prof. dr hab. Bolesław Niemierko
Akademia Marynarki Wojennej w Gdyni

Jawne i ukryte błędy pomiaru dydaktycznego

Błądzić jest rzeczą ludzką
Seneka Starszy

Pedagogika jest niechętna teorii błędu, gdyż błąd pomiaru wydaje się godzić w autorytet nauczyciela. Tolerowany bywa ogólnopopulacyjny błąd systematyczny, wynikający z trudności diagnozy istotnych właściwości ucznia, ale nie błąd losowy, narażający go na przypadkowe decyzje edukacyjne. Silnie piętnowany jest błąd lokalny, mający źródło w zmiennych warunkach egzaminu, zróżnicowanym warsztacie punktujących zadania i dowolnej interpretacji wyniku, a więc w czynnikach zależnych od autorów i administratorów egzaminu.

Autor referatu jest zdania, iż pomiarowe kompetencje pedagogów można podnieść przez popularyzację **pomiaru różnicującego** w ujęciu klasycznej teorii testu, zaniedbanego na rzecz, naturalnego dla edukacji i dominującego od ćwierćwiecza, **pomiaru sprawdzającego**, opartego wyłącznie na standardach programowych.

Odświeżanie świadomości błędu pomiaru dydaktycznego oraz umiejętności jego szacowania warto zacząć od korekty formuły współczynnika rzetelności **alfa**, tak by odpowiadała powszechnej w naszym systemie egzaminacyjnym praktyce budowania arkuszy egzaminacyjnych z dwu rodzajów zadań: zamkniętych i otwartych. **Współczynnik Raju** uwzględnia zmienną długość skali punktowania zadań, co pozwala na dokładniejszą analizę wewnętrznej zgodności wyników egzaminów zewnętrznych w Polsce.

W niedawno zakończonych Mistrzostwach Europy w Piłce Nożnej angielski obrońca wybił z wnętrza swojej bramki piłkę, ale sędzia tego nie widział i nie uznał prawidłowo zdobytego przez Ukraińców gola, mimo że, jak wkrótce dowiodły analizy filmowe, piłka całym obwodem przekroczyła linię bramki. Komisja sędziowska Mistrzostw wyraziła ubolewanie z powodu pomyłki sędziego, ale decyzji nie zmieniła.

Dlaczego UEFA wciąż wzbrania się przed elektroniczną kontrolą położenia piłki, choć zapewne będzie musiała wkrótce, wzorem paru innych dyscyplin sportowych, w końcu na nią się zgodzić? Chodzi o **władzę sędziego**, który musi na boisku panować nad rozgrzanymi, rozgorączkowanymi, pełnymi „sportowej złości" zawodnikami i, by nie zwalniać tempa gry, szybko podejmować trudne, a przy tym stanowcze decyzje.

Prawna ochrona **nauczycielskiej oceny szkolnej** ma podobne podłoże. Gdyby możliwe były uczniowskie i rodzicielskie odwołania od tej oceny, kierowanie wychowawcze w klasie szkolnej, i dzisiaj przecież niełatwe, byłoby mocno utrudnione.

Założenie jednoinstancyjności oceny szkolnej przenosi się na egzaminy zewnętrzne. Przyznanie się organizatorów do niepewności pomiarowej byłoby dla laika przekreśleniem egzaminu, a wielkość **błędu losowego**, szacowanego w procedurach oceny rzetelności, poraża nawet obytych pomiarowo pedagogów. Czy lepiej jest wobec tego pomijać tę informację, zachowywać ją tylko dla wtajemniczonych? Uważam, że na dłuższą metę nie lepiej.

Znaczenie kontroli błędu

W dyscyplinach praktycznych, czy też, jak chciał Heliodor Muszyński (1970), „teoretyczno-praktycznych", świadomość błędu diagnozy i prognozy **odróżnia profesjonalistę od amatora**.

Lekarze zdają sobie sprawę z zawodności leków i zabiegów, a co więcej - ze skutków ubocznych ich zastosowania. W poważniejszych przypadkach szanse wyleczenia są określane przez prawdopodobieństwo, nierzadko stosunkowo niskie. Psychologowie nie pozwalają stosować diagnozy i terapii psychologicznej niespecjalistom, a wydawnictwa Pracowni Testów Psychologicznych Polskiego Towarzystwa Psychologicznego są zaopatrywane w krytyczne dyskusje trafności, normy empiryczne, oszacowania błędu pomiaru oraz interpretacje otrzymanego wyniku za pomocą statystycznych przedziałów ufności. Natomiast testowe wydawnictwa dydaktyczne, w które nasz rynek obfituje, są pozbawione nie tylko norm empirycznych, lecz także jakiejkolwiek informacji o błędach pomiaru.

A może ratują nas przynajmniej podręczniki metodologii badań pedagogicznych? Tu zawód jest szczególnie przykry. Oto w niedawno wydanej *Metodologii badań nad edukacją* Krzysztofa Rubachy (2008) znajdziemy (s. 198-199) następującą procedurę:

a. Na podstawie współczynnika rzetelności pomiaru testem poczucia skuteczności (TPS), oszacowanego na 0,88, błąd standardowy pomiaru oszacowano na 0,69 **stena**[1].
b. Ten błąd pomnożono przez wartość $z_{0,01}$ wynoszącą ±2,58, otrzymując oszacowanie półprzedziału ufności jako ±1,78 **stena**.
c. Tę wielkość wykorzystano do zbudowania przedziału ufności na skali wyników surowych testu TPS wokół wyniku średniego, wynoszącego 33 **punkty**. Czytamy (s. 199): „Po wykonaniu obliczeń dolny przedział ufności[2] wynosi 31, a górny 35. Oznacza to, że przedział ufności <31; 35> zawiera wartość wyniku prawdziwego z prawdopodobieństwem 99%".

Komentarz: Między krokiem *b* a krokiem *c* powyższej procedury nastąpiło przejście (powrót) od skali stenowej do skali surowych wyników testowania, czego autor podręcznika nie odnotował. Odchylenie standardowe stenów wynosi (z definicji) 2,0, ale odchylenie standardowe wyników surowych wynosiło

[1] Skala stenowa, stosowana głównie przez psychologów, ma 10 przedziałów. Jej odchylenie standardowe, podobnie jak w skali staninowej, wynosi 2 steny, ale średnia wynosi 5,5 stena.

[2] To bardzo dziwne wyrażenie. Powinno być: „dolna granica przedziału" lub „dolny kres przedziału".

9,5 punktu (s. 79)[3]. Zatem błąd standardowy pomiaru i wybrany przedział ufności są **niemal pięciokrotnie większe** w punktach niż w stenach.

Prawidłowo oszacowane granice przedziału ufności 99% dla wyniku 33 punkty w TPS wynoszą 24,5 punktu (dolna) i 41,5 punktu (górna). Popełniony przez autora błąd oszacowania wielkości przedziału ufności wynosi aż 475 procent! Taki błąd z pewnością zauważy każdy badacz jako tako doświadczony w pomiarze, bo tak mały przedział ufności możliwy byłby do uzyskania w teście o współczynniku rzetelności powyżej 0,99, co nie zdarza się w praktyce nauk społecznych. Jak mógł być popełniony przez autora podręcznika i przeoczony przez jego recenzentów?

Podany przykład można zinterpretować dwojako:

1. Jako dowód niewielkiego obycia pomiarowego wśród pedagogów, nie wyłączając autorów podręczników metodologii tej dyscypliny.
2. Jako dowód małej precyzji pomiaru psychologicznego i dydaktycznego.

Jeżeli przedział ufności 99% dla badanego uzyskującego średni wynik w teście cytowanym przez Rubachę lub w innym teście o podobnej rzetelności ma granice 24,5 oraz 41,5 punktu, to w rozkładzie normalnym przy odchyleniu standardowym wynoszącym 9,5 punktu obejmuje środkowe 62% badanych, a więc tylko po około 14% wyników wysokich i niskich **istotnie różni się od średniej**. Znaczy to, że z tak wysoką pewnością możemy sklasyfikować badanych pod względem poczucia skuteczności działania tylko w trzech grupach: wyników wysokich (14%) o średniej około 7,5 stena, przeciętnych (62%) o średniej około 5,5 stena i niskich (14%) o średniej około 2,5 stena. Zarówno powyższe, jak i bardziej szczegółowe, obejmujące więcej kategorii klasyfikacje obarczone są ryzykiem, o którego wielkości (istotności) użytkownik wyniku powinien być uczciwie poinformowany[4].

Tolerancja błędów

Błędu pomiaru nie pozbędziemy się nigdy nie tylko w naukach społecznych, ale w nich właśnie zagadnienie **tolerancji błędu**, jako dopuszczalnej rozbieżności między stanem faktycznym a stanem zarejestrowanym, urosło do rangi racji bytu metod ilościowych. To zagadnienie przedstawię według koncepcji Michaela Kane'a, znanego nam jako twórcę procedur **dyskursu trafności**, w którym sprawdza się, krok po kroku, „trafność argumentacyjną" pomiaru, odpierając kolejne możliwe zarzuty (Kane, 2007; Niemierko, 2009, s. 148-149).

W artykule, którego tytuł mogę przetłumaczyć jako „Błędy naszych szlaków" (2011), Kane wychodzi z założenia, że **błędy są skutkiem uproszczeń, jakich dokonujemy, budując modele rzeczywistości**, zwłaszcza tak skomplikowanej

[3] Autor dodaje do tej informacji, że wynosi ono „po zaokrągleniu 9, bo nie mamy wyników surowych co 0,5". Tę operację bardzo trudno jest zrozumieć statystykowi.

[4] Tym zagadnieniem zajmowano się w początkowych latach rozwoju teorii pomiaru sprawdzającego. Na przykład Reiner Fricke (1974) ustalił, że średnio rzetelny test musiałby zawierać co najmniej 20 zadań punktowanych 0-1, by rozkłady błędu 5% trzech ocen nie zachodziły na siebie wzajem, a co najmniej 80 zadań, by uzyskać sześć rozłącznych ocen. W Polsce podobne rozumowanie przedstawili Ignacy Stępniowski i Wojciech Małecki (1974) i doszli do wniosku, że w przypadku testu złożonego z 30 zadań wyboru wielokrotnego maksymalna liczba stopni szkolnych wynosi cztery.

jak zachowania ludzkie. Gdyby jednak modele nie były dostatecznie proste, nie nadawałyby się do zastosowań praktycznych. Nauczyciel i trener sportowy, psycholog i lekarz, a nawet meteorolog i geolog - nie są w stanie zinterpretować wszystkich dostępnych danych, więc uciekają się do oszacowań stanu zjawisk i do mało dokładnych prognoz. Żaden z nich nie może uważać się za nieomylnego, nie narażając się na kompromitację w swoim środowisku.

Jaka niedokładność oszacowań jest dopuszczalna? To zależy od rodzaju podejmowanych decyzji: porad, prognoz, przydziału dóbr i innych. Kane wprowadza pojęcie **tolerancji błędu** (*tolerance for errors*), rozumiane jako „wielkość, przy której błędy zaczynają kolidować (*interfere*) z zamierzonymi interpretacjami i zastosowaniami wyników pomiaru" (s. 19). Można więc zauważyć, że tolerancja błędu ma istotne **znaczenie psychologiczne**, zależy od kompetencji użytkownika. Specjalista, który ma świadomość niepewności wyniku i ostrzega przed jego absolutyzowaniem, wykazuje niewątpliwie większą tolerancję błędu niż laik, który wierzy w sytuację zasadniczo „czarno-białą" i skłonny jest traktować wszelką niejednoznaczność oceny jako niefachowość, nieszczerość i manipulację.

Przypuśćmy, że dane zamieszczone w podręczniku Rubachy dotyczą nie poczucia skuteczności działania, lecz przedmiotu szkolnego, np. matematyki, a naszą intencją jest dokonanie zróżnicowania osiągnięć uczniów za pomocą skali stopni szkolnych. Łącząc po dwa kolejne steny z wyjątkiem krańcowych, pozostawionych pojedynczo, otrzymujemy sześć przedziałów skali, którym możemy nadać nazwy stopni szkolnych. Otrzymujemy po około 2% ocen „celujących" i „niedostatecznych", po około 14% ocen „bardzo dobrych" i „dopuszczających" i po około 24% ocen „dobrych" i „dostatecznych"[5].

Ta operacja nie ma sensu w świetle założeń **pomiaru sprawdzającego** wielostopniowego, w którym normy stopni szkolnych są wyznaczone wymaganiami programowymi. Jednak psychologowie, opierając się na teorii różnic indywidualnych, od dawna i niezmiennie (Pieter, 1969; M. Ledzińska i E. Czerniawska, 2011) zalecają krzywą normalną prawdopodobieństwa jako podstawę oceniania szkolnego. Stopnie szkolne mają tu tylko **znaczenie pozycyjne**, co jednak pozwala na zastosowanie standardowych **skal przedziałowych** i bardzo ułatwia kalkulacje statystyczne.

Przyjmijmy tolerancję błędu w następującej złagodzonej postaci: ocena osiągnięć ucznia na podstawie wyniku pomiaru nie powinna w 95% przypadków różnić się więcej niż o jeden stopień od **wyniku prawdziwego** tego pomiaru, niezależnego od czynników losowych[6]. Jaka musiałaby być rzetelność pomiaru, by można było mieć 95% pewności, że wynik prawdziwy osiągnięć ucznia mieści się w dwustenowym przedziale oceny? Ponieważ przedział punktowy oceny osiągnięć wynosi, jak założyliśmy, dwa steny, to jego granice są o jeden sten odległe od środka tego przedziału. Zakładając normalny rozkład błędów

[5] Patrz: B. Niemierko (1975), s. 148-149.

[6] Takie założenie jest naturalne i dość często przyjmowane w pomiarze dydaktycznym. By zwiększyć dopuszczalną liczbę kategorii (stopni), zakłada się też bardzo niski poziom ufności: p = 0,50 (Kolen i Brennan, 2004, s. 346-347), co radykalnie, około trzykrotne, wydłuża skale stosowane w klasyfikacji.

pomiaru, ustalamy oczekiwaną wielkość błędu standardowego na σ_e = 0,5 stopnia szkolnego. Stosując podstawowy wzór alternatywny na **współczynnik rzetelności** (Niemierko, 1975, s. 225), otrzymujemy:

$$r_{tt} = 1 - \frac{\sigma_e^2}{\sigma_t^2} = 1 - \frac{0{,}5^2}{1^2} = 0{,}75$$

Taka wielkość współczynnika jest osiągana nie tylko w standaryzowanych testach osiągnięć szkolnych, lecz także w ramach **egzaminu powszedniego**, stosowanego przez nauczycieli z zastosowaniem pojedynczych testów lub innych metod o zadowalającej rzetelności. Przekształcając odpowiednio podany wzór i podstawiając do niego typową wartość r_{tt} = 0,80, otrzymujemy σ_e= 0,45 stopnia i przedział ufności równy 97% powierzchni pod krzywą normalną.

Nauczyciele często stosują **rozwiniętą skalę stopni szkolnych**, zaopatrywaną w znaki **„+”** i „-”**,** co daje zwykle około dziesięciu odrębnych ocen. Przy tej liczbie kategorii, gdy r_{tt} = 0,80 i σ_e = 0,45 stopnia, a przy skali długiej na sześć odchyleń standardowych przedział ufności oceny ma mieć długość 0,6 σ_t będzie on odpowiadał około 81% powierzchni pod krzywą normalną, a więc szansom adekwatności oceny około 4 : 1.

Zapytajmy na koniec, jaka jest szansa adekwatności oceny, gdy zastosujemy stustopniową skalę ocen, jaka kiedyś dominowała w Stanach Zjednoczonych (Brookhart, 2004), a i teraz ma pewne analogie w postaci wyników surowych długich testów lub baterii testów oraz, ostatnio odżywających w naszej praktyce egzaminacyjnej, rang centylowych[7]. Przy tych samych założeniach i długości przedziału ufności 1/10 stena = 0,06 σ_t, będzie on odpowiadał około 10% powierzchni pod krzywą normalną, a więc szansa dokładnej zgodności oceny z wynikiem prawdziwym pomiaru wynosi około 1 : 10. Stąd tendencja, by do sprawozdań indywidualnych wprowadzać tak małe jednostki pełnymi dziesiątkami, co dotyczy skali akademickiej[8] i centylowej (decyle).

Konstrukcja testu a błąd losowy pomiaru

Dotychczas zakładaliśmy milcząco, że test (arkusz egzaminacyjny) składa się z pewnej liczby prostych (jednoczynnościowych) zadań o średniej trudności, co uprawdopodobniało rozkład normalny wyników pomiaru. W krajowej praktyce egzaminacyjnej przyjęto jednak - ze względu na różnorodność treści programów kształcenia - równoległe stosowanie **zadań zamkniętych**, z gotowymi odpowiedziami do wyboru, punktowanych z zasady 0-1, i **zadań otwartych**, wymagających budowania odpowiedzi, punktowanych zwykle skalą wielopunktową. Ponieważ te drugie znacznie silniej różnicowały badanych (Niemierko, 2007, s. 357), nabierały charakteru „rozrywającego” rozkład wyników z jednomodalnego na dwumodalny (Stożek, 2008). Ich wprowadzenie obniżało oszacowania współczynnika rzetelności pomiaru (α Cronbacha), gdyż niejednorodność części testu prowadząca do zróżnicowanego skalowania zadań **zakłóca wewnętrzną zgodność pomiaru**.

[7] Centyle tworzą skalę przedziałową liczebności uczniów, ale tylko skalę porządkową mierzonych osiągnięć, bowiem przedziały środkowe skali są dużo mniejsze niż krańcowe.

[8] B. Niemierko, 2007, s. 322.

Ten problem był dotychczas rozwiązywany przez wyodrębnianie w złożonych zadaniach otwartych poszczególnych **czynności** (kroków) rozwiązywania, punktowanych 0-1 lub niewiele dłuższą skalą. Potraktowanie tych czynności jako odrębnych zadań ułatwiało punktowanie i sztucznie, gdyż czynności w obrębie zadania były silnie wzajemnie zależne, podnosiło wartość współczynnika rzetelności. Wprowadzało jednak do oceniania osiągnięć uczniów **schematyzm**, bo wiele oryginalnych i poprawnych rozwiązań nie mieściło się w ustalonych modelach odpowiedzi (Chodnicki, 2012). Z tego powodu stosuje się obecnie **punktowanie holistyczne** takich zadań, stopniujące osiąganie celu czynność złożonej, a nie poprawność wykonania przewidzianych kroków. To wymaga zastosowania nowej, a przy tym bardziej poprawnej metodologicznie procedury szacowania zgodności wewnętrznej testu[9].

Dokonamy porównania dwu procedur, dawnej i nowej.

Współczynnik ***alfa* Cronbacha** zakładał ***zasadniczą równoległość wyników prawdziwych*** (*essential tau-equivalence*) części testu X, co znaczy, że te wyniki mogły różnić się systematycznie średnimi, ale nie wariancjami, i kowariancja zadań testu X mogła być podstawą szacowania kowariancji części (zadań) testu X z częściami (zadaniami) testu Y i w rezultacie - kowariancji części oraz rzetelności testu Y.

Współczynnik Raju zakłada tylko, że wyniki części testu są **współpochodne** (*congeneric*), co znaczy, że mogą różnić się systematycznie nie tylko średnimi, ale i wariancjami, a to pozwala na wykorzystanie kowariancji nierównoległych części (zadań) testu X jako podstawy oszacowania kowariancji z podobnie ustrukturowanymi częściami testu Y i w rezultacie - kowariancji oraz rzetelności testu Y.

Założeniem N.S. Raju (1970, 1977) była tożsamość różnic między wariancjami części testu z kwadratem różnic między długościami ich skal. **Względną wielkość części** (*lambda*) określamy jako:

$$\lambda_i = k_i / \Sigma k_i$$

gdzie k_i jest długością skali *i*-tej części testu, a $\Sigma k_i = 1$.

Współczynnik wewnętrznej zgodności testu złożonego z nierównoległych części (współczynnik Raju) ma postać:

$$_R\alpha = \frac{\sigma_x^2 - \sum \sigma_i^2}{(1 - \sum \lambda_i^2)\sigma_x^2},$$

gdzie R^{α} oznacza współczynnik Raju, σ_x^2 - wariancję wyników testu, $\Sigma\sigma_i^2$ - sumę wariancji części (zadań), $\Sigma\lambda_i^2$ - sumę kwadratów względnej wielkości części (zadań).

[9] Na tę konieczność zwrócił mi uwagę dr Henryk Szaleniec, prezes Polskiego Towarzystwa Diagnostyki Edukacyjnej, wskazując zarazem współczynnik Raju jako optymalny do oceny rzetelności testu złożonego z nierównoległych części.

Ponieważ $\frac{1}{1-\sum \lambda_i^2} \geq \frac{m}{m-1}$, gdzie *m* jest długością (liczbą zadań) testu, współczynnik R^{α} testu złożonego z nierównoległych części (zadań) ma wartość liczbową większą niż współczynnik α, oparty na założeniu zasadniczej równoległości wyników prawdziwych tych części. Kowariancja części testu X z częściami tak samo ustrukturowanego testu Y jest wyższa niż kowariancja między częściami testu X, obniżona nierównoległością tych części.

W przypadku sprawdzianu po szkole podstawowej o typowej budowie (21 zadań jednopunktowych, 2 zadania dwupunktowe, 1 zadanie pięciopunktowe, 1 zadanie dwunastopunktowe) i współczynniku α = 0,80, współczynnik R^{α} = 0,88, a więc wzrasta o 10%. To skutkuje redukcją *standardowego błędu pomiaru* z 3,6 punktu do 2,8 punktu i zmniejszeniem przedziału ufności 95% z 14 punktów do 11 punktów.

Zastosowanie podwójnego oszacowania do pięciu (*m* = 5) obszarów umiejętności[10] sprawdzianu z 2004 roku zmniejsza różnicę wyniku między dawną i nową procedurą do 5%, gdyż w tym przypadku α = 0,80, natomiast R^{α} = 0,84. Jak z tego wynika, różnice konstrukcyjne zadań oddziałują w sprawdzianie po szkole podstawowej dwukrotnie silniej niż różnice treściowe umiejętności w pięciu częściach.

Błąd systematyczny i lokalny

Niestety, pomiar w naukach społecznych cechuje się nie tylko brakiem dokładności -ograniczoną rzetelnością skutkującą dużym błędem losowym - lecz także **błędem systematycznym**, powtarzającym się w każdym zastosowaniu procedur i narzędzi lub w ich zastosowaniu w pewnych warunkach lub z udziałem niektórych osób. Ten ostatni, występujący sporadycznie, będę nazywał **błędem lokalnym**. Jego przezwyciężanie pod hasłem „porównywalności wyników" było koronnym argumentem autorów systemu egzaminacyjnego w Polsce.

Przeglądu błędów pomiaru dydaktycznego dokonam według **taksonomii niezbędnych właściwości sprawdzania**, która obejmuje pięć właściwości uporządkowanych od najniższej, najbardziej elementarnej, do najwyższej, najtrudniejszej do uzyskania (Niemierko, 2007, s.275-282):

1. bezstronność sytuacji sprawdzania,
2. dokładność punktowania wyników sprawdzania,
3. rzetelność sprawdzania,
4. trafność sprawdzania,
5. obiektywizm sprawdzania.

Bezstronność sytuacji sprawdzania polega na stwarzaniu wszystkim uczniom jednakowo sprzyjających warunków wykonania czynności przewidzianych do opanowania. Te warunki mają charakter zarówno materialny (pomieszczenie, wyposażenie, oświetlenie, wyciszenie), jak i metodyczny (jasność celów, określony materiał, właściwa instrukcja, odpowiednia atmosfera, zachęta do pracy).

[10] (1) czytanie, (2) pisanie, (3) rozumowanie, (4) korzystanie z informacji, (5) wykorzystanie wiedzy w praktyce.

Najczęstsze przypadki naruszania bezstronności sytuacji sprawdzania to:

1. **błąd kulturowy** treści sprawdzania, polegający na dobieraniu zadań trudniejszych (tematycznie, językowo) dla pewnych grup uczniów (Hornowska, 2001),
2. nieprzyjazna atmosfera, upośledzająca neurotyków (Mietzel, 2002, s. 413-421),
3. nierówne traktowanie wybranych uczniów,
4. różne formy **oszustwa egzaminacyjnego**, skrytego posługiwania się informacją uzyskaną z innych źródeł niż własna wiedza i praca (Cizek, 1999).

Stronniczość sytuacji egzaminacyjnej ma zawsze charakter lokalny. Centralna Komisja Egzaminacyjna dokłada starań, by ją wyeliminować, co nie jest łatwe, bo kompensacja dysleksji lub niepełnosprawności może wchodzić w kolizję z równym traktowaniem uczniów[11], a zwalczanie oszustwa może prowadzić do zaostrzania sytuacji sprawdzania i pogarszania atmosfery. Na tym polu jednak nie mamy ani znaczących zaniedbań teoretycznych, ani masowego ukrywania naruszeń regulaminu.

Dokładność punktowania wyników sprawdzania, jako stosowanie pomocniczej charakterystyki liczbowej wyników, jest także w Polsce centralnie regulowana, a nadto wzmacniana intensywnym kształceniem egzaminatorów. W przypadku złożonych zadań otwartych budzi jednak zasadnicze kontrowersje, bo **schematy punktowania**, zawierające uznaniowe skale jakości odpowiedzi, są często traktowane jako **klucze punktowania**, ściśle wyznaczające punkty za określone odpowiedzi i nadające się głównie do zadań zamkniętych.

Schematy punktowania nie zapewniają zgodności między sędziami. Wykonano ogromną liczbę badań nad zgodnością między punktującymi wypracowania uczniów. Okazało się w nich, że nawet wysoko wykwalifikowany zespół egzaminatorów uzyskuje co najwyżej umiarkowaną zgodność ocen (Niemierko, 1999, s. 201-203). Punktujący zadania, zwani też zasadnie **sędziami**, różnią się (a) surowością ocen, (b) rozrzutem ocen i (c) uporządkowaniem ocen, to jest związkiem (korelacją) z ocenami innych sędziów. Jest wiele prawidłowości psychologicznych wyjaśniających te rozbieżności (Tyszka, 1999).

Ponieważ dokładność punktowania silnie zależy od osobowości sędziego, ma charakter **lokalny**, ale założenie różnorodności (dywergencyjności) rozwiązań i błędy konstrukcyjne zadań (wieloznaczność) mogą prowadzić do błędów **systematycznych**. Opinia publiczna w kraju jest zbyt skąpo informowano o tego rodzaju kosztach pomiaru twórczości uczniów.

Rzetelność pomiaru, jako niezależność wyniku od wybranych czynników kontekstowych - zwykle wersji testu, często upływu czasu (pomiar psychologiczny) i doboru sędziów (pomiar dydaktyczny), niekiedy formy zdań, obycia testowego badanych i organizacji testowania - była rozważana w poprzednich częściach tego studium. **Rzetelność pomiaru jest oszacowaniem niezależności wyniku pomiaru od błędu losowego**, opartym na matematycznie ścisłych założeniach i na rachunku prawdopodobieństwa. Najszerzej ujmuje ją **teoria uniwersalizacji**

[11] Zapobiegliwi rodzice mogą usilne starać się o uznanie ich dzieci za obciążone dysleksją rozwojową, co pozwoliłoby im skorzystać z wydłużonego czasu egzaminowania.

(generalizability theory), uwzględniająca wszystkie ujawnione źródła zmienności wyników pomiaru (Brennan, 1983), co skutkuje wydatnym zwiększeniem błędu standardowego („generycznego") w stosunku do oszacowań dokonywanych na podstawie wewnętrznej zgodności wyników (współczynnika α).

Największe i najmniej eksponowane w sprawozdanych egzaminacyjnych są błędy spowodowane niezadowalającą trafnością i brakiem obiektywizmu w interpretacji wyników pomiaru.

Trafność sprawdzania jest, najogólniej biorąc, użytecznością jego wyników. Tę użyteczność rozumiemy w dydaktyce bardzo szeroko, rozciągając ją na teraźniejszość i przyszłość ucznia i społeczeństwa. Według Samuela Messicka, którego analizy trafności teoretycznej pomiaru są od dwu dekad uważane za wzorcowe, trafność pomiaru jest „zintegrowanym sądem wartościującym na temat stopnia, w jakim dane empiryczne i uzasadnienia teoretyczne stanowią o tym, że wnioski i działania oparte na wynikach testowania i innych sposobach oceniania są odpowiednie (*adequate*) i właściwe (*appropriate*)" (1989, s. 13).

Ponieważ edukację traktujemy rozwojowo (Kohlberg i Mayer, 1993), wnioski z egzaminu powinny dotyczyć głównie **umiejętności uczenia się**, rozumianej jako ogólna zdolność do opanowywania nowych czynności, uzyskiwana przez dobrze zorganizowane ćwiczenie podobnych czynności. **Samoregulacja uczenia się**, obejmująca procesu emocjonalne i poznawcze, jest szczególnie złożoną umiejętnością (Ledzińska i Czerniawska, 2011, s. 108n), wymagającą długookresowej diagnozy.

Egzaminy zewnętrzne w Polsce są silnie zorientowane na umiejętność uczenia się. Coraz szerzej obejmują różne formy pracy ze źródłami - z tekstem historycznym, literackim i użytkowym, z tabelami i wykresami, z opisem doświadczeń - co wywołuje protesty konserwatywnych dydaktyków przedmiotowych, którzy chcieliby raczej rejestrować wiadomości. Zdolność do systematycznego wysiłku nie może być jednak zmierzona jednorazową mobilizacją, więc reprezentacja umiejętności uczenia się jest w egzaminach zewnętrznych niepełna.

Błąd oceny kompetencji ucznia na podstawie rozwiązywania zadań przedmiotowych ma charakter **systematyczny**, bo dotyczy, w większym lub mniejszym stopniu, wszystkich egzaminowanych. Co więcej, ten błąd rośnie w miarę standaryzacji testu, bo obszar mierzonej kompetencji ulega zwężeniu. „Im bardziej standaryzujemy, tym silniej zmierzamy do zmniejszenia błędu losowego i zwiększenia błędu systematycznego" - ostrzega Kane (2011, s. 18), a „błędy systematyczne mogą być bardziej kłopotliwe niż błędy losowe" (tamże, s. 25).

Najważniejszym walorem sprawdzania i najwyższą kategorią taksonomii jego niezbędnych właściwości, jest **obiektywizm sprawdzania**, rozumiany jako dokładność, z jaką wyniki sprawdzania są wyznaczone wymaganiami programowymi. Gdy o uczniu uzyskującym dowolny wynik sprawdzania osiągnięć możemy powiedzieć, jakie czynności opanował, a jakich nie opanował, ten wynik ma określone **znaczenie treściowe** (*content meaning*).

Postulat znaczenia treściowego wyniku sprawdzania pojawił się w teorii pomiaru sprawdzającego już u jej zarania (Ebel, 1962). Wydaje się oczywisty, ale jest trudny do spełnienia. Nawet najstaranniej zaplanowany i skonstruowany test dopuszcza wyniki o niskim i zmiennym znaczeniu treściowym, gdy **norma ilościowa**, czyli procent punktów, jakie uczeń musi zdobyć podczas egzaminu, by go zaliczyć, jest niska. Zauważmy, że stosowana w Polsce norma 30% pozwala uczniowi pominąć ponad 2/3 zakresu programowego, w tym często kluczowe umiejętności. Można także zaliczyć egzamin na trzy różne sposoby, rozwiązując całkowicie różną 1/3 zadań, choć można dość dokładnie przewidzieć, które zadania są dla większości uczniów łatwiejsze, a które trudniejsze[12].

Błąd systematyczny wynikający z niskiej normy ilościowej jest bardzo duży. Egzamin, w którym norma jakościowa (wykaz treści) jest ambitna, a normę ilościową obniżono do 30% punktów, wykazuje **trafność fasadową** (*face validity*) - pozorowanie wymagań, które uspokaja opinię publiczną. To, że większość wymagań programowych (standardów egzaminacyjnych) jest przez większość uczniów niespełniona, uchodzi uwagi niespecjalistów pomiaru. Wytwarzane są złudzenia pozwalające zachować wyobrażenia o szerokim zakresie umiejętności i wiadomości uczniów.

Błąd fasadowy, jakim jest rozbieżność między planem narzędzia a wynikiem pomiaru uznawanym za zadowalający, może być oszacowany przez porównanie średnich osiągnięć uczniów w wybranych punktach skali, co pojawiało się w niektórych sprawozdaniach Centralnej Komisji Egzaminacyjnej. W tym zakresie jednak jest jeszcze wiele do naprawienia przez wykorzystanie psychometrycznych metod ustalania normy, zwłaszcza najbardziej nowoczesnych, jak **metoda zakładkowa** (*bookmark method*; Lewis i in, 1999), polegająca na przeglądzie katalogu zadań zawierającego ich parametry probabilistyczne (IRT), i **metoda instruktażowa** (*briefing method*; Haertel, 2002), polegająca na porównaniu standardów programowych, treści zadań i wyników zadań (Niemierko, 2009, s. 61-63). Te metody wymagają zaangażowania dużych grup ekspertów przedmiotowych, których poglądy stanowią o treści i o wielkości normy mającej reprezentować spełnienie wymagań programowych.

Systematyczny błąd fasadowy dotyczy przede wszystkim egzaminu maturalnego, w którym stosuje się jednolitą krajową normę ilościową. Gdy, jak w przypadku niektórych szkół średnich (gimnazjum, liceum i technikum), poszczególne placówki same ustalają punktowe progi rekrutacji kandydatów, błąd staje się **lokalny**, bo ryzyko przyjęcia kandydatów o niższym poziomie przygotowania niż oczekiwany staje się zróżnicowane między szkołami. Tak więc brak obiektywizmu sprawdzania, podobnie jak, wcześniej analizowany, brak obiektywizmu punktowania zadań, może deformować nam obraz osiągnięć uczniów na różnych szczeblach systemu oświatowego.

[12] Na rozróżnieniu poziomów wymagań i, w konsekwencji, trudności zadań opiera się model testu sprawdzającego wielostopniowego (*TSW*, Niemierko, 1990).

Informacja o błędach pomiaru jako nakaz etyczny

Zagadnienie błędów pomiaru dydaktycznego jest szeroko objęte diagnostyczno-edukacyjnymi **kodeksami etycznymi**, czyli zbiorami norm moralnych i zasad postępowania w psychologii i pedagogice, dyscyplinach traktowanych łącznie. Te przepisy, po raz pierwszy opublikowane w Stanach Zjednoczonych w 1954 roku, są stale rozwijane i ulepszane. Z ich aktualnej wersji (*Standards…*, 1999)[13], obejmującej 264 standardy wraz z bardzo szczegółowymi objaśnieniami, zaczerpniemy szesnaście skróconych pozycji, dobranych do interesującej nas tematyki (podkreślenia własne - BN; w nawiasie podano numer standardu w wydawnictwie amerykańskim):

- Należy podać nazwiska i kwalifikacje niezależnych **ekspertów oceniających trafność testu**, a procedury przez nich zastosowane powinny być opisane. (*1.7*)
- Każdy wynik testowania, w tym także wynik każdej wyodrębnionej części testu, powinien być zaopatrzony w informację o **rzetelności i błędzie pomiaru**. (*2.1*)
- W przypadku zadań praktycznych i pisemnych zadań otwartych niezbędna jest informacja o **dokładności punktowania tych zadań przez niezależnych sędziów**. (*2.10*)
- Ogólne znaczenie i wszelkie **ograniczenia interpretacji zastosowanych skal pomiarowych** powinny być objaśnione w sposób zrozumiały dla użytkownika. (*4.1*)
- Procedura ustalania oraz interpretacji ewentualnej **normy ilościowej wymagań egzaminacyjnych** powinna być starannie uzasadniona i objaśniona. (*4.19*)
- Przy uzasadnianiu normy ilościowej wymagań programowych pożądane jest odwołanie się do **danych empirycznych** dotyczących konsekwencji stosowania normy. (*4.20*)
- Organizatorzy egzaminu powinni starannie przestrzegać **standaryzowanych procedur testowania**, chyba że niepełnosprawność ucznia wymaga odejścia od tych procedur. (*5.1*)
- Powinny być zastosowane właściwe środki do **zapewnienia uczciwości egzaminacyjnej** i minimalizacji oszustwa. (*5.6*)
- Dla uczniów, rodziców, nauczycieli, administracji szkolnej i środków masowego przekazu trzeba przygotować odpowiednią informację o egzaminie, o jego wynikach oraz o **najczęstszych błędach interpretacji tych wyników**. (*5.10*)
- Dane statystyczne, obejmujące analizę zadań, rozkłady wyników surowych, skale pochodne, normy testowe, **rzetelność i błąd pomiaru** oraz równoległość wersji testu, powinny być dostępne dla zainteresowanych. (*6.5*)
- Należy sprawdzić, czy jakieś lokalne, etniczne, inaczej wyposażone genetycznie lub inaczej kształcone grupy uczniów nie są **dyskryminowane przez treść i formę testu**. (*7.10*)
- **Trafność programowa testu** jest szczególnie ważna dla egzaminu doniosłego. (*13.5*)

[13] Wcześniejsza wersja tych standardów została opublikowana w języku polskim przez Polskie Towarzystwo Psychologiczne (Standardy…, 1985).

- Uczeń powinien mieć możliwość co najmniej jednorazowego **powtórzenia egzaminu doniosłego** przy zastosowaniu równoległej wersji testu. (*13.6*)
- **Informacja o błędzie pomiaru** powinna być przedstawiana każdemu użytkownikowi jego wyników w takiej formie, jaka odpowiada jego kompetencji statystycznej. (*13.14*)
- Udostępniając lub ogłaszając wyniki testowania, jego organizator powinien przedstawić wszelkie dane, które mogą **zapobiec błędom w interpretacji tych wyników**. (*15.11*)
- Organizator egzaminu doniosłego powinien spowodować, by ci, którzy podejmują decyzje o uczniach i szkołach **na podstawie wyników egzaminu**, byli do tego odpowiednio przygotowani i zaopatrzeni we właściwe materiały. (*15.13*)

Jak widać, kodeksowe przestrogi obejmują cały zakres błędów analizowanych w tym opracowaniu. Podobnie brzmią wybrane paragrafy „Kodu profesjonalnej odpowiedzialności w pomiarze pedagogicznym" (Schmeiser i in., 1995), przedstawionego w naszym kraju w nieco zmodyfikowanej formie następująco (Niemierko, 1997, r. 12):

A. Odpowiedzialność sprawdzającego i oceniającego osiągnięcia uczniów:

Unikaj niesprawdzonych narzędzi pomiaru oraz narzędzi, co do których masz istotne wątpliwości, a narzędzia i procedury standaryzowane stosuj dokładnie według instrukcji.

B. Odpowiedzialność komunikującego ocenę osiągnięć uczniów:

Wyniki oceniania przedstawiaj zawsze z komentarzem na temat sposobu ich uzyskania i z ostrzeżeniem o ich przybliżonej wiarygodności. Uczciwie przyznaj się do własnych pomyłek i rozczarowań.

C. Odpowiedzialność konstruktora narzędzi pomiaru osiągnięć uczniów:

Nie ukrywaj ograniczeń i słabości narzędzi pomiaru osiągnięć uczniów przed ich użytkownikami. Usilnie zapobiegaj wszelkim nadużyciom narzędzi i błędnym interpretacjom uzyskiwanych wyników.

D. Odpowiedzialność wydawcy i użytkownika narzędzi ewaluacji osiągnięć uczniów:

Nie reklamuj wydawanych narzędzi w sposób, który może wprowadzić ich użytkownika w błąd co do zasięgu i wartości ich stosowania. Przedstawiaj uczciwe dane o procesie standaryzacji i ostrzegaj przed możliwymi nadużyciami tych narzędzi.

Etyczny wymiar egzaminowania zewnętrznego stał się przedmiotem refleksji w dziele Marii Groenwald *Etyczne aspekty egzaminów szkolnych* (2011). Autorka dostrzega „egzaminacyjne dobro" w postaci otwierania drogi do kariery zawodowej, pracy ucznia nad przygotowaniem, wzrostu zaufania do nauczycieli. Przeważa jednak „egzaminacyjne zło": ukryty program kształcenia, uprzedmiotowienie uczniów i nauczycieli, bezwzględność rywalizacji, stygmatyzacja i wykluczenie, a w skrajnych przypadkach - oszustwo.

Ta kontrowersja aspektów etycznych może być skutkiem nieświadomości błędów pomiaru dydaktycznego, zbyt sztywnej interpretacji jego wyników, przeceniania edukacyjnej roli egzaminów szkolnych, a nawet „diagnostycznej fikcji uprawianej w wymiarze ogólnopolskim" (tamże, s. 180). Z braku krytycznych analiz „egzamin odbywa się we własnym obrębie, w sobie samym się zamyka i sam sobie dostarcza dowodów własnej niezbędności" (s. 178).

Podsumowanie

1. **Kontrola błędu** stosowanych procedur odróżnia profesjonalistę od amatora w każdej dziedzinie działalności.
2. W edukacji panuje niechęć do ujawniania **niepewności pomiarowej**, kierowana obawą o utratę autorytetu.
3. Ograniczona rzetelność pomiaru dydaktycznego skutkuje dużym **błędem losowym** wyników egzaminów szkolnych.
4. W ocenie rzetelności egzaminów powinien być stosowany **współczynnik Raju**, uwzględniający różną wagę zadań.
5. **Błąd lokalny** może obniżyć bezstronność, jakość punktowania i obiektywizm pomiaru sprawdzającego.
6. **Błąd systematyczny** obniża jakość punktowania, trafność i obiektywizm pomiaru sprawdzającego.
7. Ujawnianie błędów pomiaru jego użytkownikom stanowi **powinność moralną** nauk społecznych.

Bibliografia:

1. Brennan, R.L. (1983). *Elements of generalizability theory*. Iowa City: ACT.
2. Brookhart S. (2004). *Grading*. Upper Saddle River: Pearson Education.
3. Chodnicki, J. (2012). *Wpływ schematów oceniania zadań otwartych na trafność teoretyczną egzaminu zewnętrznego na przykładzie matematyki*. Nieopublikowana praca doktorska.
4. Cizek, G.J. (1999). *Cheating on tests. How to do it, detect it, and prevent it.* Mahwah: Erlbaum.
5. Ebel R.J. (1962) *Content-standard test scores*. Educational and Psychological Measurement, 1, 12-15.
6. Fricke, R. (1974). *Kriteriumsorientierte Leisungsmessung*, Stuttgart: Kohlhammer.
7. Groenwald, M. (2011). *Etyczne aspekty egzaminów szkolnych*. Glans: Wyd. UG.
8. Haertel E.H. (2002) *Standard setting as a participatory process: Implications for validation of standard-based accountability programs*. Educational Measurement: Issues and Practice, *1*.
9. Hornowska, E. (1999). *Stronniczość testów psychologicznych. Problemy - kierunki -kontrowersje*. Poznań: Humaniora.
10. Kane, M.D. (2007). Validation [w:] R.L. Brennan (red.) *Educational Measurement. Fourth Edition*. Westport: American Council on Education - Praeger.
11. Kane, M. (2011). *The errors of our ways*. Journal of Educational Measurement, 1, 12-30.

12. Kohlberg, L., Mayer R. (1993) *Rozwój jako cel wychowania* [w:] Z. Kwieciński i L. Witkowski (red.). *Spory o edukację. Dylematy i kontrowersje we współczesnych pedagogiach*. Toruń: Edytor.
13. Kolen, M.J., Brennan, R.L. (2004). *Test equating, scaling, and linking. Methods and practices* (2nd e.). New York: Springer-Verlag.
14. Ledzińska, M., E. Czerniawska (2011). *Psychologia nauczania. Ujęcie poznawcze*. Warszawa: PWN.
15. Lewis D.M., Mitzel H.C., Green D.R. (1999) *The bookmark standard setting procedure*.Monterey: McGraw-Hill.
16. Mietzel, G. (2002). *Psychologia kształcenia. Praktyczny podręcznik dla pedagogów i nauczycieli*. Gdańsk: GWP.
17. Muszyński, H. (1970). *Wstęp do metodologii pedagogiki*. Warszawa: PWN.
18. Niemierko, B, (1975). *Pomiar osiągnięć szkolnych. Podstawowe pojęcia i techniki obliczeniowe*. Warszawa: WSiP.
19. Niemierko, B. (1990) *Pomiar sprawdzający w dydaktyce. Teoria i zastosowania*. Warszawa: PWN.
20. Niemierko, B. (1997). *Między oceną szkolną a dydaktyką. Bliżej dydaktyki*. Warszawa: WSiP.
21. Niemierko, B. (1999). *Pomiar wyników kształcenia.* Warszawa: WSiP.
22. Niemierko, B. (2007). *Kształcenie szkolne. Podręcznik skutecznej dydaktyki*. Warszawa: WAiP.
23. Niemierko, B. (2009). *Diagnostyka edukacyjna. Podręcznik akademicki*. Warszawa: PWN.
24. Pieter, J. (1969). *Egzamin obiektywny*. Warszawa: Nasza Ksiegarnia.
25. Raju, N.S. (1970). *New formula for estimating total test reliability from parts of unequal length*. Proceedings of the 78th Annual Convention of the American Psychological Association, 5, 143-14.
26. Raju, N.S. (1977). *A generalization of the coefficient alpha*. Psychometika, 42, 549-565.
27. Rubacha, K. (2008). *Metodologia badań nad edukacją*. Warszawa: WAiP.
28. Schmeiser, C. B., Geisinger, K. F., Johnson-Lewis, S., Roeber, E. D., Schafer, W. D. (1995) *Code of professional responsibilities in educational measurement*. Washington: NCME.
29. *Standards for educational and psychological testing* (1999). Washington: American Educational Research Association.
30. *Standardy dla testów stosowanych w psychologii i pedagogice* (1985). Warszawa: Polskie Towarzystwo Psychologiczne.
31. Stępniowski, I, W. Małecki (1974). *Probabilistyczne własności metody wyboru a dobór skali ocen*. Dydaktyka Szkoły Wyższej, 1.
32. Stożek, E. (2008). *Zadania rozrywające w testach*. Edukacja, 1.
33. Tyszka T. (1999). *Psychologiczne pułapki oceniania i podejmowania decyzji*. Gdańsk: GWP.