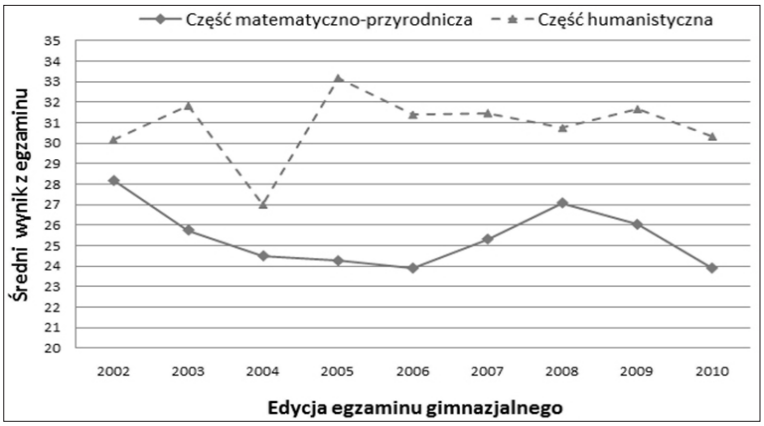
Bartosz Kondratek
Instytut Badań Edukacyjnych

Zrównywanie wyników egzaminów zewnętrznych z wykorzystaniem modeli IRT

Wprowadzenie

Mimo wysiłków twórców arkuszy egzaminacyjnych, aby utrzymać stały poziom trudności egzaminów zewnętrznych, nie sposób uniknąć występowania fluktuacji trudności między kolejnymi sesjami. Niezależnie od zmian trudności egzaminów między sesjami mogą również występować zmiany w poziomie umiejętności uczniów. W konsekwencji pokrywania się (ang. *confounding*) tych dwóch zjawisk nieobciążone porównywanie wyników uczniów między sesjami nie jest możliwe bez zastosowania statystycznej procedury zrównywania wyników (ang. *test equating*), która rozdzielałaby różnice w trudności testów od różnic w poziomie umiejętności uczniów.



Rysunek 1. Średnie z egzaminu gimnazjalnego na przestrzeni lat 2002-2010

Przykładowo średnie wyniki egzaminu gimnazjalnego w latach 2002-2010 zmieniały się w zakresie od 24 do 28 oraz od 27 do 33 dla części matematyczno-przyrodniczej oraz humanistycznej odpowiednio (rys. 1.). Bez zrównania wyników nie jesteśmy w stanie stwierdzić, w jakim stopniu zaobserwowane zmiany średnich są konsekwencją różnic w trudnościach testów, a w jakim odzwierciedlają zmiany w poziomie umiejętności uczniów. Uniemożliwia to wyznaczenie trendu zmian poziomu umiejętności jak i porównywanie wyników pojedynczych uczniów przystępujących do egzaminów w różnych latach. Problem jest zatem doniosły.

W artykule zostaną przybliżone podstawowe metody zrównywania wyników obserwowanych z wykorzystaniem modeli IRT (*Item Response Theory*) dla planu nierównoważnych grup z testem kotwiczącym (ang. *Nonequivalent Groups with Anchor Test*, NEAT). Następnie zostanie opisany prosty eksperyment Monte Carlo, ilustrujący właściwości zrównywania IRT metodą kalibracji łącznej.

Zrównywanie wyników w planie NEAT

Plan NEAT jest podstawowym planem zbierania danych umożliwiającym zrównanie wyników dwóch różnych testów, *X* oraz *Y*, rozwiązywanych przez rozłączne populacje, P oraz Q. Schematycznie przedstawiono go w tab. 1.

Tabela 1. Plan nierównoważnych grup z testem kotwiczącym

Test / Populacja	*X*	*Y*	*A*
P	√		√
Q		√	√

Odpowiedzi uczniów na testy *X* oraz *Y* w sposób bezpośredni pozwalają oszacować dystrybuanty $F_{x|P}$ oraz $F_{x|Q}$, jednak nie mamy informacji o dystrybuantach $F_{x|Q}$ oraz $F_{x|P}$ opisujących rozkład wyników testu *X* w populacji *Q* oraz testu *Y* w populacji *P*. Aby oszacować $F_{x|Q}$ oraz $F_{x|P}$ dane muszą być połączone (ang. *linked*), co w planie NEAT następuje poprzez wykorzystanie dodatkowego zestawu zadań rozwiązywanego przez uczniów z obu populacji – nosi on nazwę kotwicy (ang. *anchor*) i w tab. 1. został oznaczony literą *A* . Kotwica może stanowić odrębny test, może również być zbiorem zadań znajdujących w tym samym arkuszu co zrównywane testy, ale nie uwzględnianych przy obliczaniu punktacji w testach *X* oraz *Y*.

Dla ustalenia wagi przyjmijmy, że testem „bazowym", do którego pragniemy odnosić wyniki drugiego testu, jest test *X* . Zadaniem stojącym przed procedurą zrównywania jest oszacowanie dystrybuanty $F_{x|Q}$ opisującej rozkład testu *X* w populacji *Q*, w której nigdy nie był on administrowany. Znając dystrybuantę $F_{x|Q}$ możemy wyznaczyć ekwicentylową funkcję zrównującą eq_x (ang. *equating function*) przypisującą każdemu wynikowi *y* w teście *Y* jego odpowiednik *x* w teście *X*:

$$eq_X(y) = {}^{(cont)}F^{-1}_{X|Q}\left({}^{(cont)}F_{Y|Q}(y)\right). \qquad (1)$$

Górny lewy indeks $^{(cont)}$ przy dystrybuantach został wprowadzony na oznaczenie uciąglenia i wygładzenia dystrybuant. Operacja taka jest konieczna, ponieważ wynik w teście jest zmienną dyskretną i odpowiadająca mu dystrybuanta ma postać funkcji skokowej – takiej funkcji nie można odwracać. Najpopularniejszymi metodami uciąglania skokowych dystrybuant jest lokalna interpolacja liniowa oraz wygładzanie za pomocą estymatora jądrowego (*kernel smoothing*). Pogłębiony przegląd pierwszego podejścia można znaleźć u Kolen & Brennan (2004), a drugiego u von Davier et. al. (2004). Za szczególny, zdegenerowany, sposób wygładzania dystrybuant w równaniu (1) można również uznać (Dorans et. al., 2011) zrównywanie liniowe, które jest oparte jedynie na dwóch pierwszych momentach rozkładów:

$$eq_X(y) = \frac{\sigma_{Y|Q}(x - \mu_{X|Q})}{\sigma_{X|Q}} + \mu_{Y|Q}.$$

Zrównywanie wyników z wykorzystaniem IRT

Bez wprowadzania dodatkowych założeń dotyczących modelu pomiarowego leżącego u podstaw udzielania odpowiedzi na zadania testu zrównanie wyników dla planu NEAT jest możliwe np. poprzez zastosowanie zrównywania łańcuchowego (ang. *chained equating*) lub post-stratyfikacyjnego (ang. *post-stratified equating*). Czytelnik zainteresowany tymi metodami może sięgnąć do wspomnianej monografii von Davier et. al. (2004). Niniejszy artykuł zajmie się natomiast metodami opartymi na IRT, w których mechanizm udzielania odpowiedzi na zadania testów jest modelowany *explicite*.

W IRT odpowiedzi na zadania testu u_i są zmiennymi losowymi opisanymi poprzez parametry charakteryzujące właściwości zadań oraz ukrytą zmienną losową θ modelującą poziom umiejętności uczniów. Zmienna θ może być wielowymiarowa (również w kontekście zrównywania – Reckase, 2009), jednak my skoncentrujemy się na rozwiązaniach z pojedynczym wymiarem. Rozkład θ najczęściej przyjmuje się jako normalny $N(\mu, \sigma_2)$.

Jednowymiarowym modelem dla zadań ocenianych dychotomicznie jest, przykładowo, dwuparametryczny model logistyczny (2PLM), w którym prawdopodobieństwo udzielenia prawidłowej odpowiedzi na zadanie i w zależności od poziomu umiejętności θ jest określone funkcją, *nomen omen*, dwójki parametrów $\beta_i=(a_i,b_i)$:

$$\mathbb{P}(u_i = 1|\theta, \boldsymbol{\beta_i}) = p_i(\theta) = \frac{1}{1 + e^{-a_i(\theta - b_i)}}. \qquad (2)$$

Szczegółowe omówienie właściwości modelu danego wzorem (2) można znaleźć u Kondratka (2007) oraz praktycznie w każdym innym tekście wprowadzającym do tematyki IRT.

Dla dychotomicznych zadań funkcja $\rho_i(\theta)$ w elegancki sposób pozwala określić prawdopodobieństwo dowolnej odpowiedzi (czyli 0 lub 1) w zależności od θ:

$$\mathbb{P}(u_i|\theta, \boldsymbol{\beta_i}) = p_i(\theta)^{u_i}(1 - p_i(\theta))^{u_i - 1},$$

co przy uwzględnieniu założenia o jednowymiarowości testu (precyzyjniej: *lokalnej niezależności* – Lord & Novick, 1968), pozwala przedstawić prawdopodobieństwo dowolnego wektora odpowiedzi $\boldsymbol{u}$ na wszystkie n zadań testu w zależności od θ jako następujący produkt:

$$\mathbb{P}(\boldsymbol{u}|\theta, \boldsymbol{\beta}) = \prod_{i=1}^{n} p_i(\theta)^{u_i}(1 - p_i(\theta))^{u_i - 1}.$$

Całkując powyższe warunkowe prawdopodobieństwo przez rozkład umiejętności, otrzymujemy sparametryzowane modelem IRT prawdopodobieństwo uzyskania dowolnego wektora odpowiedzi w danym teście w określonej populacji:

$$\mathbb{P}(\boldsymbol{u}) = \int \mathbb{P}(\boldsymbol{u}|\theta, \boldsymbol{\beta}) N(\mu, \sigma^2)\, d\theta. \quad (3)$$

Rozkład wektora odpowiedzi podany wzorem (3) określa wszystkie statystyczne właściwości testu w populacji, jednak do zrównania testów potrzebny jest jedynie rozkład wyniku sumarycznego. Konkretnie, dla wyznaczenia funkcji zrównującej (1) potrzeba prawdopodobieństw określających dystrybuanty:

$$F_{Y|Q}(y) = \sum_{i=1}^{n_Y} \mathbb{P}(Y = y|Q), \quad (4a)$$

$$F_{X|Q}(x) = \sum_{i=1}^{n_X} \mathbb{P}(X = x|Q). \quad (4b)$$

Uzyskanie prawdopodobieństw $P(Y=y|Q)$ oraz $P(X=x|Q)$ z rozkładów wektorów odpowiedzi:

$$\mathbb{P}(\boldsymbol{u_Y}|Q) = \int \mathbb{P}(\boldsymbol{u_Y}|\theta, \boldsymbol{\beta_Y}) N(\mu_Q, \sigma_Q^2)\, d\theta, \quad (5a)$$

$$\mathbb{P}(\boldsymbol{u_X}|Q) = \int \mathbb{P}(\boldsymbol{u_X}|\theta, \boldsymbol{\beta_X}) N\left(\mu_Q, \sigma_Q^2\right) d\theta, \quad (5b)$$

oznacza rozpatrzenie wszystkich takich wektorów odpowiedzi u_y oraz u_x, które dają sumaryczny wynik y oraz x. Rekursywny algorytm pozwalający na rozwiązanie tego kombinatorycznego problemu został podany przez Lorda i Wingersky'ego (1983). Alternatywnie, w celu uniknięcia konieczności numerycznego całkowania skomplikowanych funkcji można na bazie parametrów modelu IRT wygenerować wystarczająco dużą liczbę wektorów odpowiedzi u_y oraz u_x i zliczyć proporcję wektorów dających określone wyniki sumaryczne (Glas & Béguin, 1996).

Ostatecznie, wyzwaniem, z jakim trzeba się zmierzyć, chcąc zrównać ze sobą testy X oraz Y z wykorzystaniem IRT, jest umieszczenie wszystkich parametrów modelu, β_y, β_x, μ_Q, σ^2_Q, μ_P, σ^2_P na wspólnej skali. W planie NEAT koniecznej do tego celu informacji dostarcza link utworzony przez test kotwiczący A rozwiązywany przez uczniów z obu populacji. W literaturze można znaleźć różne podejścia do tego problemu, z których najbardziej prominentne to:

1. Kalibracja łączna (ang. *concurrent callibration*), która wykorzystuje zdolność procedury stosowanej do estymacji parametrów IRT do radzenia sobie z brakiem danych występującym w modelu NEAT. Parametry β_A,β_y, β_x, μ_Q, σ^2_Q, μ_P, σ^2_P są oszacowane jednocześnie i znajdują się na wspólnej skali bez konieczności stosowania dodatkowych przekształceń.
2. Kalibracja oddzielna (ang. *separate callibration*), w której w pierwszym kroku dla każdej populacji niezależnie szacowane są parametry testów rozwiązywanych przez uczniów z nich pochodzących. Następnie, uzyskane w pierwszym kroku zestawy parametrów – $\beta_{A|P}$, $\beta_{X|P}$, oraz $\beta_{A|Q}$, $\beta_{Y|Q}$ i arbitralnie ustalone dla identyfikacji modelu wartości μ oraz σ^2 sprowadza się

do wspólnej skali. Sprowadzenie do wspólnej skali opiera się na fakcie, że $\beta_{A|P}$ i $\beta_{A|Q}$ są parametrami tego samego testu i powinny w wyniku przeskalowania mieć takie same wartości. Przeskalowanie parametrów odbywa się poprzez liniową funkcję, której parametry można uzyskać wieloma sposobami. Najważniejsze z nich to metody: średnia/średnia, średnia/sigma, metoda Haebary, metoda Stockinga-Lorda – ich opis można znaleźć u Kolen & Brennan (2004).

3. Metoda ustalonych parametrów (ang. *fixed parameters*) wymaga również przeprowadzenia dwóch kalibracji. Przy czym w pierwszym kroku kalibruje się test i kotwicę tylko w jednej z populacji, załóżmy *P*, uzyskując parametry $\beta_{A|P}$, $\beta_{X|P}$, μ_P, σ^2_P. Następnie, przy kalibracji testu i kotwicy w populacji *Q*, parametry kotwicy ustala się na wartościach z pierwszego kroku, $\beta_{A|Q}= \beta_{A|P}$. W drugim kroku z danych estymuje się zatem jedynie parametry $\beta_{Y|Q}$. W ten sposób wszystkie parametry $\beta_A, \beta_y, \beta_x, \mu_Q, \sigma^2_Q, \mu_P, \sigma^2_P$ znajdują się na wspólnej skali, ustalonej w kroku pierwszym.

Wymienione metody sprowadzania parametrów IRT do wspólnej skali dla planu NEAT były wielokrotnie porównywane poprzez symulacje. Badania Hanson & Béguin (1999) oraz Kang & Petersen (2009) wskazują na dużą zgodność między wymienionymi metodami z wyjątkiem kalibracji oddzielnej z przeskalowaniem średnia/średnia lub średnia/sigma (wypadały gorzej). Oczywiście wnioski wyciągane z takich badań są ograniczone do konkretnych warunków eksperymentalnych (wielkość próby, parametry testu itp.) lub szczegółów implementacji algorytmów szacujących parametry IRT w użytym oprogramowaniu.

Przykład zrównania wyników z wykorzystaniem kalibracji łącznej

Model IRT dla planu NEAT stanowi jednoznaczny opis wszystkich probabilistycznych właściwości testów *A*, *X*, *Y* w obu analizowanych populacjach. Oznacza to, że przyjmując określone wartości dla parametrów $\beta_A, \beta_y, \beta_x, \mu_Q, \sigma^2_Q, \mu_P, \sigma^2_P$ możemy generować odpowiedzi uczniów wylosowanych z *P* lub *Q* na dowolny z trzech testów *A*, *X*, *Y*. Stwarza to możliwość przeprowadzenia eksperymentu Monte Carlo sprawdzającego efektywność danej metody zrównywania wyników testowych w planie NEAT. W dalszej części, jako ilustracja możliwości zrównywania wyników obserwowanych z wykorzystaniem IRT, opisane zostaną wyniki takiego eksperymentu.

Zmienne niezależne w przeprowadzonym eksperymencie były dwie i każda z nich przyjmowała dwa poziomy:

1. wielkość próby: N =1 000 lub N =10 000 uczniów z każdej populacji;
2. średnia θ w populacji Q:μ_Q=0.4 lub μ_Q=0.8.

Ustalonymi warunkami w eksperymencie były:

1. parametry $\beta_A, \beta_y, \beta_x$, opisujące właściwości testów w modelu 2PLM (2): tab. 2;
2. rozkład θ w populacji P: μ_P=0 oraz σ^2_P=1;
3. wariancja θ w populacji Q: σ^2_Q=1;
4. metoda sprowadzania parametrów modelu do wspólnej skali: łączna kalibracja z wykorzystaniem programu MIRT (Glas, 2010);
5. liczba replikacji dla każdej kombinacji poziomów zmiennych. niezależnych: R = 150.

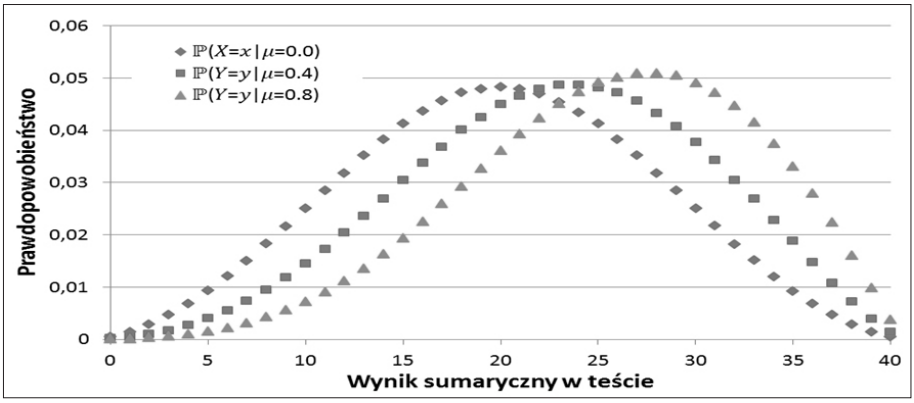
Jako zmienną zależną przyjęto dobroć oszacowania rozkładu testu *X* populacji *Q* w porównaniu do dobroci oszacowania rozkładu testu *Y* w populacji *Q*, gdzie dobroć rozumiana jest zarówno jako wariancja, jak i obciążenie estymatora.

Zaproponowane zrelatywizowanie miary efektywności zrównywania wymaga pewnego komentarza. Rozkład testu *X* populacji *Q* oraz rozkład testu *Y* w populacji *Q* zostały określone dystrybuantami (4b) oraz (4a). Oba rozkłady są konieczne do wyznaczenia funkcji zrównującej (1), jednak kwintesencja procesu zrównywania dotyczy pierwszego z wspomnianych rozkładów, tj. $F_{X|Q}$ – żaden uczeń z populacji *Q* faktycznie nie rozwiązuje ani jednego zadania z testu *X*. Koncentrując się jedynie na dobroci oszacowania rozkładu $F_{X|Q}$, zgubilibyśmy jednak istotny kontekst, mianowicie, że szacowany z ograniczonej próby rozkład $F_{Y|Q}$ też jest obarczony błędem. Jakość oszacowania $F_{Y|Q}$ stanowi kryterium mówiące o tym, jak dobrze można oszacować rozkład odpowiedzi, gdy dane są bezpośrednio dostępne i dopiero przy uwzględnieniu tej informacji możemy stwierdzić, o ile gorsze jest pośrednie (poprzez wykorzystanie łącza zadaniami kotwiczącymi *A*) oszacowanie rozkładu $F_{X|Q}$.

Tabela 2. Parametry zadań użyte do generowania danych symulacyjnych

	zad.	a_i	b_i	zad.	a_i	b_i	zad.	a_i	b_i	zad.	a_i	b_i
Parametry dla testów *X* oraz *Y*	1	1	-2.0	11	1	-1.0	21	1	0.2	31	1	1.2
	2	1.5	-2.0	12	1.5	-1.0	22	1.5	0.2	32	1.5	1.2
	3	1	-1.8	13	1	-0.8	23	1	0.4	33	1	1.4
	4	1.5	-1.8	14	1.5	-0.8	24	1.5	0.4	34	1.5	1.4
	5	1	-1.6	15	1	-0.6	25	1	0.6	35	1	1.6
	6	1.5	-1.6	16	1.5	-0.6	26	1.5	0.6	36	1.5	1.6
	7	1	-1.4	17	1	-0.4	27	1	0.8	37	1	1.8
	8	1.5	-1.4	18	1.5	-0.4	28	1.5	0.8	38	1.5	1.8
	9	1	-1.2	19	1	-0.2	29	1	1.0	39	1	2.0
	10	1.5	-1.2	20	1.5	-0.2	30	1.5	1.0	40	1.5	2.0
Parametry dla testu *A*	1	1.5	-2.0	6	1.5	-1.0	11	1.5	0.2	16	1.5	1.2
	2	1.5	-1.8	7	1.5	-0.8	12	1.5	0.4	17	1.5	1.4
	3	1.5	-1.6	8	1.5	-0.6	13	1.5	0.6	18	1.5	1.6
	4	1.5	-1.4	9	1.5	-0.4	14	1.5	0.8	19	1.5	1.8
	5	1.5	-1.2	10	1.5	-0.2	15	1.5	1.0	20	1.5	2.0

Należy również skomentować „podstępny" dobór parametrów dla testów *X* oraz *Y* widoczny w tabeli 2. Oba testy mają po 40 zadań o identycznych parametrach, zatem testy *X* oraz *Y* są sobie w pełni równoważne. W tej sytuacji żadne zrównywanie nie jest faktycznie potrzebne – uczeń z populacji *Q* w teście *X* powinien dostać dokładnie tyle punktów, ile dostał w teście *Y*, a rozkład $F_{X|Q}$ jest dokładnie taki sam jak rozkład $F_{Y|Q}$. Nie znając prawdziwych wartości parametrów, jednak tego nie wiemy, a ponieważ średnie $\mu_{\theta|P}$ oraz $\mu_{\theta|Q}$ są różne, na poziomie obserwacyjnym uzyskamy różne rozkłady wyników sumarycznych w *X* oraz *Y* – rozkłady testu *X* dla $\mu_{\theta|P}=0$ oraz testu *Y* dla $\mu_{\theta|Q}=0.4$ lub $\mu_{\theta|Q}=0.8$ ukazano na rys. 2. Do stwierdzenia równoważności testów *X* oraz *Y* trzeba zastosować procedurę zrównującą wyniki.



Rysunek 2. Rozkłady wyników w testach *X* oraz *Y* w zależności od μ_θ

Ustalenie identycznych parametrów dla testów X oraz Y ma natomiast bardzo duże znaczenie w kontekście zrelatywizowanej oceny dobroci oszacowania $F_{X|Q}$ względem dobroci oszacowania $F_{Y|Q}$. Bez identyczności testów X oraz Y ewentualne różnice między dobrocią oszacowania $F_{X|Q}$ oraz $F_{Y|Q}$ mogłyby wynikać nie tylko z tego, czy rozkład jest szacowany bezpośrednio/pośrednio, ale także z różnic w psychometrycznych właściwościach dwóch testów.

Omówienie wyników przeprowadzonych symulacji rozpoczniemy od zanalizowania dokładności, z jaką zastosowana metoda IRT potrafi oszacować średnie wyników sumarycznych: $\mu_{X|Q}$ oraz $\mu_{Y|Q}$. Jeżeli oznaczymy poprzez $\hat{\mu}^r_{X|Q}$ oszacowanie średniej $\mu_{X|Q}$ uzyskane w replikacji *r*, to następująca średnia:

$$\overline{\hat{\mu}_{X|Q}} = \frac{\sum_{r=1}^{R} \hat{\mu}^r_{X|Q}}{R},$$

stanowi oszacowanie Monte Carlo wartości oczekiwanej rozpatrywanego estymatora $\hat{\mu}_{X|Q}$. Oszacowanie Monte Carlo wariancji estymatora $\hat{\mu}_{X|Q}$ obliczane jest w następujący sposób:

$$s^2\left(\hat{\mu}_{X|Q}\right) = \frac{\sum_{r=1}^{R}\left(\hat{\mu}^r_{X|Q} - \overline{\hat{\mu}_{X|Q}}\right)^2}{R-1}.$$

Analogicznie można określić $\overline{\hat{\mu}_{Y|Q}}$ oraz $s^2\hat{\mu}_{Y|Q}$.

Otrzymane średnie i wariancje estymatorów $\hat{\mu}_{X|Q}$ oraz $\hat{\mu}_{Y|Q}$ zebrano w tab. 3. Wspomniana tabela zawiera również prawdziwe wartości $\mu_{X|Q}$ oraz $\mu_{Y|Q}$ wyliczone na podstawie założonych parametrów – odpowiadają one rozkładom na rys. 2. Można również zauważyć, że w konsekwencji równoważności testów mamy $\mu_{X|Q}=\mu_{Y|Q}$ przy ustalonej wartości $\mu_{\theta|Q}$.

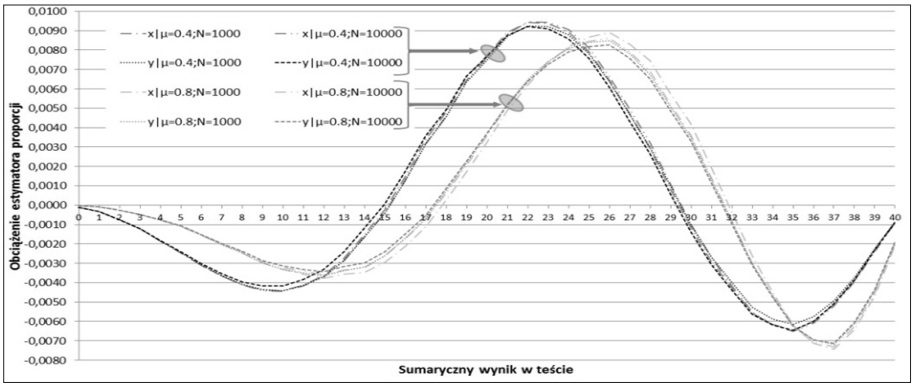
Dla wszystkich warunków eksperymentalnych oszacowane średnie $\overline{\hat{\mu}_{X|Q}}$ oraz $\hat{\mu}_{Y|Q}$ sugerują brak obciążenia estymatorów do pierwszego miejsca po przecinku. Najsilniejsze sygnały o występowaniu obciążenia na drugim miejscu po przecinku pojawiają się, paradoksalnie, przy oszacowaniu $\mu_{Y|Q}$, czyli dla średniej w teście

faktycznie administrowanym uczniom z Q i to przy większej próbie N=10 000. Analizując rozproszenie estymatora, widzimy, że wariancja $s^2(\hat{\mu}_{X|Q})$ dla N=1 000 jest tylko nieznacznie większa niż wariancja $s^2(\hat{\mu}_{Y|Q})$, a dla próby N=10 000 stosunek tych wariancji jest poniżej 2 : 1. Uwzględniając, że między N=1 000 a N=10 000 wariancje różnią się o ponad rząd wielkości, należy stwierdzić, że wariancja oszacowania średniej w teście zależy przede wszystkim od wielkości próby, a nie od tego, czy test był lub nie był w danej populacji rozwiązywany. Ogólne spostrzeżenie jest takie, że zastosowana procedura zrównywania pozwala dla uczniów z populacji Q oszacować średnią w teście nierozwiązywanym przez nich ($\mu_{X|Q}$) praktycznie równie dobrze jak w teście faktycznie przez nich rozwiązywanym ($\mu_{Y|Q}$).

Tabela 3. Oszacowania średniego wyniku sumarycznego w testach X oraz Y w populacji Q w zależności od warunków eksperymentalnych (opis w tekście)

N	$\mu_{\theta\|Q}$	$\mu_{X\|Q}$	$\overline{\hat{\mu}_{X\|Q}}$	$s^2(\hat{\mu}_{X\|Q})$	$\mu_{Y\|Q}$	$\overline{\hat{\mu}_{Y\|Q}}$	$s^2(\hat{\mu}_{Y\|Q})$
1 000	0.4	22.87	22.84	0.171	22.87	22.88	0.160
	0.8	25.67	25.71	0.230	25.67	25.67	0.132
10 000	0.4	22.87	22.86	0.016	22.87	22.80	0.012
	0.8	25.67	25.66	0.017	25.67	25.64	0.009

Założonym celem symulacji było sprawdzenie zdolności procedury zrównywania wyników do oszacowania całego rozkładu testu X w populacji Q, a nie tylko jego średniej. Analogicznie do $\overline{\hat{\mu}_{X|Q}}$ oraz $\overline{\hat{\mu}_{Y|Q}}$ obliczono dla każdego wyniku x w teście X oraz y w teście Y oszacowanie Monte Carlo wartości oczekiwanej estymatorów proporcji: $\hat{p}_{x|Q}$ oraz $\hat{p}_{y|Q}$ (gdzie $p_{x|Q}$ oraz $p_{y|Q}$ są skróconym zapisem prawdopodobieństw pojawiających się z prawej strony wzorów na dystrybuanty (4a) i (4b) lub na wykresie na rys. 2. Podobnie obliczono oszacowania Monte Carlo wariancji estymatorów proporcji: $s^2(\hat{p}_{x|Q})$ oraz $s^2(\hat{p}_{y|Q})$ Od proporcji $\hat{p}_{x|Q}$ oraz $\hat{p}_{y|Q}$ odjęto prawdziwe (wynikające z założonych parametrów β_y, β_x, μ_Q, σ^2_Q) wartości $\hat{p}_{x|Q}$ oraz $\hat{p}_{y|Q}$ (rys. 2.), uzyskując oszacowania Monte Carlo obciążenia estymatorów proporcji: $B(\hat{p}_{x|Q})$ oraz $B(\hat{p}_{y|Q})$. Uzyskane obciążenia estymatorów proporcji zebrano na wykresie na rys. 3., a wariancje estymatorów proporcji zebrano na wykresie na rys. 4.



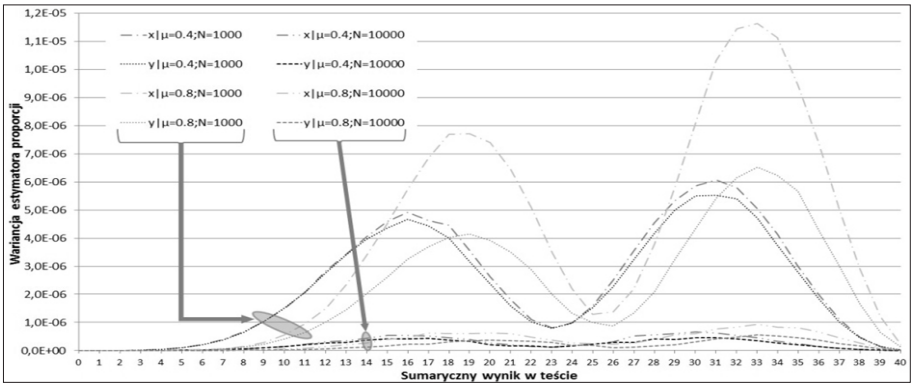
Rysunek 3. Obciążenie estymatorów $\hat{p}_{x|Q}$ oraz $\hat{p}_{y|Q}$ w zależności od wartości wyniku

sumarycznego dla różnych warunków eksperymentalnych

Wykres z obciążeniami (rys. 3.) potwierdza obserwację poczynioną dla średniej (tab. 3.), że dla ustalonej wartości $\mu_{\theta|Q}$ takie czynniki jak wielkość próby oraz to, czy test był, czy nie był faktycznie rozwiązywany przez uczniów z Q nie miały praktycznie żadnego znaczenia dla obciążenia estymatora. Cztery analizowane krzywe obciążeń dla $\mu_{\theta|Q}$=0.04 praktycznie pokrywają się ze sobą, tak samo dla $\mu_{\theta|Q}$=0.08.

Bardzo interesujące jest występowanie istotnego obciążenia estymatora, zmieniającego się w zależności od wyniku sumarycznego. Obciążenie zmienia dwukrotnie swój kierunek – dla najniższych i najwyższych wyników w teście proporcje są zaniżone, natomiast dla wyników przeciętnych są zawyżone. Maksimum zawyżenia proporcji przypada w rejonie średniego wyniku sumarycznego w teście, tj. (tab. 3.) w rejonie 22.87 dla $\mu_{\theta|Q}$=0.04 oraz w rejonie 25.67 dla $\mu_{\theta|Q}$=0.08 – stąd dwie wiązki krzywych na rys. 3. są przesunięte względem siebie. Lokalnie znaczne obciążenie umknęło przy analizie zdolności procedury zrównywania do oszacowania średniej w teście (tab. 3.), właśnie ponieważ obciążenie proporcji zmienia swój znak i znosi się przy obliczaniu średniej. Wykres obciążeń sugeruje, że oszacowana za pomocą programu MIRT wariancja $\mu^2_{\theta|Q}$ była ujemnie obciążona – rozkład θ w populacji Q legł w procesie estymacji zbytniemu „skurczeniu" względem parametrów β_y oraz β_x i uzyskaliśmy zbyt dużą koncentrację uczniów o przeciętnych wynikach z niedoszacowaniem udziału uczniów o wynikach skrajnych.

Wykres z wariancjami (rys. 4.) natomiast potwierdza, że wariancja oszacowań zależy przede wszystkim od wielkości próby (por. tab. 3.). Dla warunku N=10 000 mamy wiązkę czterech wariancji niewykraczających ponad wartość $1{\cdot}10^6$, gdy dla warunku N=1 000 odpowiednie wariancje są od kilku do kilkunastu razy większe. W dalszej kolejności istotnym czynnikiem wpływającym na wielkość wariancji oszacowań proporcji jest, czy dany test był administrowany w populacji, czy też proporcje obliczono na podstawie łącza stworzonego przez test kotwiczący$A$. Dla ustalonego N wariancje $s^2(\hat{p}_{x|Q})$ były dla większości punktów sumarycznych większe od odpowiadających im $s^2(\hat{p}_{y|Q})$ – to też jest zgodne z obserwacjami poczynionymi dla średniej w całym teście (tab. 3.).



Rysunek 4. Wariancja estymatorów $\overline{\hat{p}_{x|Q}}$ oraz $\overline{\hat{p}_{y|Q}}$ w zależności od wartości wyniku sumarycznego dla różnych warunków eksperymentalnych

Obserwujemy także specyficzną zmienność wariancji w zależności od wyniku sumarycznego: najniższe wariancje przypadają dla wyników skrajnych oraz dla wyników zbliżonych do średniego wyniku w danym teście. Oba zjawiska łatwo można wytłumaczyć poprzez odwołanie się do klasycznego wzoru na wariancję estymatora proporcji: $p(1-p)/n$. Na skrajach rozkładu wyników sumarycznych proporcje są bardzo małe, co się przekłada na to, że $p(1-p)$ przyjmuje niewielkie wartości. W centrum rozkładu wyrażenie $p(1-p)$ jest relatywnie większe, jednak znajduje się tam również dużo wyników (n) powodujących, że proporcja jest oszacowana z większą precyzją.

Wnioski

Z przeprowadzonego symulacyjnego badania właściwości zrównywania z wykorzystaniem kalibracji łącznej modelu IRT można dojść do następujących konkluzji:

1. Zastosowana metoda zrównywania szacuje parametry rozkładu testu X uczniów z populacji Q w planie NEAT (tab. 1.) porównywalnie dobrze do tego, jak szacuje parametry rozkładu testu Y dla tej samej populacji.
2. Ze względu na obciążenie oszacowań największej wiarygodności parametrów modelu IRT (Lord, 1983) oszacowania proporcji wyników sumarycznych również wykazują obciążenie. Na poziomie wyniku średniego obciążenia się znoszą, jednak może mieć to istotne znaczenie na poziomie konkretnych sumarycznych wyników, czyli na poziomie, na którym liczona jest funkcja zrównująca dana wzorem (1). Potencjalne niebezpieczeństwo obciążenia samej funkcji zrównującej może jednak być załagodzone przez fakt, że obciążenia dla rozkładu testu Y mają podobny wzór jak obciążenia dla rozkładu X i mimo występujących obciążeń $F_{Y|Q}$ oraz $F_{X|Q}$ samo przekształcenie zrównujące może ${}^{(cont)}F_{X|Q}^{-1}\left({}^{(cont)}F_{Y|Q}(y)\right)$ być nieobciążone. Problem wymaga dalszych badań.
3. Oszacowania proporcji sumarycznych wyników wykazują zmienną wariację. Oznacza to, że błąd zrównywania powinien być raportowany w sposób warunkowy.

Bibliografia:

1. Dorans, N. J. & Moses, T. P. & Eignor, D. R. (2011). *Equating Test Scores: Toward Best Practices*. In von Davier, A. A. (Ed.), *Statistical Models for Test Equating, Scaling, and Linking* (pp. 21-42). New York, NY: Springer-Verlag.
2. Glas C. A. (2010). *Preliminary Manual of the software program Multidimensional Item Response Theory (MIRT)*. (University of Twente)
3. Glas C. A. W. & Béguin A. A. (1996). *Appropriateness of IRT Observed-Score Equating* (Research Report 1996-2).
4. Hanson, B. A. & Béguin A. A. (1999). *Separate Versus Concurent Estimation of IRT Item Parameters in the Common Item Equating Design* (ACT Research Report Series, 1999-8). Iowa City, IA: ACT, Inc.

5. Kang, T., Petersen N. (2009). *Linking Item Parameters to a Base Scale* (ACT Research Report Series, 2009-2). Iowa City, IA: ACT, Inc.
6. Kolen, M. J., & Brennan R. L. (2004). *Test equating, scaling, and linking: Method and practice* (2nd ed.). New York, NY: Springer-Verlag.
7. Kondratek B. (2007). *Teoria odpowiadania na pozycje testowe oraz klasyczna teoria testów. Porównanie w kontekście modelowania statystycznego sytuacji eksperymentalnej badania testem*. *Egzamin*, 9, 76-104.
8. Lord, F. M. (1983). *Statistical bias in maximum likelihood estimators of item parameters*. *Psychometrika*, 48 (3), 425–435.
9. Lord, F. M. & Novick, M . R. (1968). *Statistical Theories of Mental Test Scores*. Reading, Massachusetts: Addison-Wesley.
10. Lord, F. M., Wingersky, M. S. (1983). *Comparison of IRT Observed-Score and True-Score "Equatings"*. ETS Research Report: ETS-RR-83-26-ONR.
11. Reckase, M. D. (2009). *Multidimensional Item Response Theory*. New York, NY: Springer-Verlag.
12. von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating*. New York, NY: Springer-Verlag.