Edukacja 2014, 3(128), 112–130
ISSN 0239-6858

Zrównanie ekwicentylowe na tle innych metod zrównywania na przykładzie sprawdzianu i egzaminu gimnazjalnego

Tymoteusz Wołodźko, Bartosz Kondratek, Henryk Szaleniec
Instytut Badań Edukacyjnych*

Artykuł przedstawia wyniki zrównania ekwicentylowego wyników trzech testów: sprawdzianu oraz części matematyczno-przyrodniczej i części humanistycznej egzaminu gimnazjalnego z lat 2002–2012. W latach 2011–2014 przeprowadzone zostały cztery sesje, podczas których uczniowie z reprezentatywnej próby polskich szkół rozwiązywali arkusze zadań pochodzących ze sprawdzianu w szóstej klasie szkoły podstawowej i egzaminów gimnazjalnych, w warunkach możliwie zbliżonych do rzeczywistej sesji egzaminacyjnej. Dane te posłużyły do oszacowania funkcji zrównujących, które zostały wykorzystane do zrównania wyników rzeczywistych egzaminów. Zrównania przeprowadzone za pomocą metody ekwicentylowej, zrównania liniowego i metod wywodzących się z *item reponse theory*, dały zbliżone wyniki. Uzyskane rezultaty omówione zostały w kontekście planowania zrównań testów.

Słowa kluczowe: psychometria, zrównywanie wyników obserwowanych, zrównywanie ekwicentylowe.

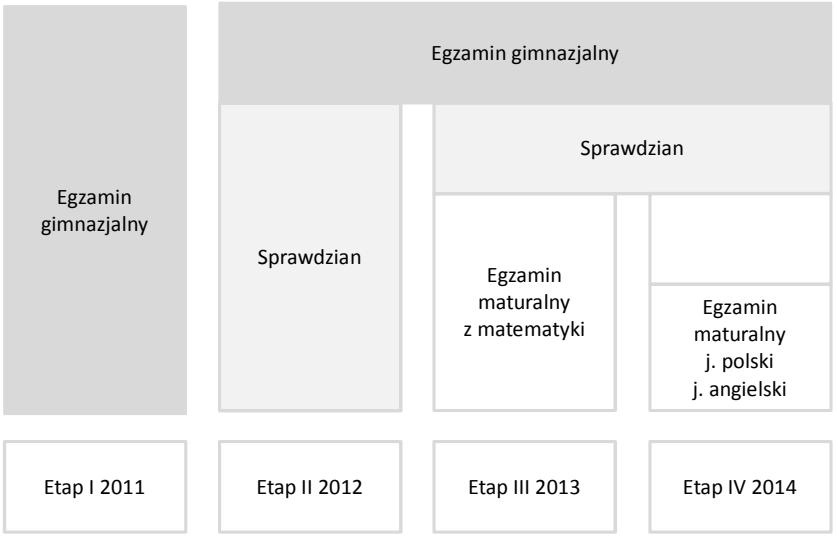
Wyniki testów często wykorzystywane są jako punkt odniesienia dla ważnych decyzji podejmowanych zarówno na poziomie indywidualnym, instytucjonalnym, jak i politycznym (Kolen i Brennan, 2004). Istotne jest więc, aby były one porównywalne nie tylko w skali kraju w danym roku, ale także między latami. Ma to szczególne znaczenie wtedy, gdy absolwenci danego etapu edukacyjnego z różnych lat biorą jednocześnie udział w tej samej rekrutacji do szkoły wyższego szczebla.

Na świecie przy zrównywaniu obserwowanych wyników egzaminacyjnych wykorzystuje się zarówno metody nieodwołujące się do żadnego specyficznego modelu pomiarowego, zwane klasycznymi, jak i oparte na modelowaniu *item response theory* (IRT). Samo zaś zrównywanie stosowane jest co najmniej od lat 40. XX w. (Holland, 2007)[1]. Chociaż system egzaminów zewnętrznych

Artykuł powstał w ramach projektu systemowego „Badanie jakości i efektywności edukacji oraz instytucjonalizacja zaplecza badawczego" finansowanego ze środków Europejskiego Funduszu Społecznego w ramach Programu Operacyjnego Kapitał Ludzki (Priorytet III: Wysoka jakość systemu oświaty, Poddziałanie 3.1.1. Tworzenie warunków i narzędzi do monitorowania, ewaluacji i badań systemu oświaty).

[1] Przegląd zastosowań zrównywania można znaleźć w publikacjach Artura Pokropka i Bartosza Kondratka (2012) i Artura Pokropka (2011). Zrównywaniu poświęcony był też cały numer „Egzaminu" (Lisiecka i Szaleniec, 2007), w którym znalazły się artykuły zarówno na temat metod klasycznych, jak i opartych na IRT.

* Adres do korespondencji: ul. Górczewska 8, 01-180 Warszawa. E-mail: t.wolodzko@ibe.edu.pl



Rysunek 1. Egzaminy włączone do badań w poszczególnych etapach studium zrównującego.

został w Polsce wprowadzony w 2002 r., do tej pory nie wdrożono rozwiązań umożliwiających porównywanie wyników egzaminów zewnętrznych w kolejnych latach, dających możliwość kontrolowania zmian w poziomie trudności egzaminów. Oznacza to, że wyniki oparte na skalach corocznie komunikowanych przez Centralną Komisję Egzaminacyjną (CKE), nie są ekwiwalentne. Brak porównywalności wyników uczniów zdających egzaminy w różnych latach w kontekście kwalifikacji do dalszych etapów kształcenia można postrzegać jako dyskryminujący dla uczniów zdających trudniejszą wersję egzaminu bez odpowiedniego zrównania względem uczniów zdających wariant łatwiejszy.

Pierwsze próby zrównania wyników prowadzone w ramach projektów badawczych CKE (mające charakter eksperymentalny), wykorzystujące zarówno metody klasyczne, jak i z zastosowaniem modeli IRT, rozpoczęły się już dwa lata po wdrożeniu egzaminów (Niemierko, 2004; 2007; Szaleniec, 2005; 2007; Smolik; 2007). Nie doprowadziły one jednak do rozwiązań systemowych i wdrożenia zrównywania wyników równolegle z prowadzonymi egzaminami w danej sesji. Systematyczne badania mające na celu zrównanie wyników egzaminacyjnych sprawdzianu na zakończenie szkoły podstawowej, egzaminu gimnazjalnego w części humanistycznej i matematyczno-przyrodniczej na zakończenie III klasy gimnazjum (wyniki z lat 2002––2014) oraz egzaminu maturalnego z języka polskiego, języka angielskiego i matematyki (wyniki z lat 2010–2014) zostały podjęte w 2010 r. przez Zespół Analiz Osiągnięć Uczniów Instytutu Badań Edukacyjnych (ZAOU IBE; Szaleniec i in., 2012; 2013). Studium badawcze, które w 2014 r. dobiegnie końca, obejmuje 4 etapy. Pierwszy dotyczył tylko egzaminu gimnazjalnego, podczas gdy kolejne obejmowały dodatkowe egzaminy, kontynuując bieżące zrównanie dla egzaminu gimnazjalnego i sprawdzianu z poprzedniego etapu. Rysunek 1 przedstawia kumulacyjny charakter poszczególnych etapów studium zrównującego.

Badania w ramach studium zostały przeprowadzone na próbach reprezentatywnych

dla kraju, z zastosowaniem schematu losowania podwójnie warstwowego: proporcjonalnego do liczebności uczniów w oddziałach w szkole, zespołowego oraz wielostopniowego. Rezultaty z pierwszych dwóch etapów zrównania egzaminów gimnazjalnych i sprawdzianu uzyskane metodą łącznej kalibracji modelu IRT (Szaleniec i in., 2012; 2013) zostały wykorzystane w szczególności do doskonalenia metodologii zastosowanej w kolejnych dwóch etapach: w 2013 i 2014 r.

Przedstawione tutaj analizy są próbą zastosowania klasycznych metod zrównywania, w szczególności metody ekwicentylowej, do danych zgromadzonych w trakcie studium, które w swoich założeniach skupiało się na zastosowaniu modeli IRT. Z jednej strony przeprowadzone analizy dokonują walidacji rezultatów uzyskanych wcześniej na podstawie modelowania IRT, z drugiej strony stanowią niezależny materiał do dyskusji nad różnymi aspektami istotnymi przy implementacji zrównywania wyników obserwowanych w skomplikowanym schemacie badawczym.

Artykuł jest zorganizowany w taki sposób, aby czytelnik mógł najpierw zapoznać się z klasycznymi metodami zrównywania na podstawie przeglądu literatury, a następnie z przedstawionymi rezultatami zrównania wyników sprawdzianu i egzaminu gimnazjalnego z zastosowaniem metody ekwicentylowej, porównaniem wyników zrównania ekwicentylowego z wynikami uzyskanymi metodą zrównania liniowego oraz wynikami uzyskanymi z zastosowaniem modeli IRT. Całość kończy dyskusja na temat konsekwencji zastosowania różnych planów zrównywania dla uzyskiwanych wyników.

Klasyczne metody zrównywania wyników obserwowanych

Można wyróżnić cztery metody zrównywania klasycznego: zrównywanie na podstawie średnich (*mean equating*), zrównywanie liniowe (*linear equating*), zrównywanie ekwicentylowe (*equipercentile equating*) oraz metoda *circle-arc* (Livingston i Kim, 2009).

Zanim omówimy klasyczne metody zrównywania, zacznijmy od sytuacji braku zrównania. Jeśli wyniki dwóch testów są takie same, nie jest potrzebne zrównanie, a ich relację odzwierciedla funkcja tożsamościowa:

$$e_Y^{(ident)}(x) = x\,, \tag{1}$$

gdzie e_Y jest ogólnym oznaczeniem dla funkcji zrównującej wyniki obserwowane testu X na skalę wyników obserwowanych testu Y. Stworzenie dwóch testów, lub dwóch wersji testu, w których uczniowie uzyskiwać będą identyczne wyniki, nie jest jednak w praktyce możliwe. Stąd powstaje potrzeba zrównywania wyników testowych. Funkcja tożsamościowa jest tu wspominana, ponieważ może być punktem odniesienia dla porównywania metod zrównywania. Odzwierciedla ona też podejście, z którym często możemy spotkać się na co dzień, gdy porównywane są ze sobą surowe wyniki egzaminów zewnętrznych z dwóch różnych sesji.

Zrównanie liniowe. Metoda ta polega na sprowadzeniu wyniku testu do skali standardowej, a następnie na przekształceniu go na podstawie średniej i odchylenia standardowego do skali drugiego testu. Można je zastosować do wyników dwóch rozwiązywanych przez równoważne sobie populacje uczniów testów X przyjmującego wartość x i Y przyjmującego wartości y, które różnią się średnią, ale różnica ta nie jest stała wzdłuż całej skali. Zrównanie liniowe wykorzystuje i ujednolica dwa parametry rozkładu, jakimi są średnia μ i odchylenie standardowe σ. Opiera się ono na następującej zależności:

$$\frac{x-\mu_X}{\sigma_X} = \frac{y-\mu_Y}{\sigma_Y},$$

z czego możemy wyprowadzić wzór zrównania liniowego, testu X na skalę testu Y:

$$e_Y^{(lin)}(x) = \frac{\sigma_Y}{\sigma_X} x + \left(\mu_Y - \frac{\sigma_Y}{\sigma_X} \mu_X \right). \qquad (2)$$

W szczególnym przypadku, gdy odchylenia standardowe z obydwu testów będą identyczne, wzór na zrównanie liniowe redukuje się do postaci:

$$e_Y^{(mean)}(x) = x - \mu_X + \mu_Y,$$

którą określa się jako zrównanie metodą średniej (Kolen i Brennan, 2004).

Zrównanie ekwicentylowe. Metoda zrównania liniowego w wielu wypadkach jest odpowiednia, np. jeśli zależy nam jedynie na porównaniu pierwszych dwóch momentów rozkładów z dwóch różnych testów na wspólnej skali. Jednak w przypadku wnioskowania na poziomie pojedynczego ucznia przekształcenie zrównujące poprawnie sprowadzające do wspólnej skali średnią i odchylenie standardowe zazwyczaj nie będzie wystarczające, gdyż rozkład dyskretnych obserwowanych wyników testowych nie może być w pełni opisany przez wspomniane dwa parametry. Należy tu zaznaczyć, że wyniki egzaminacyjne często różnią się także w zakresie dalszych momentów, na przykład różnice w zakresie skośności i kurtozy można znaleźć między innymi w raportach z przeprowadzonych badań (Szaleniec i in., 2012; 2013), ale nawet uwzględnienie tych informacji w procedurze zrównywania nie wystarczałoby do pełnego opisu rozkładu wyników obserwowalnych. Metodą, która ma zapewnić poprawne zrównanie testów poprzez odwołanie się do kompletnej informacji o rozkładzie wyników obserwowanych zrównywanych testów, jest podejście nieparametryczne realizowane poprzez zrównanie ekwicentylowe.

Idea zrównywania ekwicentylowego opiera się na obserwacji, że wszystkie wartości punktowe x i y testów X i Y są ekwiwalentne, jeżeli:

$$F_X(x) = u = F_Y(y),$$

gdzie F_X oraz F_Y to dystrybuanty X i Y, a u przyjmuje wartości w zakresie [0, 1]. O równoważności możemy więc mówić, gdy każdej wartości u towarzyszą te same wartości rozkładów wartości punktowych obu zmiennych (Davier, Holland i Thayer, 2004). Opierając się na tym fakcie, możemy zdefiniować funkcję zrównania ekwicentylowego dwóch ciągłych i ściśle rosnących dystrybuant F_X oraz F_Y następująco:

$$Y = F_Y^{-1}\left[F_X(X)\right], \qquad (3)$$

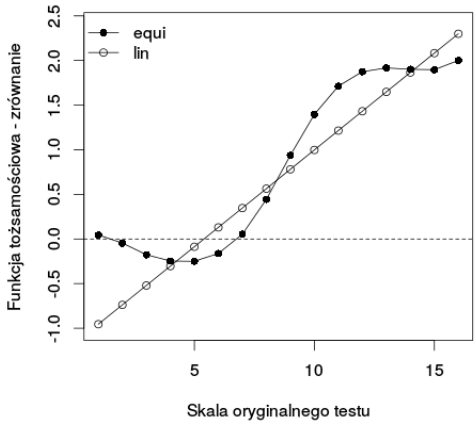
czyli złożenie $F_Y^{-1} \circ F_X$ przekształca zmienną losową X w zmienną losową Y. Niestety dystrybuanty F_X oraz F_Y dla wyników obserwowanych w testach X i Y, ze względu na dyskretność tychże wyników, są funkcjami skokowymi i wzór (3) nie może zostać bezpośrednio zastosowany. Wszystkie ekwicentylowe metody zrównywania wyników obserwowanych zmuszone są do uwzględnienia jakiejś formy odpowiedniego (skutkującego różnowartościowością) uciąglania dystrybuant, co zostało omówione w dalszej części tekstu. Funkcja zrównująca X i Y przedstawiona we wzorze (3), przyjmuje wtedy kształt:

$$e_Y^{(equi)}(x) = \tilde{F}_Y^{-1}\left[\tilde{F}_X(x)\right], \qquad (4)$$

gdzie $\tilde{F}_X$ oznacza uciągloną dystrybuantę X, a $\tilde{F}_Y^{-1}$ jest odwrotną funkcją uciąglonej dystrybuanty Y. Jak w przypadku pozostałych metod zrównywania, funkcja zrównująca (4) jest odwracalna:

$$e_X^{(equi)}(y) = \tilde{F}_X^{-1}\left[\tilde{F}_Y(y)\right]. \qquad (5)$$

Metoda ta sprowadza się więc do przypisania wartościom punktowym testu X takich wartości punktowych testu Y, które występują z takim samym prawdopodobieństwem, jak wartości testu X. Szerzej implikacje



Rysunek 2. Porównanie różnic zrównania ekwicentylowego i liniowego z funkcją tożsamościową.

wynikające z tej definicji omawiają Michael Kolen i Robert Brennan (2004) i Alina von Davier, Paul Holland i Dorothy Thayer (2004).

Na Rysunku 2 zilustrowano różnicę między zrównaniami uzyskiwanymi za pomocą metod zrównywania liniowego i ekwicentylowego przy odniesieniu do funkcji tożsamościowej. Za przykład posłużyły arkusze testowe z 2002 i 2003 r., z części matematyczno-przyrodniczej egzaminu gimnazjalnego, wypełnione przez 877 uczniów w trakcie badań.

Wyniki zrównania liniowego układają się w linii prostej, ponieważ w żaden sposób nie została uwzględniona informacja o tym, że w obu testach poszczególne wartości punktowe występują z różną częstością. Została ona uwzględniona w zrównaniu ekwicentylowym, gdzie oprócz tego, że możemy wnioskować o średnich, możemy także przekształcić poszczególne wartości punktowe zrównywanego testu na ich odpowiedniki w drugim teście.

Etapy zrównania ekwicentylowego

Zrównanie ekwicentylowe jest jednym z elementów szerzej rozumianego procesu zrównania składającego się z pięciu kroków: (1) wygładzania rozkładów, (2) szacowania prawdopodobieństw brzegowych występowania poszczególnych wartości punktowych obu testów, (3) uciąglania rozkładów, (4) zrównywania i (5) oceny błędów zrównania (Davier, 2011; Davier, Holland i Thayer, 2004). Ponieważ wszystkie etapy zostały odpowiednio zaimplementowane, w analizach omawianych w dalszej części artykułu kroki te zostaną omówione pokrótce.

Wstępne wygładzanie rozkładów (*presmoothing*) jest w swojej naturze czysto „technicznym" zabiegiem mającym na celu poprawę parametrów rozkładów dyskretnych zmiennych przez usunięcie zakłóceń losowych, a także eliminację wartości o zerowym prawdopodobieństwie wystąpienia (Davier, Holland i Thayer, 2004; Kolen i Brennan, 2004). Liczne badania pokazały, że wygładzanie rozkładów prowadzi do nieznacznego nasilenia błędów systematycznych przy zauważalnym zmniejszeniu natężenia błędów losowych, ostatecznie prowadząc do zrównań obarczonych mniejszym błędem (Kolen i Brennan, 2004). Współcześnie jest ono powszechnie stosowane, choć może być pominięte, gdy dane wykorzystywane do zrównania pochodzą z dużej próby, w związku z czym ich rozkład

w małym stopniu jest obarczony zakłóceniami wynikającymi z doboru próby (Davier, Holland i Thayer, 2004).

Powszechnie stosowaną metodą wstępnego wygładzania rozkładów jest model log-liniowy dla rozkładu Poissona (Holland i Thayer, 2000; Moses i Davier, 2006). W tym celu do danych dopasowuje się model, w którym zmienną objaśnianą są liczebności dla określonych wartości punktowych, a objaśniającą wartości punktowe. W przypadku rozkładu pojedynczej zmiennej model przyjmuje formę:

$$\log(Np_i) = \beta_0 + \sum_{k=1}^{K} \beta_{xi,k} x_i^k \,, \qquad (6)$$

natomiast dla rozkładu łącznego, tj. wyników z dwóch testów, uwzględnia się także efekt interakcyjny obu zmiennych:

$$\log(Np_{ij}) = \qquad (7)$$
$$= \beta_0 + \sum_{k=1}^{K} (\beta_{xi,k} x_i^k + \beta_{yj,k} y_j^k) + \sum_{l}^{L} \beta_{xi \times yj,l} (x_i y_j)^l ,$$

gdzie N to liczebność próby, a p_{ij} to odsetek przypadków w komórce ij tablicy kontyngencji. Istotną kwestią jest odpowiedni dobór parametrów, jakimi są rzędy wielomianów w modelu K, L. Badania symulacyjne pokazały, że najlepszym kryterium doboru parametrów modelu jest kryterium informacyjne Aikaike (AIC; Moses i Holland, 2009). AIC, podobnie jak kryterium Bayesowskie (BIC), także omawiane w tekście Tima Mosesa i Paula Hollanda. Są one powszechnie stosowane i zaimplementowane w większości programów statystycznych. Niższe wartości tych kryteriów wskazują na lepsze dopasowanie modelu.

Szacowanie prawdopodobieństw brzegowych wartości punktowych następuje na podstawie funkcji planu zrównania (*design function*). Klasycznie wyróżniane są cztery takie plany:

- plan grup równoważnych (*equivalent groups*, EG), w którym uczniowie z dwóch różnych prób rozwiązują dwa testy, przy czym zakłada się, że populacje, z jakich wywodzą się obie grupy, charakteryzują się tym samym poziomem umiejętności,
- plan pojedynczej grupy (*single group*, SG), gdzie ta sama próba rozwiązuje oba testy,
- plan zrównoważony (*counterbalanced design*, CB), to w gruncie rzeczy dwa plany SG, gdzie dwie grupy uczniów rozwiązują oba testy w różnej kolejności, oraz
- plan nierównoważnych grup z testem kotwiczącym (*nonequivalent groups with anchor test*, NEAT), w tym planie biorą udział dwie grupy, pierwsza rozwiązuje test zrównywany oraz tzw. test kotwiczący, a druga ten sam test kotwiczący i test, do którego następuje zrównanie.

NEAT można potraktować jako specyficzną formę połączenia dwóch zrównań par testów w planach SG (metoda *chained equating*, CE), choć istnieją też metody pozwalające na bezpośrednie zrównanie obu testów, jedynie uwzględniając informacje pochodzące z testu kotwiczącego (metody *frequency estimation*, FE). Tematyka ta była szerzej omawiana w innych publikacjach (Kolen i Brennan, 2004; Davier, Holland i Thayer, 2004), oraz na łamach „Edukacji" (Pokropek i Kondratek, 2012), więc nie będzie tu dalej rozwijana.

Uciąglanie rozkładów ma na celu takie przekształcenie dystrybuant zrównywanych testów, by zniwelować ich „schodkową" formę wynikającą z faktu, że wyniki testowe mają charakter dyskretny (Davier, Holland i Thayer, 2004). Klasycznie, w tym celu stosowano liniową interpolację na podstawie rang centylowych (Kolen i Brennan, 2004), współcześnie często łączy się ten etap z wtórnym wygładzaniem rozkładów (*postsmoothing*) i sięga po estymator jądrowy (*kernel smoothing*), z użyciem jądra rozkładu normalnego (Davier, 2007; Davier,

Holland i Thayer, 2004). Estymator jądrowy jest powszechnie stosowany m.in. do wygładzania rozkładów przy wizualizacji danych (Wand i Jones, 1995). By wtórnie wygładzić rozkład testu przyjmującego wartości punktowe x_j, którym towarzyszą prawdopodobieństwa brzegowe r_j, korzystamy z funkcji gęstości prawdopodobieństwa dla rozkładu normalnego, $\Phi(Z)$, ze średnią zero i odchyleniem standardowym jeden:

$$\tilde{F}_X(x) = \hat{F}_{h_X}(x) = \sum_j r_j \Phi\left[R_{jX}(x)\right], \quad (8)$$

gdzie:

$$R_{jX}(x) = \frac{x - a_X x_j - (1 - a_x)\mu_X}{a_X h_X}, \quad (9)$$

gdzie:

$$a_X^2 = \frac{\sigma_X^2}{\sigma_X^2 + h_X^2}. \quad (10)$$

Procedura ta wymaga dodatkowego argumentu w postaci szerokości pasma (*bandwidth*), h_X, który możemy ustalić odgórnie, bądź skorzystać z metod automatycznego doboru jego najlepszej wartości, które zostało szerzej opisane w pracy Davier, Hollanda i Thayer (2004).

Zrównanie. Na tym etapie następuje właściwe zrównanie na podstawie metody ekwicentylowej, omówionej wcześniej. W praktyce ten etap bardzo ściśle łączy się z poprzednim, ponieważ przy przekształcaniu wyników potrzebne jest użycie jakiejś formy interpolacji, zazwyczaj liniowej (Kolen i Brennan, 2004). Nic jednak nie stoi na przeszkodzie, by posłużyć się inną formą interpolacji.

Ocena błędów zrównania (*standard errors of equating*, SEE) jest ostatnim etapem procedury. Istnieją dwie metody szacowania błędów zrównania: analityczna oraz bootstrapowa (Davier, Holland i Thayer, 2004; Kolen i Brennan, 2004;). Użycie pierwszej z nich nie było tu możliwe ze względu na nieprzystający do niej, złożony plan badania i nie będzie szerzej omawiane. Pierwotnie metoda bootstrap została opisana i rozwijana przez Bradleya Efrona, jako sposób szacowania parametrów zmiennych o nieznanym rozkładzie (Davison i Hinkley, 2009; Efron i Tibshirani, 1993). Kolen i Brennan (2004; Wang, 2011) zaadoptowali ją do celów oceny błędów zrównania. *Bootstrap* polega na wielokrotnym zrównywaniu wyników testów pochodzących z próbek losowanych ze zwracaniem z oryginalnych danych użytych w badaniu. W najprostszej postaci algorytm szacowania błędów zrównania testów w schemacie dla jednej grupy prezentuje się w następujący sposób:

1. Z grupy N uczniów losujemy ze zwracaniem próbę liczebności N;
2. Przy użyciu próby z kroku 1 szacujemy zrównane ekwiwalenty wartości punktowych x_i, do których będziemy się odnosić jako $e_{Y_r}^{(equi)}(x_i)$;
3. Kroki 1 i 2 powtarzane są R razy, otrzymując oszacowania:
$\hat{e}_{Y_1}^{(equi)}(x_i), \hat{e}_{Y_2}^{(equi)}(x_i), \ldots, \hat{e}_{Y_R}^{(equi)}(x_i)$.

Plan niniejszego badania wymagał modyfikacji oryginalnej metody, ponieważ poddane analizie dane pochodziły z losowo wybranych oddziałów, z losowej, reprezentatywnej próby szkół na terenie Polski. Oznacza to, że aby trafnie odzwierciedlić wariancję uzyskanych wyników, konieczne było bootstrapowe losowanie z danych w taki sam sposób, w jaki dobrana została próba, czyli na poziomie szkół, a nie indywidualnych uczniów. Oznacza to, że w kroku 1 z próby M szkół, losowane było ze zwracaniem M szkół. Jest to uznany sposób losowania z danych o charakterze hierarchicznym, której trafność potwierdzają wyniki symulacyjne (Davison i Hinkley, 2009; Field i Welsh, 2007; Rena i in., 2010).

Dostępne są trzy miary błędów uzyskiwanych przez bootstrap: obciążenie (*bias*),

błąd standardowy (*SE*) i pierwiastek kwadratowy z błędu średniokwadratowego (RMSE), mierzone na poziomie wartości punktowych x_i (Albano, 2014; Kolen i Brennan, 2004; Wang, 2011):

$$bias\left(x_i\right) = \hat{\bar{e}}_Y^{(equi)}(x_i) - e_Y^{(equi)}(x_i), \qquad (11)$$

gdzie:

$$\hat{\bar{e}}_Y^{(equi)}\left(x_i\right) = \frac{1}{R}\sum_{r=1}^{R}\hat{e}_{Y_r}^{(equi)}\left(x_i\right), \qquad (12)$$

gdzie $\hat{\bar{e}}_Y^{(equi)}\left(x_i\right)$ jest uśrednionym wynikiem wyniku punktowego x_i serii zrównań w R losowaniach uzyskanych metodą bootstrap. Oraz:

$$SE\left(x_i\right) = \sqrt{Var\left[\hat{e}_{Y_1}^{(equi)}\left(x_i\right), \hat{e}_{Y_2}^{(equi)}\left(x_i\right), \ldots, \hat{e}_{Y_R}^{(equi)}\left(x_i\right)\right]}, \qquad (13)$$

$$RMSE\left(x_i\right) = \sqrt{bias^2 + SE^2}\,. \qquad (14)$$

Błąd zrównania dla całego testu rozumiemy jako średnią arytmetyczną lub średnią ważoną z błędów dla poszczególnych wartości punktowych. Obciążenie jest miarą błędów systematycznych, a *SE* błędów losowych, RMSE jest połączeniem obu miar błędów (Kolen i Brennan, 2004). W tym badaniu błędy były szacowane przy użyciu 1000 iteracji.

Przedziały ufności dla średniej wyników zrównanych szacowane są także na podstawie metody *bootstrap* i rozumiane są jako odchylenie standardowe między średnimi dla sum punktów testu zrównanego *R* razy, w sposób opisany powyżej, pomnożone przez 1,96.

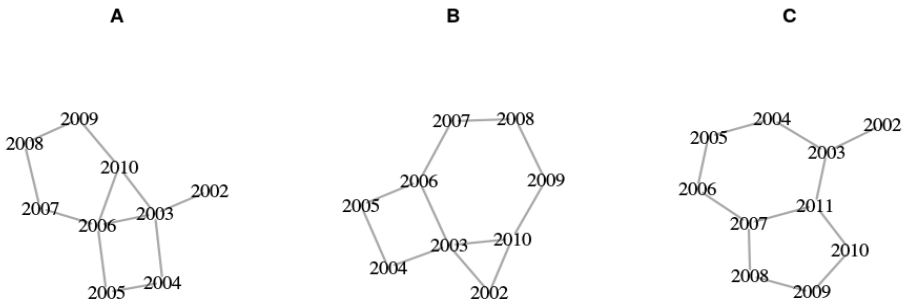
Metoda

Dane

W pierwszym etapie analiz wykorzystane zostały dane pochodzące z badań przeprowadzonych w latach 2011–2013 (Szaleniec i in., 2012, 2013). Pełne przedstawienie ich metodologii wykracza poza ramy niniejszego artykułu. Została ona szczegółowo opisana w raportach z tych badań. Badania były prowadzone na ogólnopolskich, losowych próbach szkół, gdzie w ramach każdej z wylosowanych szkół arkusze egzaminacyjne rozwiązywali uczniowie z pojedynczego, również dobranego losowo, oddziału. Cała próba podzielona była losowo na podpróby nie mniejsze niż 800 uczniów, z których każda rozwiązywała arkusz testowy składający się z wybranych zadań z egzaminów z dwóch różnych lat spośród egzaminów z lat 2002–2012 oraz dodatkowych, zewnętrznych zadań kotwiczących, które nie zostały jednak uwzględnione w zrównaniu liniowym i ekwicentylowym. Pełne

Tabela 1
Liczba szkół i uczniów włączonych do analizy w 2011, 2012 i 2013 r.

Sesja badawcza	Typ egzaminu	Liczba szkół w próbie	Liczba uczniów w próbie
2011	Egzamin gimnazjalny, część matematyczno-przyrodnicza	442	9 551
	Egzamin gimnazjalny, część humanistyczna	442	9 593
2012	Egzamin gimnazjalny, część matematyczno-przyrodnicza	80	1 682
	Egzamin gimnazjalny, część humanistyczna	80	1 689
	Sprawdzian w klasie szóstej	439	9 086
2013	Egzamin gimnazjalny, część matematyczno-przyrodnicza	80	1 891
	Egzamin gimnazjalny, część humanistyczna	80	1 912
	Sprawdzian w klasie szóstej	80	1 598



Rysunek 3. Plany badań zrównujących.

Uwaga: diagramy przedstawiają kolejno: A) egzamin gimnazjalny, część humanistyczna (badanie 2011); B) egzamin gimnazjalny, część matematyczno-przyrodnicza (badanie 2011); C) sprawdzian (badanie 2012).

schematy dostępne są w publikacjach opisujących badanie (Szaleniec i in., 2012; 2013). Na Rysunku 3 zobrazowane zostały plany badań z 2011 r. dotyczące egzaminu gimnazjalnego (w części humanistycznej i matematyczno-przyrodniczej) i z 2012 r. dotyczący sprawdzianu. Przedstawiają one testy z poszczególnych lat, których pary były rozwiązywane przez grupy uczniów, co zilustrowano jako połączenie między testami. Takie ujęcie jest ściśle związane z zastosowaną procedurą badania.

W drugim etapie badania zrównano sumaryczne wyniki egzaminacyjne pochodzące od wszystkich uczniów zdających dany typ egzaminu w głównej sesji egzaminacyjnej.

Procedura badania

Dla planu pojedynczej grupy (SG) procedura zrównywania składała się z kilku etapów. Pierwszym krokiem było oszacowanie funkcji zrównujących poszczególne pary testów z kolejnych lat na podstawie danych pochodzących z badań zrównujących. Następnie, funkcje zrównujące pary testów łączone były w łańcuchy, tak by każdy test, pokonując możliwie krótką „drogę", rozumianą jako szereg pośrednich zrównań, ostatecznie mógł zostać zrównany do testu w roku bazowym, jakim był rok 2003 w przypadku egzaminów gimnazjalnych i 2004 dla sprawdzianu w klasie szóstej. Gdy istniało kilka nieredundantnych dróg zrównania, tworzone były alternatywne łańcuchy. Na kolejnym etapie szacowane były błędy zrównania dla poszczególnych łańcuchów funkcji zrównujących. Uśrednione wyniki różnych metod predykcyjnych dają zazwyczaj lepsze oszacowania niż indywidualne metody (Clemen, 1989; Makridakis i Winkler, 1983; Winkler i Makridakis, 1983). Gdy znane były oszacowania błędów dla poszczególnych łańcuchów, alternatywne łańcuchy zrównujące testy z tych samych lat były ze sobą uśredniane (Holland i Strawderman, 2011) na podstawie wag będących odwrotnością kwadratów błędów zrównania. Uśredniony wynik punktowy zrównania $\tilde{x}_i$, to suma ważona wyników punktowych x_{ij} pochodzących z *j* zrównań:

$$\tilde{x}_i = \sum_j x_{ij} \times \left(w_{ij} \Big/ \sum_j w_{ij} \right), \qquad (15)$$

gdzie $w_{ij} = 1/\,SE\left(x_i\right)_j^2$.

W ten sposób powstawały nowe funkcje zrównujące mogące posłużyć do zrównania rzeczywistych wyników testowych. Następnie cały proces tworzenia funkcji zrównujących i uśredniania ich na podstawie wag uzyskanych na etapie 4 powtarzany był podczas

R losowań bootstrap, by w ten sposób oszacować błędy zrównania dla uśrednionych łańcuchów zrównujących. Na podstawie tych błędów oszacowano przedziały ufności dla wyników zrównania. W kolejnym kroku funkcje utworzone z uśrednionych łańcuchów zrównań posłużyły do zrównania testów pochodzących z sesji egzaminacyjnych rzeczywistych egzaminów. Żeby to mogło nastąpić, na początku pełne skale testów zrównane zostały w planie SG do skali ograniczonej do zadań rozwiązywanych przez uczniów w sesji badawczej. Takie dane zostały następnie zrównywane. Ostatecznym etapem było zrównanie w planie SG wyników w skali testu składającego się z części zadań do skali testu składającego się ze wszystkich zadań.

Dla planu równoważnych grup (EG) i zrównania liniowego procedura badania przedstawiała się prościej, ponieważ w tym przypadku wyniki wszystkich uczniów, którzy wypełnili dany test, były bezpośrednio zrównywane do testu w roku bazowym, tj. nie zachodziła konieczność uśredniania wielu funkcji zrównujących. W celu oszacowania błędów zrównywania procedura ta została następnie powtórzona podczas *R* losowań bootstrap. Za pomocą oszacowanych w ten sposób funkcji zrównano wyniki uzyskiwane na egzaminach dla podzbioru zadań egzaminacyjnych wykorzystanych na sesji badawczej. Następnie przeliczone wyniki dla podzbiorów zadań egzaminacyjnych zrównano w planie SG do pełnych skal egzaminu. Zrównanie EG możliwe było tylko dla danych pochodzących z badania z 2011 r. dla egzaminu gimnazjalnego (arkusze testowe z lat 2002–2010), a w przypadku sprawdzianu dla danych z badania w 2012 r. (arkusze testowe z lat 2002–2011), ponieważ tylko w tym wypadku ta sama populacja uczniów rozwiązywała test z roku bazowego i pozostałe arkusze, spełnione więc było założenie o ekwiwalentności grup.

Metody zrównywania użyte w tym badaniu to kolejno: zrównanie liniowe, zrównanie ekwicentylowe w schemacie SG, zrównanie ekwicentylowe w schemacie EG. Porównano zrównania metodą ekwicentylową z zastosowaniem wstępnego i wtórnego wygładzania rozkładów. Do wstępnego wygładzania rozkładów zastosowano model log-liniowy, którego parametry dobrane zostały na podstawie kryteriów AIC i BIC. Przy wtórnym wygładzaniu zastosowano estymator jądrowy, do którego parametry dobrane zostały automatycznie (Davier, 2007; Davier, Holland i Thayer, 2004). Użytą przy właściwym zrównywaniu ekwicentylowym metodą interpolacji był nieparametryczny model liniowy oparty na naturalnych sześciennych funkcjach sklejanych (Green i Silverman, 1993). Przy bardzo dobrym dopasowaniu do danych można potraktować go jako dodatkową formę wygładzania odbywającą się na etapie zrównania. Rozwiązanie to daje bardzo podobne wyniki jak interpolacja liniowa i nie jest obarczone większym błędem. Naturalne sześcienne krzywe sklejane są stosowane jako forma wygładzania wtórnego (Kolen i Brennan, 2004).

Narzędzia użyte do analizy danych

Całość analiz prowadzona była w środowisku statystycznym *R* (R Core Team, 2014). Początkowo analizy prowadzone były przy użyciu biblioteki *equate* (Albano, 2014), jednak ostatecznie zastosowano autorski pakiet *equi* (Wołodźko, 2014). Wyniki uzyskane za pomocą funkcji z obu pakietów były porównywalne i ich zestawienie nie będzie tu szerzej omawiane. Użycie autorskiego oprogramowania podyktowane było nieuniknionymi przy złożonym planie badawczym trudnościami technicznymi oraz potrzebą większej kontroli nad kolejnymi etapami procesu analiz.

Wyniki

Wyniki zrównań przedstawione są na Rysunkach 5–7. Zestawiono zrównanie

liniowe, zrównania ekwicentylowe w planach SG i EG oraz zrównanie za pomocą modelowania IRT. Celem badania było porównanie poszczególnych metod zrównywania, natomiast więcej informacji na temat trendów czasowych wyników egzaminacyjnych czytelnik znajdzie w przytaczanych już raportach (Szaleniec i in., 2012; 2013). Na wykresach przedstawiających wyniki zrównań jedynie zrównanie w planie SG przedstawiono z towarzyszącymi mu przedziałami ufności, co podyktowane jest ich czytelnością. W celu uwidocznienia różnic i zbieżności między poszczególnymi zrównaniami, zakres skali na wykresach to ±0,33 odchylenia standardowego wyników testów z lat 2002–2012 zrównanych do testu w roku bazowym. Pozioma linia przerywana to średnia testu z roku bazowego (2003 r. dla egzaminu gimnazjalnego i 2004 r. dla sprawdzianu). Oszacowania dla EG dotyczą jedynie lat 2002–2010 w przypadku egzaminu gimnazjalnego i 2002–2011 w przypadku sprawdzianu, ponieważ jedynie dane dla tych lat pozwoliły na zastosowanie takiego planu badania. Różnice między zrównaniami SG i EG zostały ocenione na podstawie 95-procentowych przedziałów ufności i nie różnią się w sposób istotny statystycznie dla wszystkich lat i wszystkich rodzajów egzaminów. Przedziały ufności dla zrównania liniowego i opartego na IRT nie zostały oszacowane, jednak wszystkie średnie wyniki dla zrównań liniowego i opartego na IRT mieściły się w przedziale ufności dla EG, natomiast porównania z SG można dokonać na podstawie Rysunków 5–7.

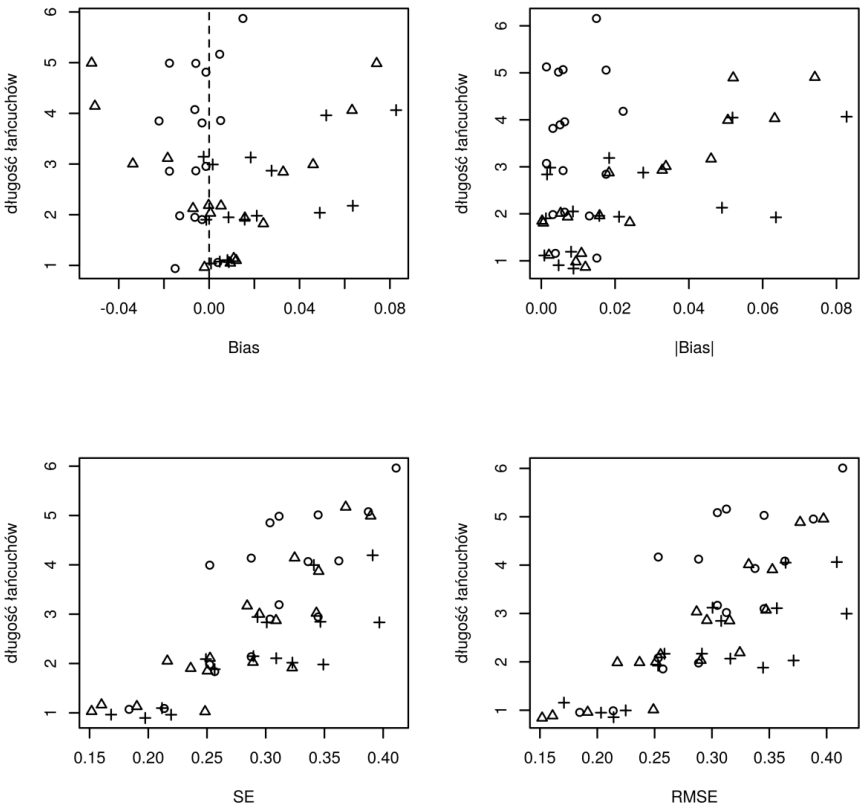
Informacje o błędach zrównania dla poszczególnych planów oraz metod wygładzania można znaleźć na Rysunkach 8–16. Są to uśrednione błędy zrównania wartości punktowych. Na wykresach nie uwzględniono testu z roku bazowego, ponieważ stanowił układ odniesienia i jego parametry w całej procedurze miały ustalone wartości. Zrównanie w schemacie EG obarczone było większym błędem niż zrównania w schemacie SG.

Sprawdzono także, jaki wpływ na wynik zrównania i towarzyszące mu oszacowania błędów ma fakt „oddalenia" od testu bazowego, rozumianego jako długość łańcucha funkcji zrównujących między zrównywanymi testami. Testy zrównywania za pomocą dłuższego łańcucha obarczone były większym błędem zrównania w przypadku miar, takich jak *SE* i RMSE, natomiast nie zaobserwowano takiej zależności w przypadku obciążenia, co możemy zaobserwować na Rysunkach 8–16 oraz na Rysunku 4. Korelacje z długością łańcucha wynoszą 0,01 dla obciążenia, 0,34 dla jego wartości absolutnej, 0,79 dla *SE* i 0,77 dla RMSE.

Dyskusja

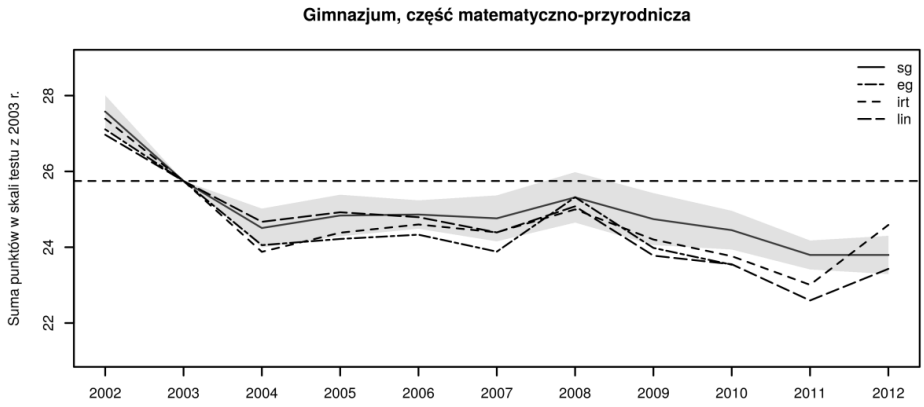
Uzyskane wyniki wskazują, że podobne efekty można uzyskać, korzystając z różnych metod zrównywania (Rysunki 5–7). Musimy jednak zdawać sobie sprawę z różnic wynikających z wyboru strategii – zrównywanie z wykorzystaniem metod opartych na modelach IRT, czy metod klasycznych, a także wyboru określonego planu zrównywania SG lub EG.

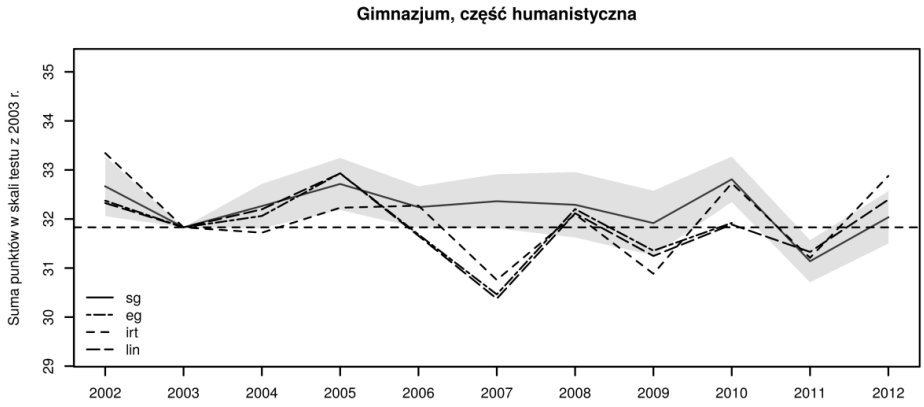
Stosując plan EG, uzyskujemy oszacowanie zrównania oparte na większych próbach niż przy SG, a więc oparte na danych, których rozkład obarczony jest mniejszym błędem wynikającym z losowania próby. Mimo tego, zrównanie ekwicentylowe w planie SG obarczone jest niższym błędem niż EG (por. Rysunki 8–16). Taki wynik nie dziwi, ponieważ przy planie SG wykorzystywana jest informacja z łącznego rozkładu zrównywanych testów, podczas gdy w przypadku planu EG jedynie przyjmujemy założenie o tożsamym poziomie umiejętności w obu grupach i fakt rozwiązywania dwóch wersji testu przez tego samego ucznia (zagnieżdżenie pomiarów w uczniu) jest pomijany. Przeprowadzone

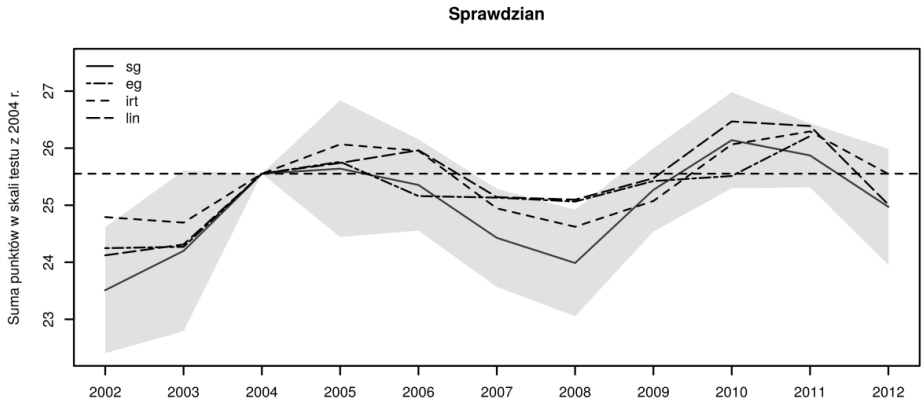


Rysunek 4. Związki długości łańcuchów funkcji zrównujących z oszacowaniami błędów: obciążeniem (*bias*), wartością absolutną obciążenia, błędem standardowym (*SE*), RMSE. Okręgami oznaczone są zrównania sprawdzianu, trójkątami – testu gimnazjalnego z matematyki, krzyżykami – testu gimnazjalnego z języka polskiego. W celu zwiększenia czytelności wykresów długość łańcuchów przedstawiona jest z dodaniem losowego „szumu".

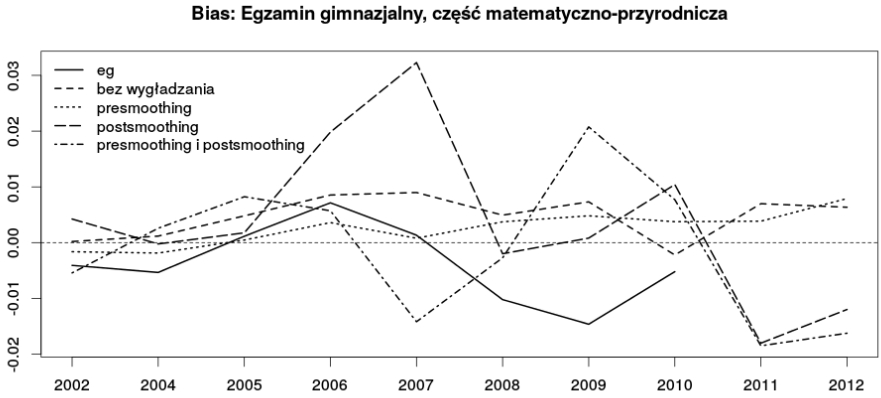
analizy pozwoliły porównać oba rozwiązania, choć nie dały jednoznacznej odpowiedzi, które rozwiązanie daje lepsze rezultaty. Najbardziej jaskrawy przykład różnic (dla planów SG w porównaniu z EG i metody opartej na modelach IRT) zaobserwowano dla zrównania części humanistycznej egzaminu gimnazjalnego odnośnie do wyników z 2007 r. (Rysunek 6), choć i tu różnica między SG i EG nie była istotna w sensie statystycznym. Różnice te najprawdopodobniej wynikają z faktu, że w przypadku SG wykorzystane zostały uśrednione wyniki pochodzące z różnych prób, a pozostałe metody korzystały z pełnych prób wszystkich osób, które rozwiązywały dany test. Fakt, że zrównanie w planie EG przyniosło podobne wyniki jak pozostałe zrównania, może również świadczyć o dobrym doborze próby, dzięki któremu poszczególne grupy były zbliżone pod względem umiejętności. Ponieważ jednak nie istnieje żadna obiektywna miara pozwalająca ocenić, które rozwiązanie jest „prawdziwe", a które „błędne", należy traktować poszczególne zrównania jako różne sposoby wykorzystania informacji pochodzących z tych samych danych empirycznych.

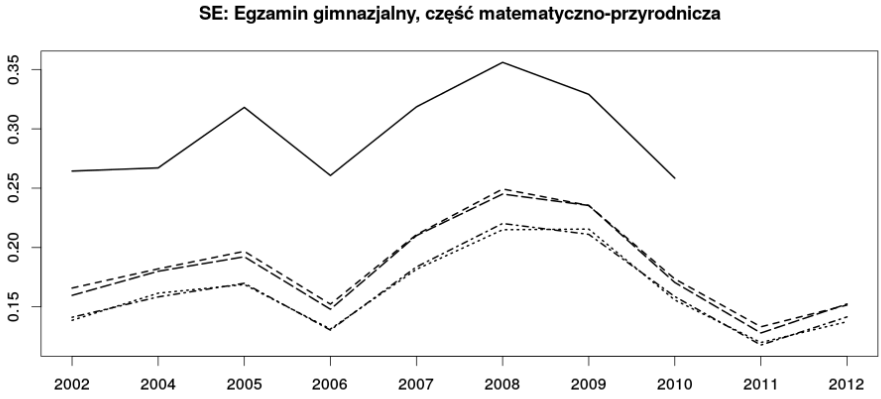


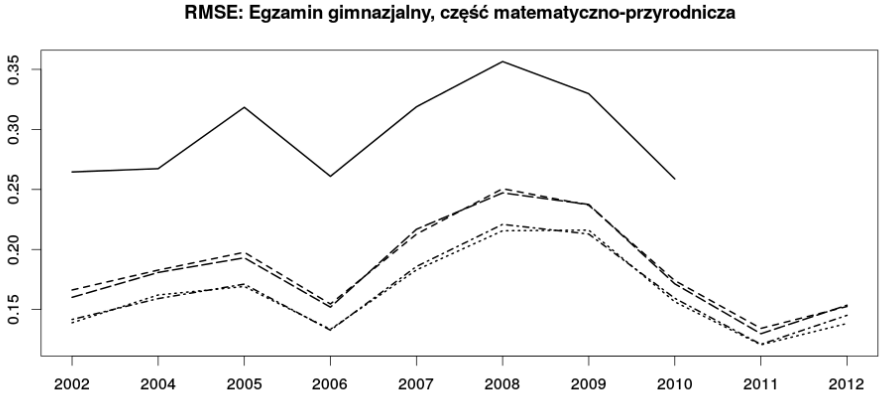




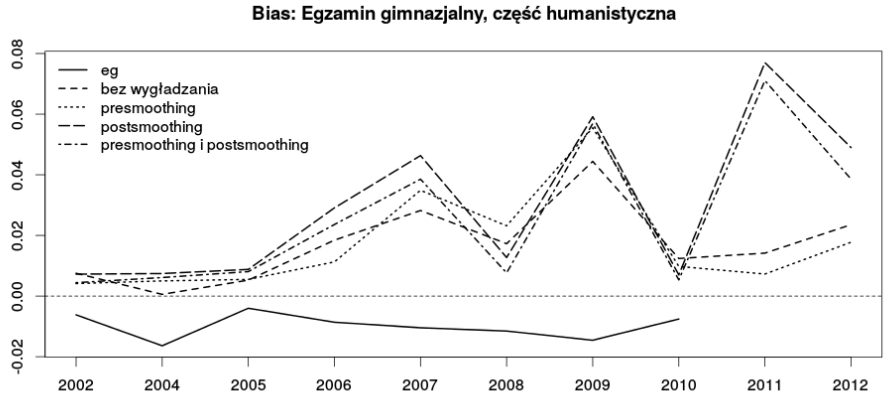
Rysunki 5–7. Wyniki egzaminu gimnazjalnego i sprawdzianu zrównanie liniowo, ekwicentylowo w planie SG i EG i przy użyciu modelowania IRT.

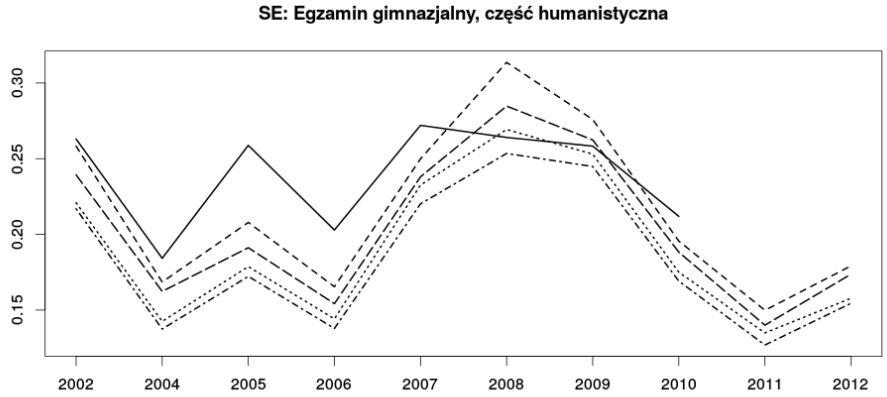


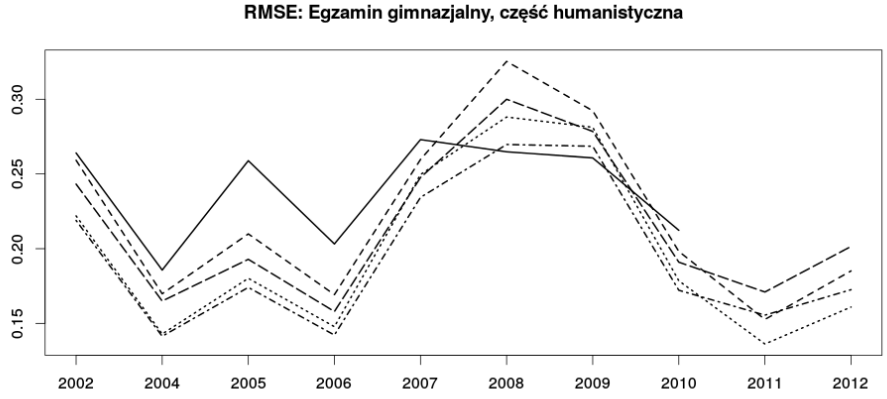




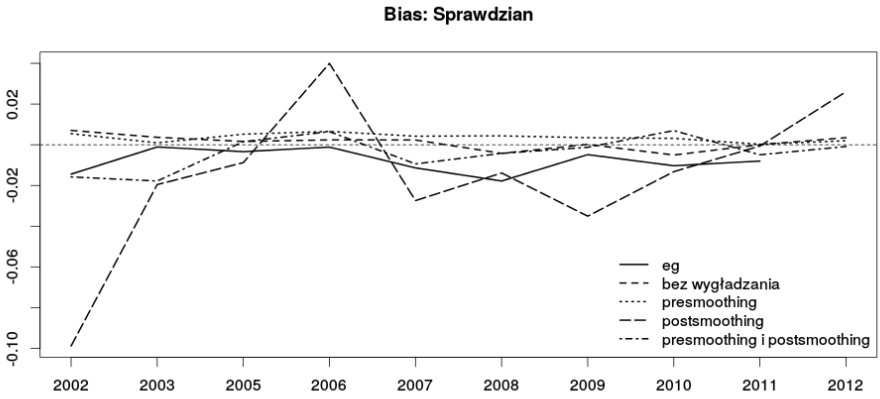
Rysunki 8–10. Błędy zrównania dla egzaminu gimnazjalnego, części matematyczno-przyrodniczej dla planu EG oraz planu SG z różnymi poziomami wygładzania.

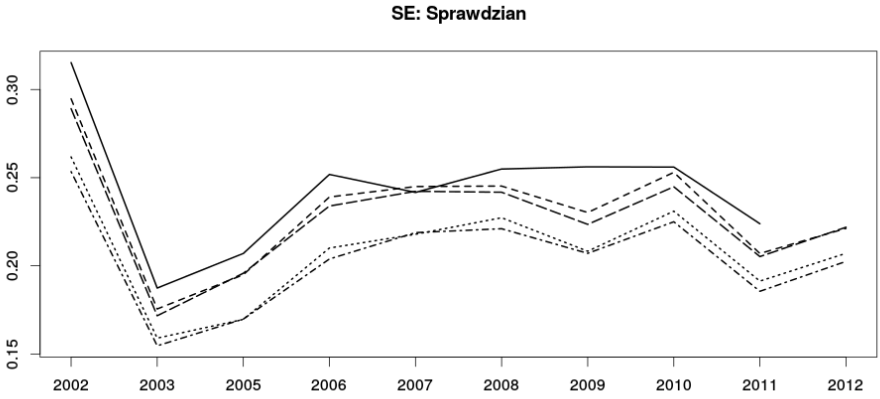


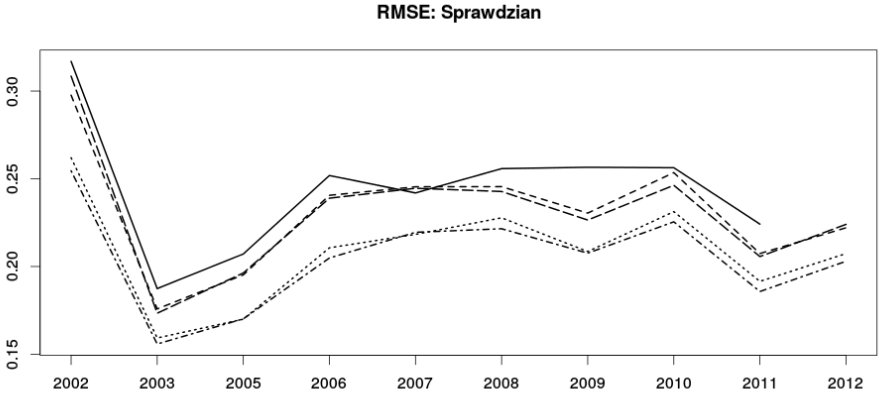




Rysunki 11–13. Błędy zrównania dla egzaminu gimnazjalnego, części humanistycznej dla planu EG oraz planu SG z różnymi poziomami wygładzania.







Rysunki 14–16. Błędy zrównania dla sprawdzianu dla planu EG oraz planu SG z różnymi poziomami wygładzania.

W trakcie analiz porównane zostały także różne sposoby wygładzania rozkładów. Uzyskane wyniki pokazują, że różnice między nimi nie są duże. Szczególnie zastanawiający jest fakt, że również zrównanie, w którym nie użyto wstępnego, ani wtórnego wygładzania rozkładów, dało wynik obarczony zbliżonym poziomem błędów. Taki rezultat wskazuje, jak ważnym elementem zrównania testów jest dobór próby oraz na to, że w niektórych sytuacjach nawet w przypadku próby mniejszej niż 20 tys. osób, wygładzanie rozkładów nie musi być konieczne (por. Davier, 2011).

Analizy zrównywania wyników sprawdzianu i egzaminu gimnazjalnego przeprowadzone z wykorzystaniem różnych metod pozwoliły zaobserwować podobne trendy, a różnice nie są duże, jeśli weźmiemy pod uwagę fakt, że każda z nich obarczona jest pewnym poziomem błędów, a więc też niepewności co do rzeczywistego wyniku. Dostarczyły one także cennych informacji na temat różnych rozwiązań analitycznych, jakie można wykorzystać w przypadku, kiedy plan badania jest złożony – obejmujący zrównywanie wyników egzaminów przeprowadzonych w okresie kilkunastu lat.

Analizy potwierdziły, jak ważny jest etap planowania badań zrównujących w odniesieniu do konkretnej metody zrównywania. W opisywanych schematach badań istniały testy, które były bardziej oddalone pod względem liczby pośrednich połączeń z testem z roku bazowego (np. testy z 2008 r.) – wyniki tych zrównań obarczone były większym błędem losowym. Wyniki badania wskazują na związek między błędami losowymi (*SE*, RMSE) a długością łańcucha funkcji zrównujących i brak takiego związku dla obciążenia, będącego miarą błędów systematycznych. Oznacza to, że przy zestawieniu w ramach łańcucha funkcji zrównujących kilku zrównań jednostkowe systematyczne odchylenia wzajemnie niwelują swój wpływ na ostateczny wynik, z drugiej jednak strony dochodzi do kumulowania się zakłóceń losowych. Wynika z tego, że przy planowaniu badań, w których byłaby zastosowana metodologia zrównywania wyników obserwowalnych, duży nacisk należy położyć na jakość „połączeń" między zrównywanymi testami. Należy przy tym unikać zrównań za pomocą długich łańcuchów funkcji zrównujących. Jeśli weźmiemy pod uwagę fakt, że zrównanie w planie SG obarczone jest mniejszym błędem, niż w planie EG, oznaczać to będzie, że planując zrównanie, najlepiej oprzeć je na bezpośrednim zrównywaniu par testów, używając schematu SG.

Mimo że tematyka ta nie była tematem analiz, warto w tym miejscu również zaznaczyć, że ważną rolę dla wyników zrównania ma jakość testów kotwiczących. Wpływ ich doboru jest tym większy, im większe są różnice w umiejętnościach grup uczniów rozwiązujących zrównywane testy, przy czym im są one większe, tym powinniśmy opierać się na dłuższych testach, o lepszych właściwościach psychometrycznych (Dorans, Moses i Eignor, 2011).

Podsumowanie

Badania zrównujące przeprowadzone w latach 2011–2014 przez Zespół Analiz Osiągnięć Uczniów IBE zaplanowane były do zastosowania strategii opartej na modelach IRT, natomiast opisane w tym artykule wyniki zrównania z zastosowaniem metod klasycznych stanowią ich uzupełnienie. W artykule przedstawiono wyniki zrównania ekwicentylowego trzech egzaminów z jedenastoletniego przedziału czasowego. Porównane zostały różne metody zrównywania wyników obserwowanych: liniowe, ekwicentylowe w planie SG, ekwicentylowe w planie EG oraz różne sposoby wygładzania rozkładów. Metody te, mimo różnic między stosowanymi algorytmami i planami badawczymi, dały zbliżone do siebie wyniki. Zrównanie liniowe pozwala dobrze oszacować

średni wynik testu po zrównaniu, podczas gdy metoda ekwicentylowa pozwala na wnioskowanie na poziomie przeliczonych wyników punktowych. SG obarczone jest mniejszym błędem zrównania, jednak wymaga także danych, w których poszczególne grupy osób badanych rozwiązują arkusze egzaminacyjne parami. EG nie wymaga tego typu danych, niesie jednak ze sobą o wiele silniejsze założenia na temat takiego samego poziomu umiejętności w grupach rozwiązują oba testy. Oba plany zrównania mają swoje plusy i minusy, które należy rozważyć, planując badania naukowe i działania prowadzące do zrównywania wyników w trakcie sesji i do komunikowania zrównanych wyników równolegle z wynikami surowymi. Szczególną uwagę należy zwrócić na kwestię kotwiczenia testów, w tym długości łańcuchów funkcji zrównujących (w przypadku, gdy planujemy zrównywanie wyników z wielu lat dla danego egzaminu), ponieważ zastosowanie bardziej skomplikowanego planu zrównywania może prowadzić do wyniku obarczonego większym błędem. Analiza przeprowadzona została z użyciem zastanych danych, jej wyniki mogą być jednak wskazówką dla tworzenia planów zrównań.

Literatura

Albano, A. D. (2014). *Equate: an R package for observed-score linking and equating*. Pobrano z http://cran.r-project.org/web/packages/equate/index.html

Clemen, R. T. (1989). Combining forecasts: a review and annotated. *International Journal of Forecasting, 5,* 559–583.

Davison, A. C. i Hinkley, D. V. (2009). *Bootstrap methods and their application.* New York, NY: Cambridge University Press.

Davier, A. A. Von (2011). A statistical perspective on equating test scores. W: A. A. von Davier (red.), *Statistical models for test equating, scaling, and linking.* New York, NY: Springer.

Davier, A. A. von, Holland, P. W. i Thayer, D. T. (2004). *The kernel method of test equating*. New York, NY: Springer.

Dorans, N. J., Moses, T. P. i Eignor, D. R. (2011). Equating test scores: toward best practices. W: A. A. von Davier (red.), *Statistical models for test equating, scaling, and linking.* New York, NY: Springer.

Efron, B. i Tibshirani, R. J. (1993). *An introduction to the bootstrap.* London: Chapman & Hall/CRC.

Field, C. A. i Welsh, A. H. (2007). Bootstrapping clustered data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *69*(3), 369–390.

Green, P. J. i Silverman, B. W. (1993). *Nonparametric regression and generalized linear models: a roughness penalty approach.* London: Chapman & Hall/CRC.

Holland, P. W. (2007). A framework and history for score linking. W: N. J. Dorans, M. Pommerich i P. W. Holland (red.), *Linking and aligning scores and scales*. New York, NY: Springer.

Holland, P. W. i Strawderman, W. E. (2011). How to average equating functions, if you must. W: A. A. von Davier (red.), *Statistical models for test equating, scaling, and linking.* New York, NY: Springer.

Holland, P. W. i Thayer, D. T. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics*, *25*(2), 133–183.

Kolen, M. J. i Brennan, R. L. (2004). *Test equating, scaling and linking*. New York, NY: Springer.

Kolen, M. J. i Jarjoura, D. (1987). Analytic smoothing for equipercentile equating under the common item nonequivalent populations design. *Psychometrika*, *52*(1), 43–59.

Livingston, S. A. i Kim, S. (2009). The circle-arc method for equating in small samples, *Journal of Educational Measurement*, *46*(3), 330–343.

Makridakis, S. i Winkler, R. L. (1983) Averages of forecasts: some empirical results. *Management Science*, *29*(9), 987–996.

Moses, T. P. i Holland, P. W. (2009). Selection strategies for univariate loglinear smoothing models and their effect on equating function accuracy. *Journal of Educational Measurement*, *46*(2), 159–176.

Moses, T. P. i Davier, A. A. Von (2006). *A SAS macro for loglinear smoothing: applications and implications.* (ETS Research Rep. No. RR-06-05). Princeton: Educational Testing Services.

Niemierko, B. (2004). Zrównywanie wyników sprawdzianu 2004 do wyników sprawdzianu 2003. W: B. Niemierko i H. Szaleniec (red.), *Standardy wymagań i normy testowe w diagnostyce edukacyjnej*. Kraków: Polskie Towarzystwo Diagnostyki Edukacyjnej.

Niemierko, B. (2007). Zrównywanie wyników sprawdzianu 2005 do wyników sprawdzianu 2003 metodą ekwicentylową. *Egzamin*, *10*, 86–104.

Pokropek, A. (2011). *Zrównywanie wyników egzaminów zewnętrznych w kontekście międzynarodowym*. Pobrano z http://www.ptde.org/file.php/1/Archiwum/XVII_KDE/pedeefy/Pokropek_2.pdf

Pokropek, A. i Kondratek, B. (2012). Zrównywanie wyników testowania. Definicje i przykłady zastosowania. *Edukacja*, *120*(4), 52–71 .

R Core Team (2014). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing. Pobrano z http://www.R-project.org/

Rena, S., Lai, H., Tong, W., Aminzadeh, M., Hou, X. i Lai, S. (2010). Nonparametric bootstrapping for hierarchical data. *Journal of Applied Statistics*, *37*(9), 1487–1498.

Smolik, M. (2007). Zrównywanie wyników sprawdzianu 2004 i sprawdzianu 2005 do wyników sprawdzianu 2003 z wykorzystaniem modelu Rascha. *Egzamin*, *10*, 86–104.

Szaleniec, H. (2005). Wykorzystanie probabilistycznych modeli zadania testowego do zrównywania wyników. W: B. Niemierko, G. Szyling (red.), *Holistyczne i analityczne metody diagnostyki edukacyjnej perspektywy informatyczne egzaminów szkolnych*. Gdańsk: Fundacja rozwoju Uniwersytetu Gdańskiego.

Szaleniec, H. (2007). Zrównywanie wyników sprawdzianu w latach 2003–2005 z wykorzystaniem probabilistycznej teorii zadania. *Egzamin*, *10*, 86–104.

Szaleniec, H., Grudniewska, M., Kondratek, B., Kulon, F. i Pokropek, A. (2012). Wyniki egzaminu gimnazjalnego 2002–2010 na wspólnej skali. *Edukacja*, *119*(3), 9–30

Szaleniec, H., Grudniewska, M., Kondratek, B., Kulon, F., Pokropek, A., Stożek, E. i Żółtak, M. (2013). *Analiza porównawcza wyników egzaminów zewnętrznych – sprawdzian w szóstej klasie szkoły podstawowej i egzamin gimnazjalny*. Warszawa: Instytut Badań Edukacyjnych.

Wand, M. P. i Jones, M. C. (1995). *Kernel smoothing*. London: Chapman & Hall/CRC.

Wang, C. (2011). *An investigation of bootstrap methods for estimating the standard error of equating under the common-item nonequivalent groups design*. Pobrano z http://ir.uiowa.edu/etd/1188

Winkler, R. L. i Makridakis, S. (1983). The combination of forecasts. *Journal of the Royal Statistical Society*, *146*(2), 150–157.

Wołodźko, T. (2014). *equi: R library for equipercntile equating*. Pobrano z https://github.com/twolodzko/equi