Sławomir Sapanowski
Okręgowa Komisja Egzaminacyjna w Łodzi

Szacowanie błędu pomiaru ze szczególnym uwzględnieniem wyników sprawdzianu w VI klasie szkoły podstawowej w latach 2002–2012

1. Wstęp

Z każdym pomiarem wiążą się pojęcia niepewności pomiarowej i błędu pomiaru. Przy czym rozumiemy tutaj błąd jako cechę narzędzia pomiarowego i właściwość mierzonej wielkości, a nie pomyłkę pomiarowca czy też błędny odczyt lub niewłaściwą interpretację otrzymanych wyników. Przez błąd pomiaru rozumiemy różnicę między wielkością rzeczywistą a zmierzoną, mimo dołożenia wszelkiej możliwej staranności podczas pomiaru. Oczywiście wielkości samego błędu nigdy nie poznamy, tak jak i nie poznamy rzeczywistej wartości mierzonej wielkości. Pozostają nam jedynie statystyczne metody pozwalające oszacować dokładność pomiaru i co za tym idzie wielkość błędu.

System egzaminów zewnętrznych dotyczy uczniów kolejnych roczników. Sprawdzany jest w ten sposób poziom umiejętności i wiedzy nabytej podczas kolejnych etapów edukacji szkolnej. Czy proces ten możemy nazwać pomiarem? Odpowiedź na to pytanie jest twierdząca. Mamy tutaj do czynienia z wszelkimi elementami wchodzącymi w skład procesu pomiarowego:

- uczeń, czyli obiekt, który będzie badany;
- jego umiejętności i wiedza, czyli cecha obiektu poddana pomiarowi;
- zbiór zadań testowych, czyli narzędzie, którym badaną cechę chcemy zmierzyć;
- zbiór danych liczbowych opisujących badaną cechę, czyli wyniki testowania.

A nawet więcej, ponieważ powstała cała gałąź nauki zwana pomiarem dydaktycznym, zajmująca się wytwarzaniem i interpretacją wyników testowania uczniów, które z kolei mają świadczyć o stopniu opanowania pożądanych treści nauczania.

Okręgowe komisje egzaminacyjne wspólnie z Centralną Komisją Egzaminacyjną powinny interpretować wyniki egzaminów również w kontekście błędów pomiarowych. Powinny, ale tak nie robią. Rezultat testowania w postaci wyniku punktowego lub procentu uzyskanych punktów przekazywany jest wszystkim zainteresowanym stronom (uczniom, rodzicom, organom prowadzącym i nadzorującym szkoły, mediom i in.) bez jakiegokolwiek komentarza na temat niepewności pomiarowej. Wygląda to mniej więcej tak jak komunikat o odległości między Łodzią i Warszawą, która to odległość jest równa 121 231,562 m. Wyniku takiego nikt oczywiście nie bierze na poważnie.

Już samo przybliżenie do pełnych kilometrów (121 km) może budzić wątpliwości, a dokładność do milimetrów powoduje jedynie uśmiech politowania i chęć zignorowania tej informacji jako zupełnie niewiarygodnej.

2. Klasyczna teoria pomiaru dydaktycznego

a) odchylenie standardowe i rzetelność testu

Klasyczna teoria testu, która znalazła szerokie zastosowanie w Polsce dopiero z wprowadzeniem systemu egzaminów zewnętrznych, oparta głównie na pracach prof. B. Niemierki, również pozwala oszacować błąd wyników testowania.

Kluczowymi pojęciami wykorzystywanymi do wyznaczania niepewności pomiarowej będą odchylenie standardowe wyników testowania oraz rzetelność testu. Z pierwszym parametrem nie ma wielkiego kłopotu – jest to doskonale znana wielkość, definiowana jako

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

gdzie:

σ – odchylenie standardowe wyników testowania

μ – średnia wyników testowania

x_i – wynik i-tego ucznia

N – liczba uczniów.

A intuicyjnie podpowiada, jak szeroko wartości jakiejś wielkości (w naszym przypadku wyniki testowania), są rozrzucone wokół jej średniej.

Mimo iż rzetelność testu wszyscy autorzy klasycznej teorii pomiaru dydaktycznego definiują podobnie, jako dokładność dokonywanego przez niego pomiaru lub powtarzalność wyników testowania, to kłopotliwe jest wybranie analitycznej wersji definicji.

Do obliczania wskaźnika[1] rzetelności testu stosowane są różne metody:

- porównanie rezultatów połówek tego samego testu – koreluje się wyniki uzyskane w odpowiedzi na parzyste i nieparzyste części testu;
- porównanie wyników uzyskanych w badaniach dwiema alternatywnymi wersjami danego testu – problemem jest tu dbałość o taką samą trafność i łatwość obu wersji, co jest często trudne do osiągnięcia;
- obliczanie właściwości statystycznych poszczególnych pozycji testowych – korelacja rezultatów poszczególnych zadań testowych z wynikiem ogólnym testu – w szczególności porównanie udziału wariancji poszczególnych pozycji testowych w wariancji wyników testowania.

Niestety nie ma jednej, uniwersalnej metody określającej rzetelność testu.

[1] Wskaźnik rzetelności testu jest liczbą z przedziału <0; 1>, przy czym wartości bliskie jedności oznaczają wysoką rzetelność, bliskie zeru – niską.

Do obliczania wskaźników rzetelności metodą połówkową stosuje się różne wzory, które stanowią odmiany podstawowej metody Spearmana-Browna. Najczęściej polecaną techniką dzielenia testu jest podział zadań na parzyste i nieparzyste. I stosowanie wzoru:

$$r_{tt} = \frac{2r_p}{1 + r_p}$$

gdzie:

r_p– korelacja pomiędzy połowami testu.

Innym wzorem, zaproponowanym przez Rulona (1937), który można zastosować do obliczenia rzetelności metodą połówkową, jest wzór:

$$r_{tt} = 1 - \frac{\sigma_p^2}{\sigma_t^2}$$

gdzie:

σ_p– odchylenie standardowe różnic między wynikami w dwóch połówkach testu

σ_t – odchylenie standardowe wyników ogólnych.

Wzór ten wynika z podstawowego pojęcia rzetelności: rzetelność jest proporcją wariancji prawdziwej w teście.

Odmianą wzoru Rulona jest propozycja Flanagana (1937):

$$r_{tt} = 2\left(1 - \frac{\sigma_1^2 - \sigma_2^2}{\sigma_t^2}\right)$$

gdzie:

σ_1^2 – wariancja pierwszej połowy testu

σ_2^2 – wariancja drugiej połowy testu

σ_t^2 – wariancja całego testu.

Badanie zgodności wewnętrznej, stanowiące odmianę metody połówkowej, można przeprowadzić, stosując wzór Kudera-Richardsona (tzw. wskaźnik KR-20):

$$r_{tt} = \frac{n}{n-1}\left(1 - \frac{\sum_{i=1}^{n} p_i q_i}{\sigma_t^2}\right)$$

gdzie:

r_{tt} – rzetelność testu (KR-20)

n – liczba zadań w teście

p_i – łatwość i-tego zadania

q_i – trudność i-tego zadania

σ_t^2 – wariancja wyników testowania.

Odmianą wzoru Kudera-Richardsona, a jednocześnie jednym z najbardziej popularnych wskaźników rzetelności jest tzw. alfa Cronbacha (1951):

$$\propto = \frac{n}{n-1}\left(1 - \frac{\sum_{i=1}^{n}\sigma_i^2}{\sigma_t^2}\right)$$

gdzie:

$\propto$ – rzetelność testu (alfa Cronbacha)

n – liczba zadań w teście

σ_i^2 – wariancja i-tego zadania

σ_t^2 – wariancja wyników testowania.

Przy czym należy pamiętać, że współczynnik Kudera-Richardsona obliczamy dla testu składającego się z pozycji (zadań) dwukategorialnych (dychotomicznych), a wskaźnik alfa Cronbacha – dla pozycji (zadań) wielokategorialnych.

b) błąd pomiaru i przedział ufności

Standardowy błąd pomiaru (*standard error of measurement*) dla wyniku otrzymanego za pomocą testu (arkusza egzaminacyjnego) możemy obliczyć za pomocą formuły:

$$SEM = \sigma_t\sqrt{1 - r_{tt}}$$

gdzie:

SEM – standardowy błąd pomiaru

σ_t – odchylenie standardowe wyników ogólnych

r_{tt} – rzetelność testu.

W sposób oczywisty należy dążyć do konstruowania testów o wysokiej rzetelności. A małe odchylenie standardowe zapewnia niewielki błąd pomiaru. I tutaj spotkamy się z pewnym konfliktem. Chcielibyśmy ograniczyć błąd pomiaru poprzez minimalizację odchylenia standardowego, a jednocześnie test charakteryzujący się odchyleniem standardowym mniejszym niż 1/5 skali pomiarowej jest zbyt mało różnicujący (Niemierko, 1999). Musimy również pamiętać, że w przypadku egzaminów zewnętrznych ograniczamy się tylko do jednokrotnego badania populacji uczniów i arkusz egzaminacyjny generujący wyniki o znikomym odchyleniu standardowym (wyniki uczniów są zbliżone do średniej) jest bezużyteczny.

Jeśli użyty test posiada wysoką rzetelność i jednocześnie odchylenie standardowe otrzymanych wyników spełnia nasze oczekiwania (jest małe, ale nie jest mniejsze niż 1/5 skali pomiarowej), możemy oszacować szerokość przedziału ufności dla wyniku konkretnego ucznia.

$$\langle W - z_{\propto/2}SEM;\ W + z_{\propto/2}SEM\rangle$$

gdzie:

$z_{\propto/2}$ – wartość statystyki „z" rozkładu normalnego dla danego poziomu ufności

SEM – standardowy błąd pomiaru

W – wynik otrzymany w teście dla danej osoby.

Wartości $Z_{\alpha/2}$ dla różnych poziomów istotności:

α =	0,15	0,10	0,05	0,01
$Z_{\alpha/2}$ =	1,44	1,64	1,96	2,56

c) błędy pomiaru dla sprawdzianu w szóstej klasie szkoły podstawowej[2]

System egzaminów zewnętrznych obejmuje swoim zasięgiem uczniów kończących kolejne etapy kształcenia (szkoła podstawowa, gimnazjum, szkoła ponadgimnazjalna). Doniosłość egzaminów jest z oczywistych względów różna. Wpływ wyniku egzaminu maturalnego na przyszłość abiturienta jest nieporównywalnie większy niż wynik sprawdzianu w szóstej klasie szkoły podstawowej. Dlaczego więc analiza błędów pomiarowych tutaj przedstawiona dotyczy tego ostatniego? Taki wybór uzasadnić można, podając kilka argumentów, ale najważniejsze z nich to:

- sprawdzian funkcjonuje w szkołach od początku reformy oświaty zapoczątkowanej w latach 90. ubiegłego wieku;
- sprawdzian jako jedyny egzamin zewnętrzny zachował swoją formę i zawartość treściową.

Egzamin gimnazjalny w tym czasie poddany został reformie polegającej na podziale arkusza humanistycznego na zakres dotyczący j. polskiego i zakres sprawdzający wiedzę z historii i wiedzę o społeczeństwie, natomiast arkusz matematyczno-przyrodniczy ewoluował w kierunku egzaminu z matematyki i przedmiotów przyrodniczych. Ponadto zamiast oceniania analitycznego wprowadzono tzw. ocenianie holistyczne.

Jeśli zaś przyjrzymy się egzaminowi maturalnemu, to ilość zmian wprowadzanych jest porównywalna z liczbą lat funkcjonowania tego egzaminu. Dlatego też wybór sprawdzianu w szóstej klasie szkoły podstawowej nie jest wyborem przypadkowym. W poniższej tabeli zestawiono odchylenia standardowe oraz rzetelność testów[3] z lat 2002–2012.

	2002	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**
α =	0,86	0,87	0,89	0,87	0,90	0,89	0,87	0,86	0,87	0,82	0,84
σ_t =	6,58	6,63	7,63	7,29	8,42	7,73	7,43	7,44	7,93	7,41	7,58

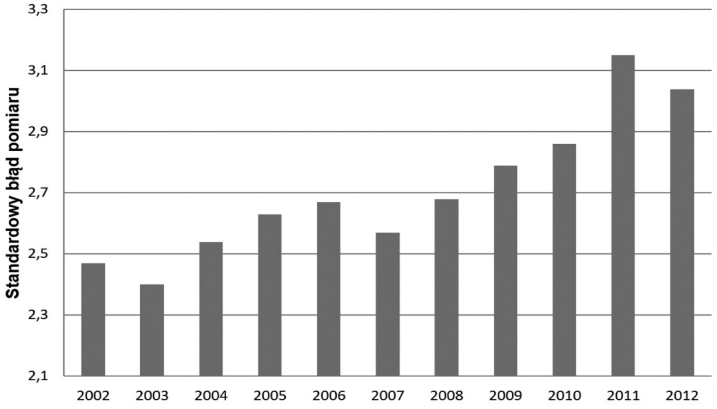
Jeśli dla tych danych policzymy SEM w kolejnych latach, to otrzymamy:

	2002	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**
SEM	2,47	2,40	2,54	2,63	2,67	2,57	2,68	2,79	2,86

[2] Wszystkie parametry statystyczne tutaj przywoływane dotyczą populacji uczniów objętych zasięgiem działania Okręgowej Komisji Egzaminacyjnej w Łodzi (woj. łódzkie i woj. świętokrzyskie).

[3] Ze względu na brak dychotomiczności zadań rzetelności liczone są za pomocą wskaźnika alfa Cronbacha.

Zaskakujące jest to, że w czasie, kiedy system egzaminów zewnętrznych był „młody", a **polscy konstruktorzy zadań i testów nie mieli obecnych doświadczeń**, standardowy błąd pomiaru był mniejszy. Wraz z kolejnymi latami błąd pomiaru wzrastał [sic!]. Jest to jeszcze lepiej widoczne, gdy rezultaty przedstawimy w postaci graficznej.



3. Probabilistyczna teoria testu (IRT) a ilość informacji i błąd pomiaru

Informacja i wiedza są obecnie uważane za nowy towar na rynku, podobny do dóbr materialnych czy energii. Równocześnie, ze względu na internet i inne masowe źródła informacji, obecne społeczeństwo globalne nazywane jest też społeczeństwem informacyjnym (Information Society). Tak zastosowane pojęcie informacji dotyczy również wiedzy faktycznej lub domniemanej, a także reguł preferencji w różnych dziedzinach ważności i użyteczności. W tym sensie, informując kogoś o kimś lub o czymś, zawiadamiamy go o faktach lub dzielimy się naszą wiedzą albo preferencjami na dany temat[4].

Informacja jest pojęciem intuicyjnie dość zrozumiałym i wynikającym z potrzeby komunikowania się. Przez informację rozumiemy interpretację ciągu sygnałów (literowych, liczbowych, werbalnych, świetlnych, dźwiękowych i in.) opisującym stan jakiejś dziedziny. Może ona (informacja) być zapisywana i przechowywana. Jednakże trudności pojawią się w momencie, gdy chcemy zmierzyć ilość informacji. Tutaj sama intuicja zawodzi, a chcąc określić ilość informacji (w sposób dość popularny), czujemy, że zależy ona od tego, jak często (z jakim prawdopodobieństwem) dany komunikat, niosący informację, występuje. Takim ilościowym podejściem zajmuje się teoria informacji Hartleya i Shannona. Miara ilości informacji jest w niej oparta właśnie na prawdopodobieństwie zajścia zdarzenia. Jako miarę ilości informacji przyjmuje się wielkość niepewności, która została usunięta w wyniku zajścia zdarzenia (otrzymania komunikatu). Komunikaty mniej prawdopodobne dają więcej informacji.

[4] http://pl.wikipedia.org/wiki/Informacja.

Ilość informacji zawartej w komunikacie jest definiowana wzorem (Hartley, 1928):

$$I = \log_r \frac{1}{p} = -\log_r p$$

gdzie:

I – ilość informacji zawartej w komunikacie

p – prawdopodobieństwo wystąpienia komunikatu

r – podstawa logarytmu.

W teorii informacji najczęściej stosuje się logarytm o podstawie r = 2, wówczas jednostką informacji jest bit (*binary information unit*). Przy r = e (podstawa logarytmu naturalnego) jednostką jest nat (*nit – natural information unit*), natomiast przy r = 10 – dit (*digital information unit*).

Tak zdefiniowana informacja ma tę właściwość, że jest addytywna. Możemy ją dodawać pod warunkiem, że pozyskane informacje są niezależne. W związku z tym narzuca się podobieństwo do założeń probabilistycznej teorii testu (IRT, Item Response Theory), gdzie mówi się o niezależności zadań testowych i o prawdopodobieństwie udzielenia poprawnej odpowiedzi przez ucznia w zależności od poziomu umiejętności (cechy latentnej Θ). A przecież wyniki egzaminu są komunikatem niosącym informacje. Pozostaje tylko pytanie o jej ilość.

Ponieważ założenia probabilistycznej teorii testu są powszechnie dostępne, nie będziemy ich tutaj przywoływać i odeślemy czytelnika do stosownej literatury.

Stosując IRT, można zauważyć, że ilość informacji dostarczona przez komunikat o wynikach egzaminów nie jest stała[5] – zmienia się w zależności od wyniku uzyskanego przez ucznia, a mówiąc ściślej – od poziomu jego umiejętności, który na podstawie uzyskanego wyniku możemy oszacować. Dlatego też ilość informacji (dla modelu 2-parametrycznego) będzie dość złożoną funkcją umiejętności ucznia (Θ).

$$I_i(\Theta)=a_i^2 f_i(\Theta)(1- f_i(\Theta))$$

gdzie:

$I_i(\Theta)$ – funkcja informacyjna (ilość informacji) dla i-tego zadania

a_i – parametr określający moc różnicującą i-tego zadania

$f_i(\Theta)$ – funkcja charakterystyczna i-tego zadania.

Ponieważ funkcja informacyjna jest funkcją addytywną, możemy obliczyć ilość informacji dostarczoną przez wyniki całego testu:

$$I_t(\Theta) = \sum_{i=1}^{n} I_i(\Theta)$$

gdzie:

$I_t(\Theta)$ – funkcja informacyjna dla całego testu

n – liczba zadań w teście.

[5] Ma to znaczenie w kontekście szacowania błędu pomiaru. W klasycznej teorii testu obliczony błąd pomiaru jest jednakowy dla wszystkich uczniów. Gdy zastosujemy IRT i teorię informacji – tak nie jest. Błąd pomiaru (najczęściej) będzie zależał od pozycji ucznia na skali pomiarowej.

Jeżeli ilość dostarczonej przez zadania informacji dla danego poziomu umiejętności (Θ) jest duża, oznacza to, że umiejętności ucznia mogą być oszacowane z dużą dokładnością (mały błąd standardowy oszacowania). Oznacza to także, że wynik uzyskany przez danego ucznia będzie bardzo blisko jego prawdziwego poziomu umiejętności. Jeżeli ilość informacji dostarczana przez wynik egzaminowania jest mała, to i wyniki testowania mogą znajdować się dosyć daleko od prawdziwej wartości mierzonej.

A zatem błąd standardowy szacowany na podstawie IRT będzie równy:

$$\mathrm{SEM}(\Theta) = \frac{1}{\sqrt{\mathrm{I_t}(\Theta)}}$$

W sposób naturalny pojawia się pytanie o to, czy błąd standardowy liczony w sposób klasyczny jest mniejszy, czy większy, a może porównywalny z liczonym w sposób probabilistyczny.

4. Przykład zastosowania IRT do oszacowania SEM dla sprawdzianu 2007[6]

Jeżeli zadamy sobie trochę trudu[7] i wykorzystamy możliwości obliczeniowe dowolnego arkusza kalkulacyjnego, możemy wykreować dla każdego zadania krzywą charakterystyczną (ICC), a także obliczyć ilość informacji, którą niesie konkretne zadanie[8].

Zamieszczanie w tym opracowaniu krzywych dotyczących wszystkich zadań oraz funkcji opisujących ilość informacji naraziłoby czytelnika na żmudne przedzieranie się przez gąszcz wykresów i diagramów, a w efekcie mijałoby się z naszym celem, czyli oszacowaniem błędu pomiaru podczas sprawdzianu w szóstej klasie szkoły podstawowej w roku 2007. Dlatego też ograniczę się do podania funkcji informacyjnej dla całego testu, która – dla przypomnienia – jest sumą funkcji informacyjnych dla poszczególnych zadań.

Przyjmuje się, że zadowalający poziom informacji, dla testów stosowanych w pomiarze dydaktycznym, wynosi 2. A zatem, jak łatwo zauważyć, w przedziale punktowym od 6 do 35 punktów analizowany sprawdzian niesie dostateczną ilość informacji[9]. Warto tutaj zauważyć, że ze względu na strukturę arkusza egzaminacyjnego (20 zadań zamkniętych), wyniki poniżej 6 punktów wynikają z przypadkowych odpowiedzi uczniowskich i nie niosą ze sobą żadnej wartościowej wiedzy na temat poziomu osiągnięć zdającego.

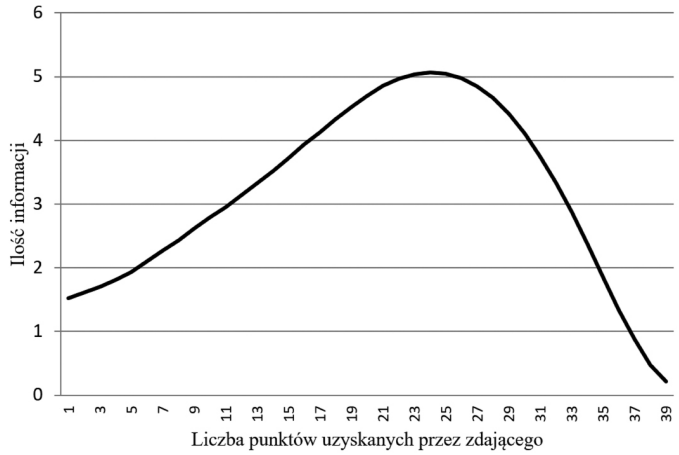
[6] Wybrano ten rocznik ze względu na to, że system egzaminów zewnętrznych „okrzepł" już nieco, a ponadto w tym roku błąd pomiaru był jednym z najmniejszych.

[7] Musimy oczywiście mieć dostęp do indywidualnych wyników testowania, a te z kolei (właściwie nie wiadomo dlaczego) są przez system niezbyt chętnie udostępniane.

[8] Mówiąc zupełnie na marginesie, informacja obliczona w ten sposób mogłaby służyć do określenia, czy zadanie jest wartościowe. Czy powinno znaleźć się w teście? Czy wreszcie nie zaburza ono wyników testowania i byłoby najwłaściwiej usunąć je z analiz?

[9] Ze względu na własności funkcji logistycznej nie można określić poziomu umiejętności ucznia dla krańców skali pomiarowej (0 i 40 punktów). Dlatego też nie jest możliwe podanie wartości funkcji informacyjnej ani oszacowanie błędu dla tych wyników.

Z drugiej strony, dla uczniów osiągających wyniki najwyższe sytuacja jest podobna, gdzie tym razem strata punktów ma charakter dość przypadkowy i niewynikający wprost z wiedzy i umiejętności uczniowskich.

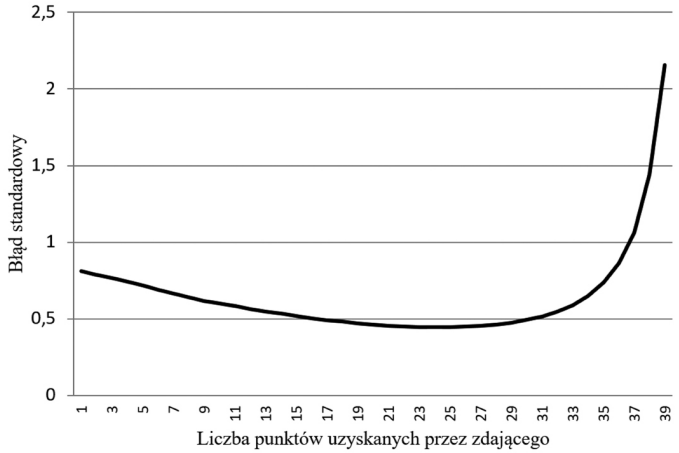


Na szczęście niski poziom informacji dotyczy niewielkiej grupy osób (ok. 12%). Aby dokładniej określić ich pozycję na skali umiejętności (Θ), należałoby przeprowadzić dodatkowe, specjalnie skonstruowane testy dla uczniów najsłabszych i najwybitniejszych. Co oczywiście na potrzeby badaczy byłoby pożyteczne, ale dla systemu oświaty całkowicie zbędne.

Wykorzystajmy, podawany już wcześniej, wzór na wielkość błędu standardowego:

$$\mathrm{SEM}(\Theta) = \frac{1}{\sqrt{I_t(\Theta)}}$$

Proste przekształcenie prowadzi nas do wykresu:



W porównaniu z przewidywaniami klasycznej teorii testu wynik jest zaskakujący. Praktycznie w całym zakresie skali pomiarowej (z wyjątkiem wyników najwyższych) błąd standardowy nie jest większy niż 1 punkt!

Czyżby wyniki uzyskane klasycznie i probabilistycznie tak bardzo się różniły? Jak to możliwe? Która metoda daje dobre rezultaty? Przecież rezultaty testowania z przedziału od 15 do 30 punktów obciążone są błędem mniejszym niż 0,5 pkt. A zatem wyniki tych uczniów określone są bardzo precyzyjnie, biorąc pod uwagę, że testowanie odbyło się jednokrotnie.

Kiedy jednak przyjrzymy się dokładniej powyższemu wykresowi, zauważymy, że największy błąd, przewidywany przez IRT, wynosi 2,16. Teoria klasyczna podaje tutaj wartość błędu – 2,57. Jeśli weźmiemy pod uwagę to, że szacując błąd pomiaru, nigdy i na żadnym etapie nie możemy pozwolić sobie na jego niedoszacowanie, oraz to, że klasycznie obliczona wartość jest jedną liczbą, obejmującą cały zakres wyników testowania, dochodzimy do wniosku, że zachodzi pewna spójność między obiema teoriami. Pierwsza podaje jedną wartość błędu, która jest jednocześnie wartością maksymalną, druga – generuje funkcję określającą błąd przypisany danemu wynikowi testowania. Przy czym obie metody podają podobną liczbową wartość błędu maksymalnego.

5. Podsumowanie

Analizy rezultatów testowania w kolejnych latach potwierdzają, że za pomocą klasycznych metod w stosunkowo prosty sposób otrzymamy wartość błędu standardowego. Metoda probabilistyczna zaś, bardziej czasochłonna i wymagająca wspomagania komputerowego, daje precyzyjniejsze wyniki, jednakże nie różni się od metod klasycznych, jeśli chodzi o przewidywanie błędu maksymalnego, występuje tutaj istotna zbieżność.

Bibliografia

Ciżkowicz B., *Klasyczna a probabilistyczna teoria testu. Analiza porównawcza* [w:] Biuletyn Badawczy nr 9, CKE, Warszawa 2007.

Ciżkowicz K., Niemierko B., *Elementy statystyki w klasycznej teorii testu: sto terminów i sto zadań,* WSiP, Warszawa 1991.

Mynarski S., *Elementy teorii systemów i cybernetyki*, Warszawa 1979.

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

Sapanowski S., *Oszacowanie umiejętności „teta" oraz wyskalowanie osi w metodzie IRT dla potrzeb obliczania parametrów zadań* [w:] *Uczenie się i egzamin w oczach nauczycieli,* XIV KDE, Opole 2008.

Sztejnberg A., Hurek J., *Zastosowanie teorii analizy zadania testowego (IRT) w procesie oceniania zewnętrznego* [w:] *Teoria i praktyka oceniania zewnętrznego* pod red. nauk. B. Niemierki i M.K. Szmigel, Wydawnictwo PANDIT, Kraków 2001.

Verhelst N., *Probabilistyczna teoria wyniku zadania testowego* [w:] Biuletyn Badawczy nr 9, CKE, Warszawa 2007.