczna teoria pomiaru liczy sobie około 100 lat, a jej punktem startu była analiza wpływu błędów pomiaru na współczynnik korelacji dwu zmiennych

(Spearman, 1904). Pedagogiczna teoria pomiaru liczy sobie co najwyżej połowę tego okresu, a każdy jej krok ku samodzielności był okupiony ostrą, a nawet miażdżącą krytyką.

Opiszę cztery takie kroki, wszystkie dokonane w Stanach Zjednoczonych i stamtąd promieniujące na resztę świata. Trzy wcześniejsze zostały dość dokładnie scharakteryzowane w mojej książce *Pomiar sprawdzający w dydaktyce* (1990), a czwarty, najnowszy, w książce *Pomiar wyników kształcenia* (1999, rozdz. X).

1. Rewolucja w planowaniu pomiaru. Gdy w połowie lat 50. ubiegłego stulecia krajowy komitet pedagogów, psychologów i politologów amerykańskich ogłosił, po paru latach pracy, **niemożność** pojęciowego ogarnięcia dziedziny celów kształcenia, kilkuosobowy zespół Benjamina Blooma zaryzykował przedstawienie dorobku tego komitetu w kilkunastu hierarchicznie uporządkowanych kategoriach (Bloom, 1956). Tak powstała przesławna **taksonomia Blooma** o dwu niewątpliwych właściwościach:

- nieustająca krytyka ze strony przeciwników behawioryzmu, przeciwników klasyfikacji celów kształcenia i przeciwników szkoły;
- nieprzebrane liczby naśladowców i przypadków systematycznego stosowania przez konstruktorów testów i programów kształcenia na całym świecie.

Na terenie dydaktyki fenomen taksonomii Blooma może być porównany z karierą **stopni formalnych** Herbarta: prawie wszyscy są im przeciwni i prawie wszyscy je stosują. Najbardziej zacięci przeciwnicy planowania testów osiągnięć szkolnych według taksonomii celów kształcenia chcieliby cele kształcenia wyprowadzać z prawidłowości statystycznych (współczynników korelacji), stawiając **ilość** przed **jakością** i drwiąc z pozanaukowego „folkloru nauczycielskiego". Im właśnie Robert Brennan (2001a: 11), czołowy dziś specjalista statystycznej teorii generalizacji, przypomina pogląd Lindquista (1953) sprzed pięćdziesięciu laty:

> Dobry test osiągnięć szkolnych powinien być wykładnią celów, jakie mierzy. To znaczy, że metoda skalowania osiągnięć szkolnych nie może ingerować w treść testu lub zmieniać sformułowania celów objętych testem. **Dla autora testu definicje celów są świętością** [podkreślenie moje – B. N.]; nie ma prawa się nimi zabawiać. Cele są mu przekazywane przez tych działaczy społecznych, którzy odpowiadają za cele kształcenia, a konstruktorowi pozostaje włączyć te cele możliwie jasno i dokładnie do egzaminu, jaki buduje.

2. Rewolucja w założeniach pomiaru. Gdy z początkiem lat 60. **nauczanie programowane** przyjęło postać alternatywnego systemu dydaktycznego, pomiar osiągnięć oparty na psychologii różnic indywidualnych przestał wystarczać do wykazania, że programy są niezawodne w sensie doprowadzenia uczniów do **pełnego opanowania** (ang. *mastery*) przewidzianych wiadomości i umiejętności. Pojawiła się koncepcja **pomiaru sprawdzającego** osiągnięcia uczniów, bezpośredniego i absolutnego (Glaser, 1963), jako przeciwstawienie **pomiaru różnicującego** uczniów według **zmiennej ukrytej**, pośredniego i względnego.

Taka zmiana była szokiem i zdradą. Tak to właśnie przeżyłem osobiście, gdy poznałem nową koncepcję na kursie ewaluacji pedagogicznej zorganizowanym przez IEA w Szwecji w 1971 r., a więc tuż po wyciągnięciu pomysłu Glasera na światło dzienne przez dwu jego zapalonych zwolenników (Popham, Husek, 1969). Zachwiała się podstawa wiedzy, jaką o pomiarze miałem. Uspokoiło mnie dopiero spostrzeżenie, z jak wielką skwapliwością dydaktycy przedmiotowi w Polsce i na świecie akceptują założenia pomiaru sprawdzającego.

Entuzjazmowi pedagogów towarzyszyło umiarkowane uznanie psychologów. Wielu z nich zgodziło się zaledwie na **dwie interpretacje** wyników pomiaru – według „norm" (empirycznych) i według „kryteriów" (programowych), nie dostrzegając różnic w planowaniu i budowie testów. Inni pozostali w przekonaniu, że „każdy test jednocześnie sprawdza i różnicuje", a więc nie ma powodu, by sprawdzanie wyodrębniać. Nawet po upływie ćwierćwiecza nie pozbyto się zasadniczych wątpliwości co do sensu pomiaru sprawdzającego (Linn, 1994).

3. Rewolucja w funkcjonowaniu pomiaru. W tejże dobie nauczania programowanego zauważono, że zespoły ekspertów programowych znacznie chętniej przyjmują wyniki weryfikacji empirycznej na początku niż na końcu wykonywanej pracy. Tak powstało pojęcie **ewaluacji kształtującej** (formatywnej, konstruktywnej), która – w przeciwieństwie do **ewaluacji sumującej** – dokonywana jest w toku, nie zaś na koniec działania (Scriven, 1967). Przeniesione do praktyki szkolnej znaczyło, że osiągnięcia uczniów powinny być mierzone w całym procesie kształcenia, a sprawdzanie **bieżące** tych osiągnięć jest ważniejsze od **końcowego**.

Ewaluacja kształtująca podnosiła rolę psychologów szkolnych i bardzo szybko zawładnęła wyobraźnią pedagogów, ale to wcale nie znaczy, że znalazła uznanie społeczne. „Wpierw dobrze naucz, a dopiero potem sprawdzaj" jest logiką większości niespecjalistów, a wczesne egzaminowanie uczniów uważa się za niesprawiedliwe i niepotrzebne. „Po co nam kosztowny sprawdzian po szkole podstawowej, skoro i tak wszyscy idą do gimnazjum?" – czytamy w gazetach, zaś z rozkładaniem matury na kilka końcowych lat kształcenia ogólnego nie udało się w Europie zajść daleko. Wciąż dominuje przekonanie, że **egzamin doniosły** wieńczy pracę ucznia, nie należy nim zakłócać uczenia się, a odkrycia „całej prawdy" o działaniach szkoły dokonuje się najskuteczniej przez badanie absolwentów.

4. Rewolucja w wartościowaniu pomiaru. Ta rewolucja, rozpoczęta przez Samuela Messicka (1989, 1995a, 1995b), jeszcze trwa. Poprzedziły ją głębokie zmiany w modelu i procedurach ustalania **trafności** pomiaru psychologicznego i dydaktycznego, jakie dokonują się od lat międzywojennych.

Najpierw trafność rozumiano najzupełniej technicznie, jako korelację wyników pomiaru z **kryterium** w postaci przyszłych sukcesów (szkolnych, akademickich, zawodowych, życiowych) ucznia, to jest jako **trafność prognostyczną** lub (później) **kryterialną**. Wkrótce bardziej potrzebne i szerzej dostępne okazało się pojęcie **trafności wewnętrznej** pomiaru, zwanej także **trafnością programową**. To właśnie pedagodzy „wymusili" przejście od (ilościowej) analizy związków korelacyjnych do (jakościowej) analizy czynności wykonywanych przez uczniów dla rozwiązania zadań testu.

Wysoką rangę naukową nadały trafności pomiaru prace Lee Cronbacha (Cronbach, Meehl, 1955; Cronbach, 1971). Ukształtowały one pojęcie **trafności teoretycznej**, oparte na celowej konstrukcji pojęcia, które ma być reprezentowane pomiarem (ang. *construct validity*). Na przykład „świadomość historyczna", „językowa" i „kulturowa", „naukowy obraz przyrody", „znajomość języka ojczystego" i „obcego", „rozumienie czytanego tekstu literackiego" i „użytkowego", „umiejętność dokonywania działań na liczbach i wyrażeniach" – to pojęcia, które mają bogaty sens teoretyczny i w związku z tym powinny być starannie przedyskutowane, nim powstanie zbiór zadań je reprezentujący. Nie wystarczy wszakże zbudować i poprawnie zastosować takie zadania. Jako jeden z twórców ogólnej teorii decyzji, Lee Cronbach wyjaśnił, że to właśnie **decyzje** podejmowane na podstawie wyników pomiaru wymagają właściwego uzasadnienia („walidacji"). Tak więc ciężar dowodu trafności przesunął się z „testu" na „pomiar", z „pomiaru" na „wynik pomiaru", a z „wyniku pomiaru" na „decyzje oparte na wyniku pomiaru". Czy mógł się posunąć jeszcze dalej w procesie pomiarowym?

Okazuje się, że tak. Działając w przedsiębiorstwie Testowe Usługi Edukacyjne (*Educational Testing Service* – ETS), Samuel Messick obserwował konflikty, jakie narastały wokół pomiaru dydaktycznego. Testom standaryzowanym ETS zarzucano stronniczość rasową i kulturową, zaburzanie pracy szkoły, arbitralność i jednostronność, wyobcowanie z życia społeczności lokalnej. Messick doszedł do wniosku, że to wszystko trzeba brać pod uwagę, projektując zastosowanie pomiaru dydaktycznego i psychologicznego, a przewidywane **konsekwencje społeczne** pomiaru są istotnym składnikiem jego trafności teoretycznej. Tak powstało rozszerzone, zintegrowane pojęcie trafności teoretycznej pomiaru dydaktycznego, ujęte przez jego autora w następującej „macierzy postępującej" (*progressive matrix*):

Tabela 4.
Cztery zakresy trafności teoretycznej według Messicka

Uzasadnienie trafności	Funkcje trafności	
	Interpretacja testu	Zastosowanie testu
Na podstawie zebranych dowodów	**A. Trafność teoretyczna**	**B.** Trafność teoretyczna **+ stosowność i użyteczność**
Na podstawie przewidywanych konsekwencji („trafność konsekwencyjna")	**C.** Trafność teoretyczna **+ kształtowanie wartości**	**D.** Trafność teoretyczna + stosowność i użyteczność + kształtowanie wartości **+ skutki społeczne**

Trafność teoretyczna rozszerzona obejmuje wszystkie wcześniej znane rodzaje trafności (w polach A i B), a nadto wprowadza dwa nowe postulaty: (C) pozytywne oddziaływanie wychowawcze i (D) dodatni bilans skutków dla całej edukacji. Rewelacją jest oparcie dolnego wiersza tabeli na niepewnych przewidywaniach przyszłości,

nie zaś na jakimkolwiek zbiorze danych z badań (*evidence*), co było wymagane dotychczas.

Rozluźnienie procedur oceny jakości pomiaru zaowocowało wzrostem jego znaczenia społecznego. Messick (1995b: 5) tak to ujmuje:

> [...] trafność, rzetelność, porównywalność i bezstronność są nie tylko zasadami pomiaru; one są **wartościami społecznymi** o znaczeniu i doniosłości w każdym sądzie wartościującym i każdej decyzji [edukacyjnej]. Jako naczelna wartość społeczna, trafność odgrywa rolę naukową i społeczną, której w żadnym wypadku nie może podołać zwykły współczynnik korelacji między wynikami testu a wybranym kryterium (klasyczna trafność kryterialna) lub opinia ekspertów, że treść testu odpowiada proponowanemu zastosowaniu testu (tradycyjna trafność wewnętrzna). W istocie, ujmując to szeroko, trafność jest niczym mniej niż wartościującym podsumowaniem zarówno danych pozytywnych, jak i obecnych oraz potencjalnych konsekwencji interpretacji i wykorzystania wyników pomiaru (trafność teoretyczna szeroko rozumiana).

Dzieje trafności rozszerzonej

Na temat trafności przewidywania konsekwencji pomiaru, nazwanej potem krótko **trafnością konsekwencyjną** (*consequential validity*), napisano już kilkaset rozpraw i wydano wiele monograficznych numerów profesjonalnych czasopism pomiarowych. Jednych autorów oszałamiają horyzonty nowego podejścia, lecz innych oburza naginanie ukształtowanych już pojęć do chwiejnych wyobrażeń i postaw laików.

Sam Cronbach przestrzegał przed osłabianiem pojęcia trafności, zauważając, że wciąganie wszystkiego pod wspólny parasol jest rezygnacją z porządnej teorii, a „wielu konstruktorów traktuje trafność teoretyczną testów jak kosz na śmieci" (cyt. za: Kane, 2001: 326). Zachęcał, by „myśleć raczej o *dyskursie trafności* niż o (empirycznym) *badaniu trafności*" (tamże: 329). Wkrótce dyskurs przybrał miano „sokratejskiego", a jego składniki jako: (1) wyjaśnianie, (2) ewaluację, (3) uogólnianie, (4) ekstrapolację i (5) podejmowanie decyzji. Dwa początkowe procesy można uznać za część **opisową** dyskursu, a trzy pozostałe – za jego część **projektującą**.

> Uczyniono postęp. Przebyliśmy drogę od dość prostych modeli kryterialnych do całkiem wyrafinowanych modeli teoretycznych

– konkluduje Kane (tamże: 339). Nie wszyscy specjaliści pomiaru uważają jednak tę zmianę za pomyślną. Na przykład James Popham, entuzjasta pomiaru sprawdzającego i ewaluacji kształtującej, jest zdania, że teoretycy potrzebują ściślejszych pojęć niż użytkownicy pomiaru i wszelkie koncesje populistyczne w tym zakresie są szkodliwe. Co więcej, przewiduje, że

> Messick i inni zwolennicy trafności konsekwencyjnej spowodują, że szeregowi edukatorzy nie będą mogli zrozumieć, czym trafność pomiaru naprawdę jest. Przez obciążenie pojęcia trafności ładunkiem nie do

> udźwignięcia, obrońcy trafności konsekwencyjnej zatracili przejrzystość sensu trafności oceniania (Popham, 1997: 12).

To pisze jeden z najbystrzejszych i najdowcipniejszych ludzi z branży pomiarowej, z niezwykłym uporem dążący do jasności w kwestiach niepodatnych na uproszczenia... Inny wybitny znawca pomiaru, William Mehrens (1997) wręcz sugeruje, że

> [...] społeczność psychometryczna powinna raczej *zwęzić* pojęcie trafności niż je rozszerzać.

Problematyka **wartości** (*value judgement*) w metodologii badań społecznych może być różnie traktowana: od wzgardliwej eliminacji (w orientacjach pozytywistycznych) do pełnej dominacji (w orientacjach humanistycznych). Działalność (nieżyjącego już, niestety) Samuela Messicka i jego następców przesuwa pomiar dydaktyczny ku tej drugiej. Opowiadając się za taką zmianą, nie powinniśmy jednak zaniedbywać postępu technicznego, jaki się równolegle dokonuje i jaki jest nam niezbędny w reformie egzaminacyjnej.

Probabilistyczne modele pomiaru

Gdy w końcu lat 70. ubiegłego stulecia zacząłem studiować „model Rascha" i inne **probabilistyczne modele wyniku zadania testowego** (*item response theory* – IRT), było wiele przeszkód (poza ideologicznymi, oczywiście) wobec ich upowszechniania w naszej dydaktyce. Do najważniejszych należały:

- założenie o „lokalnej niezależności" wyniku zadania testowego, to znaczy o pełnej jednorodności mierzonej właściwości, która jako „cecha ukryta" powinna dokładnie wyznaczać interkorelacje zadań, co znaczy, że moc różnicująca zadań testowych powinna być jednolicie wysoka;
- ograniczenie (pierwotnie) do zadań punktowanych 0–1;
- duża liczba potrzebnych danych (co najmniej dwutysięczne próby uczniów);
- pracochłonne (iteracyjne) szacowanie parametrów zadań;
- duża trudność matematyczna teorii.

Przez dwadzieścia lat nie udało mi się bliżej zainteresować tą teorią żadnego z dydaktyków nauk ścisłych, sam zaś ograniczyłem się do ogólnych informacji o niej w książce *Pomiar sprawdzający w dydaktyce* (1990) oraz do pewnych analogii w analizie zadań (Niemierko, 1999, rozdz. IX; Niemierko, 2002a, rozdz. 19).

Obecna sytuacja jest wszakże zasadniczo odmienna od dawnej:

1. Dostarczono wiele dowodów na użyteczność IRT – do budowania testów, tworzenia banków zadań, zrównywania wyników testowania, weryfikacji hipotez statystycznych – przy przybliżonym spełnieniu postulatu lokalnej niezależności, to jest przy ograniczonej spójności mierzonej właściwości.
2. Zbudowano modele probabilistyczne dla zadań punktowanych według skali rozwiniętej (*partial credit*: Lord, 1980; Hambleton, Swaminathan, 1985).

3. Pojawiły się w Polsce zobiektywizowane egzaminy obowiązkowe, czyli badania wyczerpujące kilku populacji uczniów, tworzące półmilionowe bazy danych.
4. Komputery nowych generacji znakomicie wyręczają nas w złożonych obliczeniach.
5. Kilka osób dobrze przygotowanych matematycznie – Henryk Szaleniec z Krakowa, Jerzy Chodnicki z Łomży, Marek Kryniewski z Gdańska – zainteresowało się IRT.

Stan obecny zastosowań probabilistycznych modeli wyniku zadania testowego w dydaktyce polskiej można określić jako okres prób laboratoryjnych. Wstępne wyniki tych prób (Szaleniec, 2002; Kryniewski, 2002) są zachęcające.

Największą atrakcją modeli probabilistycznych jest możliwość odtwarzania brakujących części **krzywych charakterystycznych** zadań, czyli zbiorów oszacowań prawdopodobieństwa rozwiązania określonych zadań testowych przez określonych uczniów, na podstawie danych fragmentarycznych: tylko pewnych zadań (np. tylko łatwych lub tylko trudnych) lub tylko pewnych uczniów (np. tylko o niskich lub tylko o wysokich osiągnięciach). Pozwala to na poczekaniu (w systemie adaptacyjnym, zindywidualizowanym) budować dla wybranych uczniów testy o żądanych parametrach **łatwości i błędu pomiaru** osiągnięć szkolnych.

Właściwości testki

Testką (*testlet*) będziemy nazywali zbiór zadań, także jednoelementowy, stanowiący część testu osiągnięć szkolnych wyróżnioną ze względu na konstrukcję, zastosowanie lub sposób punktowania (Lee i in., 2000: 10). Technologia testki budzi rosnące zainteresowanie w USA (Wainer, Lewis, 1990), a w Polsce, ze względu na łączenie zadań otwartych z zamkniętymi w sprawdzianie i egzaminach doniosłych, staje się z dnia na dzień niezbędna.

Stosujący współczynnik rzetelności „alfa" do testów o mieszanej konstrukcji z pewnością zauważyli, że ich test „tracił" na rzetelności po włączeniu zadania rozszerzonej odpowiedzi (zazwyczaj mierzącego umiejętność budowania pisania wypowiedzi ciągłej lub stanowiącego próbę pracy) pomimo bardzo wysokiej mocy różnicującej tego zadania. Dzieje się tak na skutek zakłócenia wewnętrznej zgodności testu przez łączenie zadań o różnej konstrukcji i różnym sposobie punktowania.

Konsekwencje niejednorodnej budowy testu nie ograniczają się do rzetelności. Lee, Brennan i Frisbie (2000: 11) stwierdzają, że:

> [...] wiele testów pedagogicznych lepiej byłoby traktować raczej jako złożone z testek niż jako niezróżnicowane zbiory zadań lub prób działania. Zatem analizy wielu testów powinny być świadomie dokonywane według testek. Testki mogą być pominięte tylko wtedy, gdy można wykazać, że mają one niewielki wpływ na dane zagadnienie pomiarowe (np. na trafność, rzetelność, błąd pomiaru, zrównywanie i skalowanie wyników).

Gdy chodzi o rzetelność testu złożonego z testek, to najprostszym (choć nie zawsze łatwym) rozwiązaniem problemu rzetelności jest podzielenie testu na dwie części w ten sposób, by każda z testek była dokładnie przepołowiona, ustalenie rzetelności **połówki testu**, a następnie zastosowanie wzoru „proroczego" Spearmana-Browna. Bardziej wyrafinowane rozwiązania tego problemu są oparte na modelach probabilistycznych pomiaru i na teorii generalizacji wyniku pomiaru.

Generalizacja wyniku pomiaru

Teoria generalizacji, nazwana też w Polsce (Nowakowska, 1975: 59–79) **teorią uniwersalizacji**, bada warunki uogólniania wyników pomiaru psychologicznego i dydaktycznego. Te warunki to głównie: treść zadań, forma zadań, upływ czasu, organizacja testowania i punktowanie wyników przez egzaminatora. Każdy z nich można ujednolicić lub potraktować jako źródło zmienności. Im większy jest zakres czynników zmienności, tym niższa jest powtarzalność wyników pomiaru wyrażana odpowiednim współczynnikiem rzetelności.

Jak wyjaśniałem to w książce *Pomiar wyników kształcenia* (1999, rozdz. XI), psychologowie byli przez wiele lat najbardziej zainteresowani zmiennością wyników pomiaru (1) między wersjami równoległymi testu i (2) wraz z upływem czasu, a dla dydaktyków nieważna jest stałość wyniku w czasie (starają się przecież osiągnięcia uczniów podnieść), za to przemożnym źródłem zmienności wyników bywa (3) punktowanie zadań otwartych i zadań praktycznych (dokładność punktowania). By uogólnić wynik na wybraną dziedzinę treści kształcenia i na populację wykwalifikowanych egzaminatorów (lub innych punktujących), musimy dysponować odpowiednimi powtórzeniami pomiaru – kombinacjami wartości tych warunków. To samo dotyczy ewentualnych dalszych czynników zmienności wyników.

Teoria generalizacji wiąże rzetelność pomiaru (powtarzalność) z jego trafnością (uwarunkowania) i z tego względu zakres jej zastosowań stale rośnie (Brennan, 2001b). Jej znaczenie dla diagnostyki edukacyjnej, zorientowanej na podnoszenie skuteczności uczenia się, jest bezsporne. Co może natomiast zniechęcać do tej teorii, to spadek rzetelności i wzrost błędów pomiaru przy uwzględnianiu kolejnych okoliczności, nad którymi nie w pełni panujemy. Cóż, „człowiek z jednym zegarkiem zawsze wie, która jest godzina, ale człowiek z dwoma zegarkami nigdy nie jest tego zupełnie pewien" – stwierdza filozoficznie Robert Brennan (2001a).

Zrównywanie wyników pomiaru

Kolejne zagadnienie pomiarowe wtórnie odkrywane w Polsce to **zrównywanie wyników pomiaru** (*equating*), rozumiane jako określanie wyników równoważnych dwu lub więcej wersji testu, bezpośrednio, przez dwukrotne badanie tych samych uczniów, lub przez „kotwiczenie", to jest przez wykorzystanie innego testu lub mniejszej grupy

zadań (testki) do „kalibrowania" (cechowania) skali wyników. To zagadnienie pojawiło się obecnie, w drugim roku obowiązkowego sprawdzianu w szkole podstawowej i egzaminu gimnazjalnego.

Przedtem nie mieliśmy w zasadzie (wyjątkiem są prace Instytutu Badań Kompetencji w Wałbrzychu) testowych wydawnictw cyklicznych, a większość pozycji wydawniczych o tytule „Testy... " nie zawierała nawet śladu standaryzacji. Obecnie chodzi o to, by sprawdziany i egzaminy przynosiły wyniki nadające się do porównywania osiągnięć **populacji krajowej** w kolejnych latach – zarówno ogólnie, jak i w poszczególnych obszarach wymagań egzaminacyjnych. Chcielibyśmy na przykład móc odnotować postęp w zakresie czytania i ewentualny brak postępu w zakresie wykorzystania wiedzy w praktyce. **Względne wyniki** (pozycje) osiągnięć uczniów w powiatach, gminach, szkołach i oddziałach, a także względne wyniki (pozycje) ucznia w kolejnych egzaminach, są łatwo porównywalne w **skali staninowej**, ale jeżeli staniny obliczymy co roku na nowo, to ich średnia zawsze wyniesie dość dokładnie 5,0 i nie dowiemy się niczego o postępie populacji.

Zrównywanie wyników kolejnych edycji testów jest działaniem rutynowym w USA (Angoff, 1971; Petersen i in., 1989). U nas wykonano dopiero pierwsze próby takiego podejścia (Niemierko, 2002b; 2003) i jest za wcześnie na to, by jego wynik podawać do publicznej wiadomości. Można tylko zauważyć, że wobec odstępstw rozkładów liczebności wyników egzaminów szkolnych od rozkładu normalnego żadne przekształcenia liniowe tu nie wystarczą i niezbędne jest zastosowanie **metody ekwicentylowej**, będącej zrównywaniem wyników poszczególnych wersji testu na podstawie ich rang centylowych. Ta metoda pozwala na ściślejszą korektę średniej, wariancji i kształtu rozkładu nowej wersji testu ze względu na parametry poprzedniej wersji. Dalsze próby obejmą wykorzystanie modeli probabilistycznych (IRT) do tego celu.

Normy ilościowe testów sprawdzających

Oprócz **standardów edukacyjnych (standardów wymagań egzaminacyjnych)**, rozumianych jako (werbalne) opisy oczekiwanych osiągnięć uczniów i stanowiących testowe **normy jakościowe**, w pomiarze sprawdzającym potrzebne są także **normy ilościowe** w postaci wybranych wyników surowych (*cutting scores*) lub procentu punktów możliwych do uzyskania, jakie będą podstawą decyzji o zaliczeniu egzaminu. Zagadnienie norm ilościowych należy do najbardziej wrażliwych społecznie, bo po ich ustanowieniu nawet jednopunktowe różnice wyników pomiaru mogą decydować (współdecydować) o wartości wieloletniej pracy ucznia.

Metody budowania norm ilościowych dzielą się na trzy grupy:

1. Oparte na analizie **treści** kształcenia (wymagane i niewymagane elementy).
2. Oparte na **obserwacji** (wynikach) wybranych uczniów.
3. Oparte na **świadomej decyzji zespołowej** (*informed judgement methods*).

Wykonano tysiące badań porównawczych w zakresie niearbitralnych metod (grupy 1 i 2) budowania norm. Wyniki tych badań były zarówno niezgodne wewnętrznie, jak i niepraktyczne, bo uzyskane normy były z reguły za wysokie (zob. np. Green i in., 2003). Całą nadzieję pokładamy w trzeciej grupie metod.

„Świadomość" decyzji zespołowej polega na udostępnieniu zespołowi „sędziów kompetentnych" informacji o prawdopodobnym **stosunku selekcji** (procencie uczniów nie zaliczających egzaminu, z ewentualnym warstwowaniem według płci, środowiska, lokalizacji szkoły, regionu kraju itp.) przy określonej normie. Dobierając zespół sędziów, trzeba pamiętać o wszystkich grupach **udziałowców** (*stakeholders*) podejmowanych decyzji: politykach, administracji szkolnej, nauczycielach, rodzicach i uczniach, pracownikach uczelni wyższych, dziennikarzach oświatowych (Crocker, 2002; Ryan, 2002). Możemy ich nazwać „ewaluatorami" lub, jak zaproponował Cronbach, „walidatorami" (co podkreśla związek budowania norm ilościowych z szeroko rozumianą trafnością pomiaru).

Doświadczenie wykazuje, że **każdy** zespół walidatorów egzaminu zaczyna od „myślenia życzeniowego" i proponuje nierealistycznie wysokie normy, a dopiero potem, po skonfrontowaniu ze wskaźnikami statystycznymi, stopniowo i z ociąganiem, obniża normy. Operację trzeba powtarzać wielokrotnie, by osiągnąć zadowalające, a właściwie – kompromisowe, rezultaty. Taki proces, nazywany **iteracyjnym**, jest pracochłonny i denerwujący dla obu stron: walidatorów, którzy muszą zmieniać stanowisko i „pomiarowców", których dane są często kwestionowane. W ubiegłym roku Edward Haertel (2002) zaproponował inne, niesprawdzone jeszcze w praktyce podejście, które nazwał **metodą instruktażową** (*„briefing book" method*), a proces jej zastosowania – **uczestniczącym** (*participatory*). W tym procesie sędziowie otrzymują:

- wykaz dziesięciu kolejnych **norm**, np. od 40% do 85% punktów w odstępach co 5%;
- **standardy wymagań egzaminacyjnych** zredukowane do tych części tekstu, które trafnie opisują osiągnięcia danego procentu uczniów;
- przykładowe zadania, które – według parametrów IRT, to jest według **krzywych charakterystycznych** – najlepiej różnicują uczniów na tych progach;
- opisy umiejętności mierzonych przez te zadania (wypisy z **kartoteki testu**);
- oszacowanie **procentu uczniów** niezaliczających egzaminu, ogółem i w wybranych warstwach;
- rozkład przewidywanych **osiągnięć szkół** w procentach uczniów zaliczających egzamin (od 100% do 0%).

Swoje pedagogiczne „credo" wykłada Haertel w pierwszym akapicie cytowanego artykułu:

> **Raportowanie wyników według standardów wymagań** (podkreślenie moje – B. N.) stało się powszechne w pomiarze dydaktycznym. Już nie wystarczy powiedzieć, czy wyniki testu są wysokie, czy niskie; język współczesnych reform edukacyjnych potrzebuje liczb określających proporcje uczniów o zadowalających i niezadowalających wynikach.

> Od raportu przedstawiającego wyniki według standardów oczekuje się innego rodzaju informacji niż porównania dokonywane według norm empirycznych. Jego natura polega na **osądzaniu wartości** [*value judgement*, podkreślenie moje – B. N.], często związanym z doniosłymi konsekwencjami dla uczniów i szkół. Od dawna wiemy, że wiele rodzajów osądzania wartości występuje we wszystkich etapach planowania testu, jego zastosowania i interpretacji, ale raporty według standardów stanowią potężne i bezpośrednie źródła informacji o spełnieniu wymagań, zaostrzając pytania o naturę tych sądów normatywnych, o ludzi za nie odpowiedzialnych oraz o to, czy należy i jak je brać pod uwagę przy ustalaniu trafności testu.

Sprawdzian po szkole podstawowej i egzamin po gimnazjum obywają się na razie bez norm ilościowych, co znaczy, że – wbrew nazwie – jedynie **różnicują** osiągnięcia uczniów, a nie **sprawdzają** ich i nie potwierdzają, że są one uzyskane przez poszczególnych uczniów. Jest jednak mało prawdopodobne, by egzamin maturalny mógł także obejść się bez normy ilościowej (to znaczy, by mógł być zaliczony nawet przy wyniku zerowym). Znaczy to, że trzeba będzie utworzyć **zespół sędziów kompetentnych do normowania matury**, by podzielić się odpowiedzialnością społeczną za decyzje o promowaniu jednych licealistów, a niepromowaniu innych. W ślad za maturą pójdą zapewne, prędzej czy później, egzaminy gimnazjalne. Trawestując znane powiedzenie, musimy zgodzić się, że „egzaminy doniosłe są zbyt ważne na to, by je zostawić pedagogom".

Post-totalitarna przeszłość kraju i, jako jej pochodna, **autorytarna koncepcja edukacji** mogą wydłużać proces uczenia się demokratycznego podejmowania decyzji i dzielenia się odpowiedzialnością. Wielu działaczom wydaje się, że na uczniach i na szkole można **wymusić postęp**, a lepsze jest utrzymanie „twarzy" (fasady) systemu przy pewnych przemilczeniach („wmiataniu śmieci pod dywan") niż nagłe odsłonięcie braków edukacji.

Czy tyka obecnie maturalna „bomba z opóźnionym zapłonem"? Mam nadzieję, że nie, bo procedury zapobiegania wybuchom są dość łatwo dostępne. Trzeba tylko pamiętać, by na czas podjąć następujące prace:

1. Śmiałe **zróżnicowanie** kilku (w przyszłości zapewne więcej niż dwu) poziomów matury i innych egzaminów, bowiem lepiej jest ograniczyć normy jakościowe i wprowadzić umiarkowanie wysokie normy ilościowe (np. 60–80% punktów) niż przeciwnie, zachować wysokie normy jakościowe przy żałośnie ubogiej normie ilościowej (np. 20–40% punktów).
2. Dokonanie dostatecznej liczby prób pilotażowych i przygotowanie danych instruktażowych, tak by zespoły sędziów kompetentnych mogły **świadomie** podjąć wiążącą decyzję o normie ilościowej i by mogły zaakceptować przewidywane liczby niepowodzeń egzaminacyjnych.
3. Powołanie przedmiotowych **zespołów sędziowskich**, przećwiczenie procedur stanowienia norm i podjęcie właściwych decyzji.

Konsekwencje egzaminów doniosłych

Dyskusja o ujemnych konsekwencjach testowych egzaminów pojawia się cyklicznie w wielu krajach. Jej nawrót w profesjonalnych czasopismach pomiarowych w USA wywołał starannie wyważony artykuł, którego autorem był Gregory Cizek (2001). Czytamy w nim:

> Mnożą się opowieści o rozpaczy doświadczanej przez solidnych uczniów, którym odmówiono dyplomu na podstawie doniosłego testu i przykłady na to, jak testy zawężają programy kształcenia, frustrują najlepszych nauczycieli, wywołują dławiący niepokój nawet u najbystrzejszych uczniów, a u małych dzieci powodują wymioty lub płacz lub jedno i drugie. Co wszakże wyróżnia obecne krytyki testów to intensywność, z jaką podnoszone są negatywne skutki testowania (prawdziwe i wydumane) i rzucający się w oczy brak jakiegokolwiek zainteresowania pozytywnymi skutkami (prawdziwymi czy wydumanymi) tych debat publicznych.

By uzupełnić ten brak, Cizek analizuje dziesięć następujących **pozytywnych konsekwencji** testowych egzaminów doniosłych:

1. Rozwój zawodowy nauczycieli.
2. Dostrzeganie uczniów o specjalnych potrzebach edukacyjnych.
3. Upowszechnianie wiedzy o pomiarze dydaktycznym.
4. Gromadzenie informacji o uczniach i szkołach.
5. Wykorzystywanie informacji o uczniach i szkołach.
6. Poszerzanie pola wyboru edukacji.
7. Rozliczanie szkół pod względem skuteczności kształcenia.
8. Zgłębianie treści i struktury przedmiotów szkolnych przez nauczycieli.
9. Podnoszenie jakości testów sprawdzających.
10. Podnoszenie jakości uczenia się.

Te konsekwencje nie zawsze są widoczne i nie zawsze bezdyskusyjnie pozytywne. Niewątpliwie jednak występują i zasługują, nie tylko według autora tej listy, na wsparcie.

Komunikacja i emocje

Teoria komunikacji uczy nas dostrzegać rolę emocji w porozumiewaniu się ludzi. Każdy akt komunikacji jest **wielotorowy**, co polega na powiązaniu informacji o **faktach** z informacją o **stosunku** nadawcy do odbiorcy oraz informacji **jawnej** z informacją **ukrytą** (Nęcki, 1996: 175). W warunkach szkolnych niemal każde zwrócenie się nauczyciela do ucznia niesie pewien ładunek informacji o aprobacie lub dezaprobacie jego sposobu uczenia się („lubienia" go lub „nielubienia" pod tym względem), a obok tekstu jawnego (słownego, bezpośrednio obserwowalnego), zawiera tekst ukryty (bezsłowny, emocjonalny, odbierany intuicyjnie przez analogię z wcześniejszym doświadczeniem).

Podstawą użytecznych strategii komunikacji interpersonalnej jest dostrzeganie różnic między ludźmi: wieku, płci, typu kultury, wykształcenia, doświadczenia, potrzeb,

zainteresowań, pozycji służbowej. Te różnice mają zasadnicze znaczenie dla porozumiewania się. Stephen Covey (2000: 73) stwierdza:

> Kluczem zaś do docenienia tych różnic jest zrozumienie, że wszyscy ludzie widzą świat nie takim, jakim jest, lecz takim, jacy są oni sami.

Uprzedzenia wielu dorosłych wobec szkoły, a w szczególności – wobec wymagań i egzaminów zasługują na poważne badania, zwłaszcza w kraju o tak autorytarnych tradycjach edukacyjnych jak Polska. Liczba rodziców, działaczy społecznych, dziennikarzy, a nawet profesjonalnych pedagogów, którzy źle wspominają szkołę – zbyt surową, zbyt zbiurokratyzowaną lub tylko niezdolną zaspokoić potrzeby młodzieży uzdolnionej – jest zapewne ogromna. „Złe emocje" odzywają się w ocenach procesów dydaktycznych, w dyskusjach, w decyzjach, w publikacjach popularnych i naukowych.

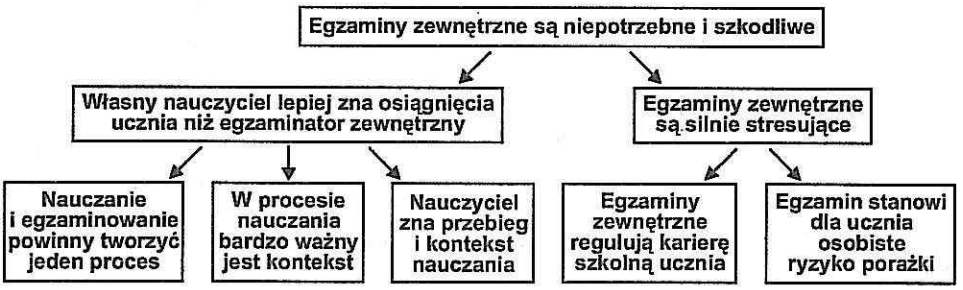
Do silnych uprzedzeń dołącza się nasza podatność na sensacje. Gdy „zadanie testowe pokona ucznia", przyjmujemy to ze smutnym zrozumieniem, ale gdy przeciwnie (jak to się zdarza nawet w najlepiej wystandaryzowanych testach) „uczeń pokona zadanie testowe" wykazując w nim błąd lub nieścisłość, zapanowuje ogólna radość – wcale nie z sukcesu ucznia, lecz z porażki konstruktora testu – i gazety piszą o tym tak, jakby człowiek pogryzł tygrysa.

Dystans między twórcami i użytkownikami testów jest stanowczo zbyt wielki. Gdyby był mniejszy, bylibyśmy skłonni martwić się, a nie cieszyć, błędami w sztuce pomiarowej, no i dokładać wszelkich starań, by na przyszłość błędom zapobiec.

Nieuczciwa strategia komunikacyjna

By dobrze komunikować się z kimkolwiek, musimy poznać się wzajemnie. Oponenci egzaminów doniosłych dysponują przecież, obok silnych emocji, także racjonalnym uzasadnieniem **niechęci do pomiaru dydaktycznego**. Ich pogląd „rozpakuję" metodą zaproponowaną przez Edwarda Haertela (1999), a potem dokonam własnej interpretacji wyniku.

Rysunek 1.
„Rozpakowanie" poglądu, że
„egzaminy zewnętrzne są niepotrzebne i szkodliwe"



Zauważmy, że uogólnienie zawarte w górnym wierszu rysunku, zapewne niezgodne z przekonaniem większości uczestników konferencji diagnostyki edukacyjnej, ma podstawę w dwu twierdzeniach środkowego wiersza, które możemy uznać za prawdziwe, a te wspierają się na przesłankach dolnego wiersza, które nie tylko są w pełni przekonywające, lecz także pedagogicznie fundamentalne. Zatem, jeżeli dedukcyjne „rozpakowanie" poglądu zamienić na jego **indukcyjne** uzasadnienie, to z prawdy (prawdziwych założeń) wynika fałsz (nieprawdziwe uogólnienie).

Gdzie zatem tkwi błąd logiczny? W **uproszczeniu**, to jest w ograniczeniu liczby przesłanek elementarnych. Gdyby w dolnym wierszu wykresu znalazły się takie przesłanki:

Nauczyciel nie jest w stanie wyrazić swej wiedzy o kontekście i wynikach nauczania w zobiektywizowanej skali	**Szkoła potrafi przygotować ucznia poznawczo i emocjonalnie do egzaminu zewnętrznego**

rozumowanie doprowadziłoby do ostrożniejszych wniosków.

Uproszczone rozwiązania złożonych zagadnień, nie tylko pedagogicznych, mogą doprowadzić do spustoszenia intelektualnego wśród odbiorców. Przypuśćmy, że utalentowany dziennikarz lub, co gorsza, literacko uzdolniony uczony (politolog, ekonomista, pedagog) przedstawia swoją „dedukcję z emocjonalnie mu bliskich poglądów" w eleganckiej, quasi-indukcyjnej formie. Odbiorcy są zachwyceni prostotą wywodu, który wciąga ich i rzekomo rozwiązuje problem, z jakim przedtem się borykali. Tak tworzy się grono entuzjastów lub może sekta wyznawców proroka, przekonanych o bezwzględnej racji pewnej tezy.

Nim zacytuję przykład działania takiej grupy, zajmę się jeszcze pewną „herezją egzaminacyjną", szerzącą się na całym bodaj świecie.

Granice odpowiedzialności szkoły

Poniższy przykład „rozpakowania" poglądu na temat przyczyn niezadowalających wyników egzaminów zewnętrznych jest dość bliski oryginalnemu rysunkowi Haertela (1999).

Rysunek 2.
„Rozpakowanie" poglądu, że
„wyniki egzaminów zewnętrznych są obrazem pracy szkoły"

Także i tym razem nietrudno jest wskazać przesłanki, które silnie warunkują prawdziwość naczelnej tezy zwolenników administracyjnych sankcji wobec szkół, których uczniowie uzyskują zbyt niskie wyniki egzaminów zewnętrznych:

Kandydaci przyjmowani do poszczególnych szkół różnią się średnim poziomem uzdolnień i przygotowania	Środowisko i rodzina ucznia mają większy wpływ na motywację uczenia się niż szkoła	Autorskie i twórcze wyniki kształcenia nie mieszczą się w zobiektywizowanym egzaminie zewnętrznym

Metoda „rozpakowywania poglądów" uzupełniona o poszukiwanie „pominiętych przesłanek" niechybnie ujawnia słabość wąskich, osobistych i środowiskowych „teorii", lansowanych – w dobrej lub nieco gorszej wierze – na różnych szczeblach kompetencji pedagogicznej.

Fanatyzm polityczno-edukacyjny

Przeciw złowrogim uproszczeniom poglądów w stosunku do egzaminów doniosłych wypowiedział się William Rich (2003), którego mrożącą krew w żyłach opowieść zacytuję w obszernym wyborze:

> [Gdy chodzi o konsekwencje testowania doniosłego, to] doniosła polityka *Wielkiego Skoku* w wykonaniu Przewodniczącego Mao [...] jest celnym przykładem.
>
> Przywództwo Przewodniczącego Mao ilustruje podejście wielkiej doniosłości w zarządzaniu uprawami, nie tylko z powodu klęski zaopatrzenia w żywność, lecz także z powodu wyniszczenia ptaków. Pod władzą Mao urzędnicy komunistyczni zapragnęli skopiować radzieckie metody rolne, które – z ideologicznego punktu widzenia – powinny były zapewnić obfitość produkcji rolnej. Część owego nieszczęsnego planu była oparta na spostrzeżeniu, że ptaki, zwłaszcza wróble, zjadały znaczną część zboża, nim mogło ono dojrzeć. Wysyłano więc wieśniaków na pola celem zabijania wróbli, by móc zbierać więcej ziarna i by rosła produkcja. We wioskach walono w garnki i patelnie tak długo, aż ptaki spadały z wyczerpania. Niestety, nieroztropni komuniści zniszczyli nie tylko ptaki, lecz także równowagę ekologiczną, która, między innymi, trzymała na wodzy owady. Gdy zabrakło ptaków, owady kompletnie zniszczyły plony, powodując wielki głód i stawiając ludzi w jeszcze gorszej sytuacji niż przed rozpoczęciem akcji zabijania ptaków.
>
> Co więcej, każdy kto poddał w wątpliwość skutki polityki Mao, był natychmiast wtrącany do więzienia. Ta praktyka zniechęcała każdego (uczonego, polityka lub wieśniaka) do głoszenia czegokolwiek, co mogłoby być odczytane jako krytyka polityki Przewodniczącego. Komuniści byli przekonani, że tacy reakcjoniści żerowali na dorobku ludu i dlatego nie byli godni nieść pochodni rewolucji (Rich, 2003: 33–34).

Rich dowodzi, że wskutek egzaminów zewnętrznych „czynności niedoniosłe" mogą zniknąć w szkole. W imię podniesienia jakości kształcenia „będzie walić się

w garnki i patelnie" tak długo, aż w kąt pójdą autorskie wizje programowe. Dziać to się będzie przy pełnej aprobacie szerokiej publiczności „nie dlatego, że lubi ona testy, lecz ze względu na prostotę informacji". Jakże łatwo jest zrozumieć szkołę według zasady „zastosuj test i poranguj szkoły od 1 do 10"! – ironizuje Rich (2003: 34).

Wiele cech, wiele miar

Wraz z rozbudową pojęcia trafności teoretycznej nową aktualność zyskują złożone metody analiz, zaproponowane jeszcze w ramach korelacyjnej kultury pomiarowej. Należy do nich **macierz wielu cech i wielu metod** (*multitrait-multimethod matrix*), zaproponowana przez Campbella i Fiskego (1959). Przedstawiłem ją w podręczniku *Testy osiągnięć szkolnych* (1975, rozdz. V) przed trzydziestu laty i od tego czasu nie słyszałem, by polscy pedagodzy się nią interesowali. Dopiero teraz dowiaduję się, że jeden z wydziałów Centralnej Komisji Egzaminacyjnej zbiera dane do takiej macierzy. Postępy teorii generalizacji i pełna komputeryzacja metod analizy wariancji powinny sprzyjać zamiarom wielostronnej analizy uczenia się i egzaminowania.

Wielkie pole badawcze rozpościera się przed zagadnieniem **motywacji egzaminacyjnej**. Dowiadujemy się otóż, że otwarte zadania tekstowe z matematyki stały się łatwiejsze dla ósmoklasistów od zadań wyboru wielokrotnego, gdy zdecydowano się płacić uczniom (niewielkie sumy) za prawidłowe rozwiązania (Lane, Stone, 2002: 25)! Wyższe wyniki egzaminów regulaminowych w stosunku do egzaminów pilotażowych i badań próbnych są zjawiskiem ogólnie znanym.

Złożoność **oceniania wewnątrzszkolnego** jest przedmiotem rosnącej liczby studiów empirycznych. Badania przeprowadzone w *Educational Testing Service* pokazują, że oszacowanie motywacji uczenia się i pilności w wykonywaniu prac domowych odgrywa w nim bardzo istotną rolę (Willingham i in., 2002: 15–20: McMillan, 2001: 23 i nast.). Podobne wyniki uzyskał James McMillan (2001), a Susan Brookhart (1991: 36) nazwała oceny nauczycielskie „luźną mieszanką opinii o postawie, wysiłku i osiągnięciach ucznia". W przypadku najsłabszych uczniów ich osiągnięcia poznawcze bywają mało znaczącym składnikiem tej mieszanki (Brookhart, 1993; Szyling, 2002).

Tym i podobnym (jak na przykład „dobre zachowanie") składnikom **oceny społeczno-wychowawczej** dziewczęta zawdzięczają przewagę na świadectwach szkolnych, podczas gdy egzaminy testowe stosunkowo lepiej zdają chłopcy. Według H. D. Hoovera, ta prawidłowość jest silnym argumentem za stosowaniem „kombinacji ocen ze szkoły średniej i wyników egzaminu testowego" wobec kandydatów na wyższe uczelnie (Hoover, 2003: 8).

Nowa federalna ustawa edukacyjna w Stanach Zjednoczonych, określona jako „Żadne dziecko nie zostaje w tyle" (*No Child Left Behind* – NCLB), zwróciła uwagę „pomiarowców" na diagnostykę edukacyjną. Hasłem dnia stało się sformułowane w tej ustawie żądanie, by stosować **wiele miar** (*multiple measures*) i nigdy nie podejmować

decyzji na podstawie jednego wyniku (Henderson-Montero i in., 2003: 7). Trzeba stosować różne źródła, metody i narzędzia, wykorzystywać różne okazje, badać różne populacje i poziomy osiągnięć, a także systematycznie powtarzać pomiar, by mieć obraz złożonej całości.

Syntetyzowanie wyników „wielu cech, wielu miar" jest podporządkowane analizie wybranego pojęcia, to jest dyskursowi **trafności teoretycznej** pomiaru. Może być dokonywane według trzech zasad (Schafer, 2003):

1. **koniunkcji**, gdy wymaga się od ucznia każdego rodzaju osiągnięć;
2. **kompensacji**, gdy wysokie osiągnięcia w jednym zakresie równoważą braki w innym zakresie;
3. **komplementarności**, gdy wystarczy jedno z dwu lub więcej osiągnięć.

Metodyka „wielu cech, wielu miar" jest jeszcze w powijakach, lecz trudno przecenić jej znaczenie dla diagnostyki edukacyjnej w USA i w innych krajach.

Odkupienie grzechów psychometrii

Pod takim właśnie tytułem James Popham (2003) zabrał głos w dyskusji o konsekwencjach testowych egzaminów doniosłych. Jego zdaniem, „grzechem śmiertelnym" edukacji jest „stosowanie podejścia różnicującego do ewaluacji jakości kształcenia". Doping, jakiemu podlegają szkoły, powoduje, że „ciężkie grzechy dydaktyczne napotyka się nagminnie w każdej klasie, która znajdzie się pod wpływem tradycyjnie skonstruowanych testów osiągnięć". Pojawia się w niej bowiem „redukcjonizm programowy" i „znika radość uczenia się na rzecz ćwiczeń w typie musztry".

„Grzech złego uczynku" w postaci kłamstwa, rabunku lub zbrodni jest, zdaniem Pophama, dość łatwy do wykrycia. Znacznie trudniej jest stwierdzić „grzech zaniedbania", nawet największego, jakim jest dostarczanie zredukowanej edukacji większości uczniom w kraju.

Za grzechy tak ciężkie nie wystarczy pokuta w postaci odmawiania gorących modlitw. Wszyscy powinniśmy **szerzyć kulturę oceniania** (*promote increased assessment literacy*) wśród polityków oświatowych, nauczycieli, publiczności, a zwłaszcza rodziców dzieci w wieku szkolnym. Służyć temu będą następujące działania:

1. Pisanie artykułów o pomiarze dydaktycznym do gazet.
2. Pisanie rozpraw o ocenianiu szkolnym do poważnych czasopism.
3. Ustne prelekcje dla różnorodnego audytorium, szkolnego i rodzicielskiego.
4. Wyszukiwanie właściwych zasobów: książek, artykułów, taśm, dysków itp.
5. Zakładanie stowarzyszeń wspierających pomiar dydaktyczny.
6. Pobudzanie aktywności komitetów rodzicielskich w zakresie kultury oceniania.
7. Znajdowanie funduszy na promocję kultury oceniania.
8. Organizowanie konferencji rodziców i działaczy oświatowych na temat oceniania.

Nic dodać, nic ująć! Taki zakres działań odpowiada, jak sądzę, uczestnikom naszej konferencji. Przewidywana dyskusja nad krakowską koncepcją Stowarzyszenia Diagnostyki Edukacyjnej obejmie główne wątki powyższego wykazu.

Jim Popham nie byłby jednak sobą, gdyby na koniec nie postraszył opornych wobec swojego programu pokuty najstraszniejszą z kar, a mianowicie – skazaniem delikwenta na syzyfową pracę **wiecznego** wyjaśniania osobom postronnym „natury rozszerzonej trafności teoretycznej" pomiaru dydaktycznego...

Podsumowanie

Założenia i ogólne prawidłowości

1. Pedagogika jest nauką praktyczną, ułatwiającą osiąganie celów edukacyjnych.
2. Przeważają w niej badania stosowane, zorientowane na usprawnienia procesów.
3. W kształceniu występuje aspekt emocjonalny, przenikający także diagnostykę edukacyjną.
4. Inaczej niż psycholog, nauczyciel zawsze ujawnia wyniki dokonywanego pomiaru.
5. Kolejne „rewolucje" teoretyczne podnoszą swoistość pomiaru dydaktycznego.
6. „Rewolucje" wywołują silny opór wśród ekspertów teorii klasycznych.

Postęp w teorii i w praktyce

1. Taksonomia celów kształcenia, pomiar sprawdzający i ewaluacja kształtująca mają wciąż więcej zwolenników wśród praktyków niż wśród teoretyków pomiaru.
2. Rozszerzone pojęcie trafności zobowiązuje do przewidywania konsekwencji pomiaru.
3. Trafność konsekwencyjna ma wielu zwolenników i sporo zajadłych przeciwników.
4. Mimo restrykcyjnych założeń matematycznych teoria pomiaru probabilistycznego (IRT) budzi nadzieje dydaktyków na użyteczne zastosowania.
5. Wyodrębnienie testek może ułatwić analizę wyników sprawdzianów i egzaminów.
6. Teoria generalizacji wiąże zagadnienia rzetelności i trafności pomiaru.
7. System egzaminów szkolnych w Polsce wymaga zrównywania ich kolejnych wersji.
8. Wyzwaniem pomiaru będzie normowanie wymagań matury i innych egzaminów.

Akceptacja społeczna egzaminu doniosłego

1. W dyskusjach o egzaminach doniosłych argumenty negatywne są bardziej wyraziste.
2. Podłożem emocjonalnym tej argumentacji bywają wspomnienia szkolne dyskutanta.

3. Każda teza da się uzasadnić indukcją z wyselekcjonowanych przesłanek.
4. Popularne poglądy na temat roli egzaminów zewnętrznych i odpowiedzialności szkoły za ich wyniki są na ogół szkodliwie uproszczone.
5. Należy przeciwdziałać obciążaniu szkoły całą odpowiedzialnością za wyniki egzaminów.
6. Jednostronność prowadzi do zgubnych skutków: i w polityce, i w edukacji.
7. Diagnostyka rozwija się pod hasłem „wielu cech, wielu miar".
8. Tylko surowa pokuta może zbawić zwolenników pomiaru różnicującego w edukacji!

Bibliografia

Angoff W. H. (1971), *Scales, Norms, and Equivalent Scores*, [w:] R. L. Thorndike (red.), *Educational Measurement. Second Edition*, ACE, Washington.

Bloom B. S. (red.) (1956), *Taxonomy of Educational Objectives. The Classification of Educational Goals*, Handbook I: *Cognitive Domain*, McKay, New York.

Brennan R. L. (2001a), *An Essay on the History and Future of Reliability from the Perspective of Replications*, „Journal of Educational Measurement", nr 4.

Brennan R. L. (2001b), *Generalizability Theory*, Springer-Verlag, New York.

Brookhart S. M. (1991), *Grading Practices and Validity*, „Educational Measurement: Issues and Practice", nr 1.

Brookhart S. M. (1993), *Teacher's Grading Practices: Meaning and Values*, „Journal of Educational Measurement", nr 2.

Campbell D. T., Fiske D. W. (1959), *Convergent and Discriminant Validation by the Multitrait-Multimehod Matrix*, „Psychological Bulletin", s. 81–105.

Cizek G. J. (2001), *More Unintended Consequences of High-Stakes Testing*, „Educational Measurement: Issues and Practice", nr 4.

Covey S. (2000), *Komunikacja synergiczna*, [w:] J. Stewart (red.), *Mosty zamiast murów. O komunikowaniu się między ludźmi*, PWN, Warszawa.

Crocker L. (2002), *Stakeholders in Comprehensive Validation of Standards-Based Assessments: A Commentary*, „Educational Measurement: Issues and Practice", nr 1.

Cronbach L. J. (1971), *Test Validation*, [w:] R. L. Thorndike (red.), *Educational Measurement. Second Edition*, ACE, Washington.

Cronbach L. J., Meehl P. E. (1955), *Construct Validity in Psychological Tests*, „Psychological Bulletin", s. 281–302.

Glaser R. (1963), *Instructional Technology and the Measurement of Learning Outcomes*, „Educational Psychologist", s. 519–521.

Green D. R., Trimble C. S., Lewis D. M. (2003), *Interpreting the Results of Three Different Standard-Setting Procedures*, „Educational Measurement: Issues and Practice", nr 1.

Haertel E. H. (1999), *Validity Arguments fo High-Stakes Testing: In search for Evidence*, „Educational Measurement: Issues and Practice", nr 4.

Haertel E. H. (2002), *Standard Setting as a Participatory Process: Implications for Validation of Standard-Based Accountability Programs*, „Educational Measurement: Issues and Practice", nr 1.

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

Henderson-Montero D., Julian M. W., Yen W. M. (2003), *Multiple Measures: Alternative Design and Analysis Model*, „Educational Measurement: Issues and Practice", nr 2.

Hoover H. D. (2003), *Some Common Misconceptions about Tests and Testing*, „Educational Measurement: Issues and Practice", nr 1.

Kane M. T. (2001), *Current Concerns in Validity Theory*, „Educational Measurement: Issues and Practice", nr 4.

Kryniewski M. (2002), *Zastosowanie probabilistycznej teorii wyniku zadania testowego do analizy testów*, [w:] B. Niemierko (red.), *Ocenianie szkolne, ekonomika i polityka oświatowa, probabilistyczne modele pomiaru*, skrypt nr 3 dla uczestników III Podyplomowego Studium Ewaluacji Dydaktycznej na Uniwersytecie Gdańskim, Międzywydziałowe Studium Pedagogiczne UG, Gdańsk.

Lane S., Stone C. A. (2002), *Strategies for Examining the Consequences of Assessment and Accountability Programs*, „Educational Measurement: Issues and Practice", nr 1.

Lee G., Brennan R. L., Frisbie D. A. (2000), *Incorporating the Tetlet Concept in Test Score Analyses*, „Educational Measurement: Issues and Practice", nr 4.

Lindquist E. F. (1953), *Selecting Appropriate Score Scales for Tests*, [w:] *Proceedings of the 1952 Invitational Conference on Testing Programs*, ETS, Princeton.

Linn R. L. (1994), *Criterion-Referenced Measurement: A Valuable Perspective Clouded by Surplus Meaning*, „Educational Measurement: Issues and Practice", nr 4.

Lord F. M. (1980), *Applications of Item Responce Theory to Practical Testing Problems*, Erlbaum, Hillsdale.

McMillan J. H. (2001), *Secondary Teacher's classroom Assessment and Grading Practices*, „Educational Measurement: Issues and Practice", nr 1.

Mehrens W. A. (1997), *The Consequences of Consequential Validity*, „Educational Measurement: Issues and Practice", nr 2.

Messick S. (1989), *Validity*, [w:] R. L. Linn (red.), *Educational Measurement. Third Edition*, ACE, Washington.

Messick S. (1995a), *Validity of Psychological Assessment: Validation of Inferences From Persons' Responses and Performances as Scientific Inquiry Into Score Meaning*, „American Psychologist", s. 741–749.

Messick S. (1995b), *Standards of Validity and the Validity of Standards in Performance Assessment*, „Educational Measurement: Issues and Practice", nr 2.

Muszyński H. (1970), *Wstęp do metodologii pedagogiki*, PWN, Warszawa.

Nęcki Z. (1996), *Komunikacja międzyludzka*, Wydawnictwo PSB, Kraków.

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. (2002a), *Ocenianie szkolne bez tajemnic*, WSiP, Warszawa.

Niemierko B. (2002b), *Zrównywanie wyników egzaminu. Materiał seminaryjny przygotowany na spotkanie w Sopocie w dniu 10 grudnia 2002 r.*, maszynopis, Gdańsk.

Niemierko B. (2003), *Zrównywanie sprawdzianu 2003 do sprawdzianu 2002*, maszynopis, Gdańsk.

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

Petersen N. S., Kolen M. J., Hoover H. D. (1989), *Scaling, Norming, and Equating*, [w:] R. L. Linn (red.), *Educational Measurement. Third Edition*, ACE, New York.

Popham W. J. (1997), *Consequential Validity: Right Concern – Wrong Concept*, „Educational Measurement: Issues and Practice", nr 2.

Popham W. J. (2003), *Seeking Redemption for Our Psychometric Sins*, „Educational Measurement: Issues and Practice", nr 1.

Popham W. J., Husek T. (1969), *Implications of Criterion-referenced Measurement*, „Journal of Educational Measurement", s. 1–9.

Rich W. (2003), *Historical High-Stakes policies realting to Unintended Consequences of High-Stakes Testing. Response to Cizek*, „Educational Measurement: Issues and Practice", nr 1.

Robitaille D. F., Garden R. A. (1989), *The IEA Study of Mathematics II: Contexts and Outcomes of School Mathematics*, Oxford, Pergamon.

Ryan K. (2002), *Assessment Validation in the Context of High-Stakes Assessment*, „Educational Measurement: Issues and Practice", nr 1.

Schafer W. D. (2003), *A State Perspective on Multiple Measures in School Accountability*, „Educational Measurement: Issues and Practice", nr 2.

Scriven M. (1967), *The Methodology of Evaluation*, [w:] R. E. Stake (red.), *Perspectives of Curriculum Evaluation*, Monograph Series on Evaluation, nr 1, Rand McNally, Chicago.

Spearman C. (1904), *The Proof and Measurement of Association Between Two Things*, „American Journal of Psychology", s. 72–101.

Szaleniec H. (2002), *Probabilistyczne modele wyniku zadania testowego (Item Response Theory)* [w:] B. Niemierko (red.), *Ocenianie szkolne, ekonomika i polityka oświatowa, probabilistyczne modele pomiaru*, skrypt nr 3 dla uczestników III Podyplomowego Studium Ewaluacji Dydaktycznej na Uniwersytecie Gdańskim, Międzywydziałowe Studium Pedagogiczne UG, Gdańsk.

Szyling G. (2002), *„Idź w świat i radź sobie, jak potrafisz". Drugi układ wymagań programowych w oczach uczniów i nauczycieli*, [w:] B. Niemierko, J. Brzdąk (red.), *Dwa rodzaje oceniania szkolnego*, Katowice, s. 73–86.

Wainer H., Lewis C. (1990), *Toward a Psychometrics for Testlets*, „Journal of Educational Measurement", s. 185–201.

Willingham W. W., Pollack J. M., Lewis C. (2002), *Grades and Test Scores: Accounting for Observed Differences*, „Journal of Educational Measurement", nr 1.