Edukacja 2014, 3(128), 95–111
ISSN 0239-6858

Modele analizy efektu oceniającego w pomiarze edukacyjnym

Filip Kulon
Instytut Badań Edukacyjnych*

Artykuł powstał w wyniku poszukiwań optymalnego modelu analizy w ramach prowadzonych badań porównywalności oceniania i efektu egzaminatora w zakresie egzaminu maturalnego z języka polskiego i matematyki. W części pierwszej przedstawiono krótko teorię dotyczącą zagadnienia efektu oceniającego (*rater effect*), odnosząc je do obszaru pomiaru edukacyjnego w Polsce, w którym otrzymało ono nazwę efektu egzaminatora. Skupiono się na zagadnieniu od strony pomiarowej i nie rozważano psychologicznych podstaw oceniania. W drugiej części artykułu przedstawiono wybrane modele analizy tego efektu i wskazano, który model pozwala na oszacowanie największej liczby różnych aspektów efektu egzaminatora. Opisane zostały również symulacje sprawdzające przydatność modelu HRM-SDT do analizy danych z polskiego egzaminu maturalnego.

Słowa kluczowe: efekt oceniającego, pomiar edukacyjny, efekt egzaminatora, ocenianie, *item response theory*, IRT.

Efekt oceniającego w ocenianiu umiejętności

Naukowe zainteresowanie czynnikami związanymi z ocenianiem sięga początków XX w. i badań Edwarda Thorndike'a nad efektem halo (Saal, Downey i Lahey, 1980). Koncepcje wywodzące się z psychologii ewoluowały i znalazły swoje zastosowanie, oprócz psychologii organizacji i zarządzania, również w pomiarze edukacyjnym. Ilekroć mamy do czynienia z ocenianiem przez ludzi, możemy mówić o efekcie oceniającego, czyli o jego wpływie na ocenę. Wpływ ten może zależeć nie tylko od indywidualnych cech oceniającego, ale może on być również powodowany procedurą oceniania czy użytą skalą oceny. Wszelkie tego typu efekty zwykło się jednak zbiorczo nazywać efektem oceniającego (*rater effect*). Jest to termin określający szeroką gamę czynników generujących wariancję ocen niezwiązaną z rzeczywistym poziomem mierzonej cechy ukrytej ocenianego, ale z oceniającym (Scullen, Mount i Goff, 2000). Wśród nich można wyróżnić kilka najczęściej badanych i opisywanych efektów (Saal i in., 1980): efekt halo, łagodność i surowość, tendencja centralna (ograniczenia skali), rzetelność (zgodność).

Artykuł powstał w ramach projektu systemowego „Badanie jakości i efektywności edukacji oraz instytucjonalizacja zaplecza badawczego" realizowanego przez Instytut Badań Edukacyjnych i współfinansowanego ze środków Europejskiego Funduszu Społecznego (Program Operacyjny Kapitał Ludzki 2007-2013, priorytet III: Wysoka jakość systemu oświaty).

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

Niewątpliwie punktem wyjścia i najistotniejszym elementem prac nad efektem oceniającego jest założenie, że każdy pomiar obarczony jest błędem. Wśród błędów składających się na szeroko rozumiany efekt oceniającego możemy wyróżnić błędy losowe i systematyczne. Przyjrzyjmy się krótko poszczególnym typom efektów wymienionym powyżej i określmy skojarzony z nimi rodzaj błędu.

Z efektem halo mamy do czynienia, gdy oceniający przypisuje ocenianemu jakiś poziom mierzonej cechy w różnych wymiarach (kryteriach) na podstawie ogólnego wrażenia, zamiast oceniać poszczególne wymiary niezależnie. Trudno jednoznacznie wskazać czy ten typ efektów należy uznać za błąd systematyczny, czy też losowy. Z jednej strony, jeśli ogólne wrażenie oceniającego tworzone jest na podstawie cechy, która koreluje z ocenianymi wymiarami, będziemy mieć do czynienia z błędem systematycznym. Z drugiej, jeśli ogólne wrażenie i poszczególne wymiary ocenianej cechy będą niezależne, oceniający będzie losowo przydzielał oceny niezwiązane z rzeczywistym poziomem poszczególnych kryteriów.

Łagodność i surowość w ocenianiu polegają na systematycznym przypisywaniu ocen niższych lub wyższych niż odpowiadające rzeczywistemu poziomowi mierzonej cechy. Jest to bardzo prosta, a zarazem jasna definicja wskazująca na charakter błędu generowanego przez ten typ efektów.

Tendencja centralna, nazywana także ograniczeniem skali, polega na przypisywaniu osobom ocenianym kategorii położonych blisko środka skali, niezależnie od rzeczywistego poziomu mierzonej cechy. Jednak ograniczenie skali można potraktować jako odrębne zjawisko, gdyż niekoniecznie musi ono następować w jej środku. Można sobie wyobrazić, że oceniający przyznaje najczęściej np. tylko oceny 1–3 z pięciostopniowej skali – mamy wówczas do czynienia z ograniczeniem skali, ale nie z tendencją centralną. Jeszcze innym aspektem tego typu błędów jest stosowanie ocen skrajnych, czyli próba uproszczenia skali oceny do dychotomii, bez rozróżniania poszczególnych poziomów cechy. Takie zjawisko nazywane jest ekstremizmem (Wolfe, 2004). Zjawiska takie jak tendencja centralna, ograniczenie skali czy ekstremizm są przykładami modyfikacji skali, na jakiej mierzona cecha powinna zostać oceniona. W celu podkreślenia negatywnego skutku takich zmian można posłużyć się nazwą „zniekształcenia skali". Ten rodzaj efektów należy zaliczyć do błędów systematycznych. W zasadzie łagodność/surowość też jest zniekształceniem skali, lecz jest to efekt na tyle specyficzny, że zasługuje na wydzielenie go z tej grupy.

Niewątpliwym oczekiwaniem, jakie mamy wobec oceniających, jest to, żeby przyznawane przez nich oceny były adekwatne do poziomu mierzonej cechy. Jesteśmy zatem zainteresowani, aby nasze narzędzie, czyli schemat oceniania (*scoring rubric*) i stosujący go oceniający, było jak najbardziej trafne i jak najbardziej rzetelne. Niestety, nawet za pomocą wielokrotnego oceniania nie jesteśmy w stanie wypowiedzieć się na temat trafności narzędzia i możemy jedynie określić jego rzetelność. Potwierdzeniem wysokiej rzetelności byłoby przyznanie przez różnych oceniających tych samych ocen danemu ocenianemu, czyli zgodność. Warto jednak zauważyć, że zgodność między oceniającymi nie jest wymagana, aby narzędzie uznać za rzetelne (Saal i in., 1980). Gdy oceniający różnią się między sobą np. łagodnością, mogą nie osiągnąć zgodności pomimo tego, że będą rzetelnie przydzielać oceny. W przypadku rzetelności mamy do czynienia z losowym składnikiem błędu. Nie jesteśmy w stanie przewidzieć, w jaki sposób oceny różnych osób przez nierzetelnego oceniającego odbiegać będą od rzeczywistego poziomu mierzonej cechy. Zgodność odnosi się do ocen bezwzględnych,

w których bierzemy pod uwagę błędy systematyczne, generowane przez oceniającego (jak np. łagodność). Rzetelność natomiast odnosi się do ocen względnych, w których różnice między oceniającymi, powodujące systematyczne zniekształcenia skali ocen, nie są brane pod uwagę. Łatwo zauważyć, że termin zgodność łączy ze sobą dwa rodzaje błędów: losowe (rzetelność) i systematyczne (np. łagodność), co nie jest pożądane. W celu uniknięcia nieporozumień narosłych wokół terminów zgodność i rzetelność można mówić o precyzji stosowania schematu oceniania. Im oceniający bardziej precyzyjnie stosuje schemat, tym z mniejszym błędem losowym przyzna ocenę, co z kolei odpowiada wyższej rzetelności oceniania. Odwrotnie, niewielka precyzja stosowania schematu odpowiada za większy błąd losowy, a więc za mniejszą rzetelność.

Podsumowując powyższe rozważania na temat różnych aspektów efektu oceniającego, można zaproponować następującą, nieco zmodyfikowaną, typologię efektów: efekt halo, łagodność/surowość, zniekształcenia skali, precyzja.

Jedną z możliwości badania umiejętności uczniów jest stosowanie zadań, w których ocena ucznia ustalana jest przez oceniającego, przeważnie za pomocą dyskretnej skali ocen (wg ustalonego schematu oceniania). Zadania takie nazywa się zadaniami otwartymi (*constructed response items*). Z punktu widzenia dalszych rozważań bardzo istotny jest fakt, że w proces ich oceniania są zaangażowane osoby mające bezpośredni wpływ na jego wynik. W konsekwencji indywidualne cechy oceniających, np. ich osobowość czy doświadczenia, mogą być źródłem obciążenia wyniku.

Dotychczasowe polskie badania efektu oceniającego odnosiły się do pomiaru edukacyjnego w ramach systemu egzaminów zewnętrznych, w którym osoby oceniające prace uczniów nazywa się egzaminatorami. Mimo uniwersalności zagadnienia i jego występowania we wszelkich sytuacjach, gdy mamy do czynienia z ocenianiem jakiejś cechy przez ludzi, na gruncie polskim opisywane zjawisko przyjęło się nazywać „efektem egzaminatora" (Dolata, Putkiewicz i Wiłkomirska, 2004; Dubiecka, Szaleniec i Węziak, 2006). Ponieważ w artykule omawiane są zjawiska dotyczące pomiaru edukacyjnego i systemu egzaminów zewnętrznych, w dalszej części używany będzie ten termin.

W przypadku zadań otwartych egzaminatorzy je oceniający są również źródłem części błędu, który jest składnikiem każdego pomiaru (Szmigel i Szaleniec, 2001). W polskim systemie oceniania zewnętrznego w ramach jednej sesji egzaminacyjnej wszyscy uczniowie piszą ten sam test, do którego stosowany jest ten sam schemat oceniania i taka sama procedura. Prace poszczególnych uczniów trafiają jednak do różnych egzaminatorów, co może mieć wpływ na końcową ocenę ucznia. Ma to znaczenie szczególnie w przypadkach, kiedy wyniki egzaminu są brane pod uwagę na dalszym etapie edukacji. Poznanie skali efektu egzaminatora jest niezbędne do zapewnienia sprawiedliwego oceniania.

Opisywane zagadnienie można modelować z użyciem różnych metod, w zależności od celu, któremu ma służyć analiza. Jednym z nich może być diagnostyka systemu egzaminacyjnego i dążenie do zwiększenia rzetelności ocen przez eliminację nierzetelnych oceniających czy udoskonalenie procedur. Innym celem może być oszacowanie wpływu egzaminatorów na oceny konkretnych uczniów i wykorzystanie tej wiedzy do uwzględnienia poprawki w ostatecznej ocenie ucznia. Oczywiście, jeden model może być odpowiedni do różnych celów, choć niektóre z nich pozwalają na oszacowanie tylko jednego typu efektu egzaminatora. Przedstawione poniżej modele można podzielić na dwie grupy: modele analizy wariancji (Dolata i in., 2004; Glas, 2012; Scheerens, Glas i Thomas, 2003) oraz

modele *item response theory* (IRT; DeCarlo, Kim i Johnson, 2011; Dubiecka i in., 2006, Patz i Junker, 1999; Patz, Junker, Johnson i Mariano, 2002). Omawiane modele nie wyczerpują wszystkich podejść do szacowania efektu egzaminatora, lecz pokazują najbardziej popularne i najczęściej stosowane. W kontekście polskiego systemu egzaminów zewnętrznych efekt egzaminatora nie był często analizowany. Wśród systematycznych badań na ten temat można wymienić prace Romana Dolaty, Elżbiety Putkiewicz i Anny Wiłkomirskiej (2004) w zakresie egzaminu maturalnego oraz Anny Dubieckiej, Henryka Szaleńca i Doroty Węziak (2006) w odniesieniu do sprawdzianu po szóstej klasie szkoły podstawowej. Prace te reprezentują wspomniane wyżej, odmienne podejścia do analizy.

Z reguły zainteresowanie badaczy pomiaru edukacyjnego skupia się na precyzji oceniających oraz na ich łagodności/ /surowości, a rzadziej na efektach zniekształcenia skali. Prawdopodobną przyczyną takiego stanu rzeczy jest trudność w dobrym rozróżnieniu między łagodnością/surowością i zjawiskami takimi jak ograniczenie skali, ekstremizm czy tendencja centralna. Bardzo rzadko wspomina się o efekcie halo, jednak w danych z polskiego systemu egzaminów zewnętrznych nie mamy do czynienia z sytuacją, kiedy jeden egzaminator dokonuje oceny znacząco różnych cech (umiejętności) ucznia. Nawet jeśli ocenie podlega kilka zadań, z reguły mierzą one te same lub bardzo podobne umiejętności, zatem efekt halo właściwie nie występuje. W związku z tym zostanie on w dalszych rozważaniach pominięty. Pominięte zostaną również takie metody jak kappa Cohena (dla dwóch oceniających) czy kappa Fleissa (dla wielu oceniających). Pozwalają one na szacowanie jedynie zgodności między oceniającymi (zwane są współczynnikami zgodności), a poszukiwany, optymalny model powinien pozwalać na oszacowanie przynajmniej łagodności/ /surowości i precyzji oceniających. Pożądaną cechą byłaby także możliwość szacowania efektów zniekształceń skali.

Modele analizy

Warto przyjrzeć się, jak do tej pory opisywany i badany był efekt egzaminatora w odniesieniu do polskiego systemu egzaminacyjnego, a w szczególności, jakie metody analizy danych zastosowano.

Dolata i współpracownicy (2004) w swoich badaniach przyjęli, że na całkowite zróżnicowanie punktacji badanych prac składa się efekt zróżnicowania jakości prac, prosty efekt egzaminatora i efekt interakcji praca–egzaminator. Oszacowania tych efektów autorzy uzyskali przez wykorzystanie jednoczynnikowych modeli analizy wariancji z użyciem oceny danej pracy jako zmiennej zależnej i, odpowiednio, numeru pracy dla efektu zróżnicowania jakości pracy oraz numeru egzaminatora dla „prostego efektu egzaminatora". Miarami tych efektów były procentowe wskaźniki η^2. Autorzy przyjęli założenie, że pozostała część wariancji ocen jest wyjaśniana poprzez efekt interakcji praca–egzaminator. Jako miarę tego efektu przyjęli różnicę całkowitej wariancji i sumy pozostałych dwóch efektów. W odniesieniu do przedstawionej wyżej typologii efektu egzaminatora, użyte wskaźniki dla prostego efektu egzaminatora i efektu interakcji praca–egzaminator można uznać za miarę precyzji – im większe procentowe wskaźniki η^2, tym mniejsza precyzja egzaminatorów. Takie wskaźniki obliczane są dla całej grupy analizowanych prac (konkretnego arkusza testowego), a nie dla pojedynczych egzaminatorów.

Drugim typem wskaźników, którym posłużyli się autorzy, jest odchylenie ocen egzaminatorów od średniej ocen danej pracy – jest to informacja o łagodności/surowości egzaminatora, którą autorzy utożsamiają z prostym efektem egzaminatora. Obliczenie

tej miary jako procentu maksymalnego odchylenia standardowego możliwego dla danego zadania zapewnia jej porównywalność między różnymi zadaniami. Autorzy dokonywali oszacowania efektu egzaminatora dla danego typu arkusza przez obliczenie rozstępu indywidualnych wskaźników egzaminatorów. Taki wskaźnik nie pozwala jednak stwierdzić nic na temat przeciętnej łagodności czy surowości egzaminatorów w danym typie arkusza.

Użyty model jest dość prosty, lecz nadaje się do celów diagnostyki systemu egzaminacyjnego, a w szczególności do oceny rzetelności systemów punktacji analizowanych zadań, co było celem autorów. Wątpliwe może być założenie, że całość wariancji ocen, która nie została wyjaśniona na podstawie zróżnicowania jakości prac i prostego efektu egzaminatora, można przypisać efektowi interakcji egzaminatora z daną pracą. Nie jest to jednak szczególnie istotne w momencie, kiedy zastanawiamy się nad typem użytego modelu i wskaźników poszczególnych efektów egzaminatora, które na jego podstawie możemy uzyskać. W zasadzie mamy w przypadku tego badania do czynienia z dwoma modelami. Z jednej strony, do oszacowania precyzji oceniania na poziomie arkusza testowego użyto analizy wariancji. Z drugiej, na poziomie egzaminatorów obliczono różnice pomiędzy ocenami poszczególnych egzaminatorów i średnią ocen danej pracy, co zostało wykorzystane dwojako: jako oszacowanie łagodności/surowości egzaminatora oraz jako miara jego precyzji oceny dla danego zadania. Niestety, brakuje jakichkolwiek oszacowań efektów związanych ze zniekształceniem skali.

W badaniu zespołu Dubieckiej (2006) użyto, zaproponowanego przez Johna Linecre'a, wieloaspektowego skalowania Rascha (*many-facet Rasch measurement*, MFRM). Jest to rozwinięcie modelu Rascha zakładające, że wynik osiągnięty przez zdającego jest zależny nie tylko od jego umiejętności i trudności zadania, lecz również od innych aspektów, np. cech egzaminatora, schematu oceniania itp. W omawianym badaniu skupiono się na jednym dodatkowym aspekcie, łagodności/surowości egzaminatora, choć model dopuszcza istnienie większej liczby aspektów. Użyty, podstawowy model, można przedstawić za pomocą następującego równania:

$$ln\left(\frac{P_{nkjr}}{P_{nk(j-1)r}}\right) = B_n - D_k - R_r - F_j, \quad (1)$$

gdzie:

P_{nkjr} – prawdopodobieństwo przyznania przez egzaminatora r kategorii punktowej j za rozwiązanie zadania k przez zdającego n;

$P_{nk(j-1)r}$ – prawdopodobieństwo przyznania przez egzaminatora r kategorii punktowej j-1 za rozwiązanie zadania k przez zdającego n;

B_n – umiejętność zdającego n;

D_k – trudność zadania k;

R_r – łagodność/surowość egzaminatora r;

F_j – parametr progu j.

Za wskaźniki łagodności/surowości egzaminatorów przyjęto wartości parametrów R_r bezpośrednio z modelu, wyrażone w logitach. Ma to z jednej strony zaletę, gdyż posługujemy się miarą na tej samej skali co trudność zadań, ale i wadę, którą jest arbitralność tej skali. Niestety, model nie pozwala na oszacowanie precyzji egzaminatorów ani efektów związanych z przekształceniem skali. Autorzy byli zainteresowani różnicą w łagodności/surowości między poszczególnymi zespołami egzaminatorów oraz zespołami koordynacji. Zagregowane wskaźniki dla poszczególnych grup uzyskano na podstawie średnich ze wskaźników dla poszczególnych egzaminatorów.

Analiza wariancji

Rozwinięciem zaprezentowanego wyżej prostego modelu analizy wariancji może być dalsza dekompozycja wariancji na różne

elementy składowe związane z ocenianiem. Naturalne wydaje się rozszerzenie modelu tak, aby uwzględniał interakcję oceniającego z uczniem (a w zasadzie z jego pracą), lecz można jeszcze wyróżnić inny składnik wariancji, np. pochodzącą ze zróżnicowania zadań i interakcji zdającego i oceniającego z zadaniem (Scheerens i in., 2003). Taki rozszerzony model można przedstawić następująco:

$$\sigma_X^2 = \sigma_n^2 + \sigma_k^2 + \sigma_r^2 + \sigma_{nk}^2 + \sigma_{nr}^2 + \sigma_{kr}^2 + \sigma_e^2, \quad (2)$$

gdzie:
σ_X^2 – całkowita wariancja oceny;
σ_n^2 – wariancja pochodząca od ucznia (pracy);
σ_k^2 – wariancja pochodząca od zadania;
σ_r^2 – wariancja pochodząca od egzaminatora;
$\sigma_{nk}^2, \sigma_{nr}^2, \sigma_{kr}^2$ – wariancja pochodząca od interakcji: ucznia i zadania, ucznia i egzaminatora, egzaminatora i zadania;
σ_e^2 – wariancja błędu.

Precyzję egzaminatorów w takim modelu, podobnie jak we wcześniejszym przykładzie użycia analizy wariancji, można obliczyć jako stosunek wariancji wyjaśnionej przez wybrane czynniki do całkowitej wariancji. Autorzy proponują dwa sposoby obliczania precyzji (zwanej przez nich rzetelnością), zależne od tego, czy wariancję zadań i egzaminatorów uznamy za błąd, czy nie, stosując je odpowiednio do ocen bezwzględnych i względnych. Wskaźnik dla ocen bezwzględnych ma postać:

$$\rho^2 = \frac{\sigma_n^2}{\sigma_n^2 + \sigma_k^2/N_k + \sigma_r^2/N_r + \sigma_{nk}^2/N_k + \sigma_{nr}^2/N_r + \sigma_{kr}^2/N_k N_r + \sigma_e^2/N_k N_r}, \quad (3)$$

gdzie:
N_k – liczba zadań;
N_r – liczba egzaminatorów.

Jeśli chcemy dokonać względnej oceny uczniów i uznamy, że wyniki te chcemy porównywać z pominięciem trudności zadań i łagodności/surowości egzaminatorów, to należy wykluczyć ich wkład w całkowitą wariancję, zatem omawiany wskaźnik ma postać:

$$\rho^2 = \frac{\sigma_n^2}{\sigma_n^2 + \sigma_{nk}^2/N_k + \sigma_{nr}^2/N_r + \sigma_e^2/N_k N_r}. \quad (4)$$

Cees Glas (2012) nazywa wskaźnik dla ocen bezwzględnych zgodnością, a dla ocen względnych rzetelnością, co jest spójne z opisem zjawiska przedstawionym w pierwszej części artykułu. Uznaje on jednak, że do całkowitej wariancji – w przypadku ocen względnych – ma wkład również interakcja egzaminatora z zadaniem. W takim wypadku do mianownika równania (4) należy dodać jeszcze wyrażenie $\sigma_{kr}^2/N_k N_r$. Tak zdefiniowane wskaźniki są według niego właściwe, gdy potraktujemy efekt zadań jako efekt losowy. W wypadku, gdy efekty zadań uznamy za stałe, wskaźnik zgodności będzie miał postać:

$$\rho^2 = \frac{\sigma_n^2 + \sigma_{nk}^2}{\sigma_n^2 + \sigma_r^2/N_r + \sigma_{nk}^2 + \sigma_{nr}^2/N_r + \sigma_{kr}^2/N_r + \sigma_e^2/N_r}, \quad (5)$$

a wskaźnik rzetelności:

$$\rho^2 = \frac{\sigma_n^2 + \sigma_{nk}^2}{\sigma_n^2 + \sigma_r^2/N_r + \sigma_{nk}^2 + \sigma_e^2/N_r}. \quad (6)$$

Pomimo możliwości obliczenia zgodności lub rzetelności egzaminatorów, taki model nie oferuje żadnych parametrów umożliwiających pomiar efektów z innych grup, jak łagodność/surowość czy zniekształcenia skali. Oczywiście, podobnie jak we wcześniejszym przypadku użycia analizy wariancji, można posłużyć się odchyleniem ocen egzaminatorów od średnich ocen danej pracy jako miarą łagodności/surowości egzaminatora.

Modele IRT

Richard Patz i Brian Junker (1999) stworzyli model bardzo zbliżony do modelu Linacre'a (MFRM), pozwalający na obliczenie parametru łagodności/surowości egzaminatora. Co bardzo istotne, w odróżnieniu od

wieloaspektowego skalowania Rascha, parametr ten odnosi się do interakcji oceniającego z zadaniem. W modelu tym uwzględniono również dyskryminację zadań i ma on następującą postać:

$$ln\left(\frac{P_{nkjr}}{P_{nk(j-1)r}}\right) = a_k(\theta_n - b_{kj} - \rho_{rk}), \quad (7)$$

gdzie:

P_{nkjr} – prawdopodobieństwo przyznania przez egzaminatora r kategorii punktowej j za rozwiązanie zadania k przez zdającego n;

$P_{nk(j-1)r}$ – prawdopodobieństwo przyznania przez egzaminatora r kategorii punktowej j-1 za rozwiązanie zadania k przez zdającego n;

θ_n – umiejętność zdającego n;

a_k – dyskryminacja zadania k;

b_{kj} – parametr progu j dla zadania k;

ρ_{rk} – łagodność/surowość oceniającego r dla zadania k.

Niestety model ten, podobnie jak MFRM, nie pozwala na oszacowanie precyzji oceniania, co jest sporą wadą. Jednym ze sposobów jej oszacowania mogłoby być użycie równolegle dekompozycji wariancji, ale wygodniejsze byłoby rozwiązanie, gdzie za pomocą jednego modelu można otrzymać jak najwięcej parametrów związanych z poszczególnymi składowymi efektu egzaminatora.

Wynikiem dalszych prac nad modelami opartymi na IRT jest *hierarchical rater model* (HRM; Patz i in., 2002). Założeniem nieulegającym zmianie w stosunku do innych modeli jest to, że uczniowie odpowiadają na zadania pod warunkiem posiadanej przez nich ukrytej cechy (umiejętności). Na drugim poziomie tego hierarchicznego modelu znajdują się jednak, również nieobserwowalne, „prawdziwe" oceny uczniów, a więc takie, które nie są obciążone efektem egzaminatora. Dopiero na najniższym poziomie egzaminatorzy dokonują obserwowalnej oceny za pomocą skali przeznaczonej dla danego zadania.

Wymienieni wyżej autorzy wskazują, że eliminacja poziomu „prawdziwych" ocen uczniów w ramach MFRM i pominięcie zagnieżdżenia ostatecznych ocen wewnątrz uczniów i egzaminatorów prowadzi do nieprawidłowego oszacowania błędów standardowych. Z tego powodu proponują oni użycie modelu z poziomem „prawdziwych" ocen uczniów, które mogą być modelowane na przykład na postawie *partial credit model* (PCM). Na najniższym poziomie takiego modelu interesują nas prawdopodobieństwa przyznania przez egzaminatora kategorii oceny j pod warunkiem kategorii oceny „prawdziwej" η. Autorzy sugerują, aby prawdopodobieństwa te były proporcjonalne do gęstości rozkładu normalnego dla danej kategorii oceny ze średnią zależną od łagodności/surowości egzaminatora i odchyleniem standardowym zależnym od jego precyzji. Można zatem najniższy poziom tego modelu zapisać następująco:

$$P(Y_{nkr} = j|\eta_{nk} = \eta) \propto \exp\left\{-\frac{1}{2\psi_r^2}[j - (\eta + \varphi_r)]^2\right\}, \quad (8)$$

gdzie:

$Y_{nkr} = j$ – ocena (kategoria) ucznia n w zadaniu k przypisana przez egzaminatora r równa j;

$\eta_{nk} = \eta$ – „prawdziwa" ocena (kategoria) ucznia n w zadaniu k równa η;

ψ_r – precyzja egzaminatora r;

φ_r – łagodność/surowość egzaminatora r.

Do zdecydowanych zalet tego modelu należy zaliczyć możliwość oszacowania zarówno łagodności/surowości egzaminatora, jak i jego precyzji. Wadą jest to, że parametry te są szacowane dla egzaminatora niezależnie od zadania. Można jednak rozszerzyć model tak, aby uwzględniał interakcję oceniającego z zadaniem przez obliczenie precyzji i łagodności/surowości egzaminatora dla każdego z zadań. Również rozszerzenie tego modelu, aby na drugim poziomie uwzględniał dyskryminację

zadań, nie nastręcza trudności – wystarczy zamiast PCM użyć *generalized partial credit model* (GPCM). Pomimo takich zabiegów model nadal nie pozwala na oszacowanie efektów egzaminatora związanych ze zniekształceniami skali.

Oprócz wspomnianych wyżej ograniczeń HRM, jego autorzy wskazują, iż dla egzaminatorów o wysokiej precyzji, parametry łagodności/surowości są obciążone dużym błędem (Patz i in., 2002). Rozwiązaniem tego problemu jest zmiana sposobu szacowania prawdopodobieństw przyznania oceny przez egzaminatora na najniższym poziomie w innym hierarchicznym modelu – *hierarchical rater model with signal detection theory* (HRM-SDT; DeCarlo i in., 2011). Jego autorzy zakładają, że decyzja oceniającego odnośnie do oceny przyznawanej uczniowi za zadanie jest uzależniona od percepcji jakości odpowiedzi ucznia na zadanie, a percepcja oceniającego jest ukrytą, ciągłą zmienną losową. W przypadku konkretnego zadania percepcja jest realizacją z rodziny normalnych lub logistycznych rozkładów prawdopodobieństwa, z innym parametrem położenia dla każdej kategorii „prawdziwej" odpowiedzi ucznia. Odległość pomiędzy poszczególnymi rozkładami (ich parametr położenia) jest zależny od zdolności oceniającego do rozróżnienia ukrytych kategorii „prawdziwych", pozwala zatem na oszacowanie precyzji oceniającego. Drugim założeniem poczynionym na potrzeby tego modelu jest to, że oceniający wyznaczają progi wykonania zadania tak, aby zaklasyfikować zadanie do odpowiedniej kategorii w zależności od tego, pomiędzy którymi progami znajduje się ich percepcja. Dla *J* kategorii, które mogą przydzielić oceniający, wyznaczanych jest *J*-1 progów – poniżej pierwszego progu przyznawana jest pierwsza kategoria, a powyżej ostatniego najwyższa kategoria. Liczba ukrytych, „prawdziwych" kategorii nie musi być równa liczbie kategorii, które mogą przyznać egzaminatorzy. Konstruowanie skal ocen opiera się jednak na przekonaniu, że można rozróżnić tyle poziomów umiejętności, ile kategorii używanej skali. Można zatem uznać, że liczba kategorii skali używanej przez oceniających jest równa liczbie ukrytych kategorii, do której należy badana umiejętność.

HRM-SDT na pierwszym poziomie można przedstawić następująco:

$$P(Y_{nkr} \leq j | \eta_{nk} = \eta) = \\ = F\left[c_{kjr} - d_{kr}(\eta - 1)\right], \tag{9}$$

gdzie:

$Y_{nkr} \leq j$ – ocena (kategoria) ucznia *n* w zadaniu *k* przypisana przez egzaminatora *r* mniejsza lub równa *j*;

$\eta_{nk} = \eta$ – „prawdziwa" ocena (kategoria) ucznia *n* w zadaniu *k* równa η;

F – dystrybuanta rozkładu normalnego lub logistycznego;

c_{kjr} – próg kategorii *j* w zadaniu *k* dla egzaminatora *r*;

d_{kr} – precyzja egzaminatora *r* w zadaniu *k*.

Jeśli przyjmiemy, że rozkład prawdopodobieństwa percepcji egzaminatorów należy do rodziny logistycznych rozkładów prawdopodobieństwa, to równanie (9) opisuje kumulatywne prawdopodobieństwa dla poszczególnych kategorii zadania w *graded response model* (GRM). Jest to, obok GPCM, szeroko stosowany model IRT dla zmiennych wielokategorialnych. Różnica w stosunku do GRM polega na tym, że w przypadku HRM-SDT mamy do czynienia z dyskretną, a nie ciągłą, zmienną ukrytą.

Przy użytej w równaniu (9) parametryzacji położenie progów zależne jest od precyzji danego egzaminatora. Z tego powodu trudno ustalić „idealne" progi, dla „idealnego" oceniającego wolnego od efektu egzaminatora (a konkretnie efektów łagodności/surowości i zniekształceń skali), gdyż one

również będą zależne od precyzji. Śmiało można przyjąć założenie, że dla „idealnego" oceniającego, moment, w którym powinna następować decyzja o przyznaniu kategorii j+1 powinien następować wtedy, kiedy prawdopodobieństwo przyznania tej kategorii staje się wyższe niż prawdopodobieństwo przyznania kategorii j. Przy równej liczbie „prawdziwych" kategorii i liczbie kategorii używanej przez egzaminatorów z takim progiem mamy do czynienia, kiedy stosunek prawdopodobieństw percepcji dla sąsiednich kategorii „prawdziwych" jest równy 1. Ma to miejsce w połowie odległości między położeniem tych rozkładów wyznaczanym przez parametr d. Warto zatem dokonać przekształcenia modelu do następującej postaci:

$$P(Y_{nkr} \leq j|\eta_{nk} = \eta) = \\ = F\left[-d_{kr}\,(\eta - 1 - c_{kjr})\right]. \quad (10)$$

Dzięki takiej parametryzacji progi będą wyznaczane na skali zmiennej η, a więc będą bezpośrednio odnosiły się do kategorii i będą niezależne od precyzji. Wartość „idealnego" progu c_j będzie w takiej sytuacji równa j-0,5.

Z przedstawionych modeli, HRM-SDT ma największy potencjał uchwycenia różnych aspektów efektu egzaminatora. Poprzez odniesienie progów wyznaczonych dla egzaminatora do „idealnych" progów daje możliwość oszacowania efektów zniekształcenia skali (ograniczenia skali czy używania ocen skrajnych), czego nie umożliwiają pozostałe modele.

Parametry modelu HRM-SDT

Z opisanego powyżej modelu HRM-SDT otrzymujemy dla każdego egzaminatora dwie grupy parametrów: d_{kr}, które oznaczają precyzję egzaminatora w danym zadaniu, a także c_{kjr}, które wyznaczają progi służące do przyznawania poszczególnych kategorii w danym zadaniu. Dzięki porównaniu położenia progów oszacowanych dla poszczególnym egzaminatorów do progów „idealnego" oceniającego, możliwe jest uchwycenie kilku istotnych efektów egzaminatora.

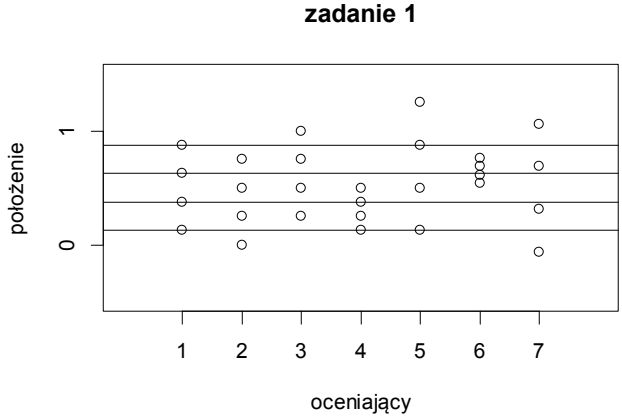
Łagodność oceniającego można definiować jako umieszczanie przez niego progów poszczególnych kategorii poniżej progów kategorii „idealnych", natomiast surowość – powyżej. Oznacza to, iż łagodny oceniający przypisuje kolejną kategorię przy niższym poziomie umiejętności ucznia niż „idealny" egzaminator. Surowy oceniający natomiast przypisuje kolejną kategorię przy wyższym poziomie tej cechy.

Jeśli egzaminator swój pierwszy próg umieści dużo przed pierwszym progiem „idealnego" egzaminatora, a ostatni dużo za ostatnim „idealnym" progiem, będzie to oznaczało, iż w praktyce nie będzie on przyznawał pierwszej i ostatniej kategorii (por. egzaminator 7 na Rysunku 1). Jest to przejaw tendencji centralnej oceniającego. Analogicznie, jeśli egzaminator skrajne progi umieści odpowiednio powyżej pierwszego „idealnego" progu i poniżej ostatniego „idealnego" progu, i jednocześnie pozostałe progi przesunie w kierunku środkowych kategorii, to będziemy mieli do czynienia z efektem używania ocen skrajnych (por. egzaminator 6 na Rysunku 1). W ten sposób można również identyfikować ograniczenia skali w dowolnym jej obszarze.

Interpretacja wartości liczbowych precyzji (d_{kr}) nastręcza jednak pewnych trudności. W przeprowadzonych przez zespół Lawrence'a DeCarlo (2011) symulacjach, wartości parametrów d_{kr} wahały się w granicach 1–6 ze średnią około 3,5 i miały rozkład zbliżony do normalnego. Nie podają oni żadnego punktu odniesienia czy zakresu dla tego parametru, który pozwalałby na stwierdzenie, że precyzja jest niska, średnia czy wysoka. Jest to parametr względny

i można jedynie z pewnością stwierdzić, że wraz z jego wzrostem rośnie precyzja oceniającego. Trudno nawet oszacować wielkość tego efektu, porównując dwóch oceniających w tym samym zadaniu. Można jednak szacować precyzję egzaminatorów w nieco inny sposób. Działanie modelu dość mocno opiera się na estymowaniu „prawdziwych" kategorii, do których należą odpowiedzi uczniów w poszczególnych zadaniach. Można ich użyć do obliczenia, jak często egzaminatorzy poprawnie przydzielili poszczególne kategorie, a w ilu przypadkach popełnili błąd. Zatem dzięki porównaniu ocen przyznanych przez egzaminatorów do „idealnych" ocen prac jesteśmy w stanie obliczyć procent poprawnych klasyfikacji, który jest bardzo intuicyjnym sposobem mierzenia precyzji. Taki wskaźnik można łatwo agregować dla zadań i egzaminatorów, np. przez obliczenie średniej. Pozwala to na identyfikację mało precyzyjnych oceniających, a dodatkowo może też wskazywać zadania sprawiające problemy z przydzieleniem kategorii adekwatnej do posiadanego przez uczniów poziomu umiejętności.

Autorzy modelu proponują wizualny sposób identyfikacji pozostałych efektów. Dla każdego zadania należy sporządzić osobny wykres, na którym umieszczane są tzw. względne położenia, a więc położenie progów dla każdego egzaminatora na skali zadania przekształconej tak, aby zakres skali wynosił 0–1. Przekształcenie skali ma na celu zapewnienie porównywalności między oceniającymi, gdyż położenie kryteriów przy parametryzacji użytej przez autorów zależy od parametru *d*. Obok reprezentowanych przez znaczniki progów poszczególnych oceniających za pomocą linii wyznaczone są progi „idealne". Jeśli użyta zostanie parametryzacja z równania (10), to nie zachodzi potrzeba przekształcania skali, gdyż parametry *c* są niezależne od parametrów *d*, a jedynie od liczby kategorii danego zadania. Przykładowy wykres przedstawiono na Rysunku 1. Widać na nim odpowiednio: „idealnego" egzaminatora (1), egzaminatora łagodnego (2) i surowego (3), łagodnego w wyższych kategoriach (4), surowego w wyższych kategoriach i ograniczającego skalę (5), używającego głównie kategorii skrajnych (6) oraz wykazującego tendencję centralną (7).



Rysunek 1. Wizualizacja położenia progów z modelu HRM-SDT na przykładzie zadania z pięcioma kategoriami i siedmioma hipotetycznymi egzaminatorami.

Okręgi oznaczają położenie progów poszczególnych egzaminatorów, a poziome linie progi „idealne". Na podstawie (DeCarlo i in., 2011).

O ile w przypadku niewielkiej liczby oceniających i niewielkiej liczby zadań wizualny sposób identyfikacji efektów egzaminatora jest akceptowalny, o tyle przy dużej liczbie oceniających i zadań zaczyna on stanowić problem, gdyż wymaga analizy każdej interakcji zadanie–egzaminator. Autorzy modelu nie proponują niestety żadnych liczbowych wskaźników dla poszczególnych efektów. W niektórych przypadkach możemy być zainteresowani również zagregowanymi miarami efektów dla grup egzaminatorów (zespołów) czy zadań. Choć można obliczyć średnie położenia progów dla zespołu egzaminatorów dla jednego zadania, to taka operacja dla zadań różniących się liczbą kategorii staje się problematyczna. Potrzebna jest w związku z tym, niezależna od liczby kategorii, ogólna miara poszczególnych efektów, aby można było dokonywać porównań między zadaniami. Oczywiście, wiąże się to z utratą części informacji, lecz w celu identyfikacji przypadków odstających egzaminatorów czy zadań można przyjrzeć się poszczególnym progom tylko w interesujących nas przypadkach (również w formie graficznej).

Sposobem na określenie ogólnego wskaźnika łagodności/surowości egzaminatora w danym zadaniu może być posłużenie się różnicą między położeniem progów „idealnych" i progów tego egzaminatora. Jeśli dla każdego progu w zadaniu obliczymy taką różnicę, a następnie obliczymy średnią, to otrzymamy wskaźnik informujący nas o tym, o ile (średnio) kategorii w tym zadaniu dany egzaminator zawyża (wartości dodatnie – łagodność) lub zaniża (wartości ujemne – surowość) ocenę uczniów. Wartość 0 oznaczałaby oceniającego niewykazującego tego efektu. Nazwijmy ten parametr średnim odchyleniem od „idealnych" progów i oznaczmy symbolem σ_{kr}.

Zniekształcenia skali można z kolei powiązać z rozstępem wartości parametrów c_{kjr} (położenia progów). Oczywiście, rozstęp zależny jest od liczby kategorii w zadaniu, co utrudnia porównywanie zadań o różnej liczbie kategorii. Można temu zaradzić przez podzielenie rozstępu przez rozstęp dla „idealnych" progów, a więc liczbę kategorii w zadaniu pomniejszoną o 2. Oznaczmy taki parametr jako r_{kr}. Parametr równy 0 wskazywałby na używanie jedynie skrajnych kategorii zadania, a więc ekstremizm. Wartość równa 1 oznaczałaby, że oceniający równomiernie używa wszystkich kategorii, choć może być łagodny lub surowy. Im wyższa wartość parametru r_{kr}, tym egzaminator bardziej ograniczałby skalę, choć nie oznaczałoby to tendencji centralnej, gdyż ograniczenie może następować w różnych miejscach skali. W połączeniu z wartością parametru σ_{kr} jesteśmy w stanie rozpoznać, gdzie następuje ograniczenie.

Tabela 1 zawiera zestawienie obydwu proponowanych parametrów dla omawianych wyżej hipotetycznych egzaminatorów, dla których progi przedstawiono na Rysunku 1. Oceniający nr 4 i nr 6 mają zbliżone wartości parametru r_{kr} – odpowiednio 0,5 i 0,3, co świadczy o częstszym używaniu ocen skrajnych. Dodając jednak informację o tym, że oceniający nr 4 jest łagodny (σ_{kr} = 0,75), a oceniający nr 6 surowy (σ_{kr} = -0,6), możemy stwierdzić, że pierwszy z nich częściej używa najwyższej kategorii niż pierwszej, a drugi odwrotnie. W przypadku egzaminatorów nr 5 i nr 7 mamy natomiast taką samą wartość parametru r_{kr} równą 1,5, a zatem ograniczają oni skalę. Egzaminator nr 5 jest jednak surowy (σ_{kr} = -0,75), zatem rzadziej przydziela najwyższe kategorie, a egzaminator nr 7 wykazuje tendencję centralną (σ_{kr} = 0).

Użycie razem wskaźników σ_{kr} i r_{kr} pozwala na identyfikację kilku istotnych efektów egzaminatora. Mogą one posłużyć do identyfikacji nietypowych oceniających, a także nietypowych zadań (na podstawie średniej z parametrów wszystkich egzaminatorów dla danego zadania). Dzięki tym wskaźnikom opisywany model można

Tabela 1
Zestawienie parametrów σ_{kr} i r_{kr} z modelu HRM-SDT dla przykładowych oceniających

Oceniający	1	2	3	4	5	6	7
σ_{kr}	0,00	0,50	-0,50	0,75	-0,75	-0,60	0,00
r_{kr}	1,00	1,00	1,00	0,50	1,50	0,30	1,50

stosować nawet w przypadku dużej liczby egzaminatorów i zadań, bez konieczności weryfikacji dużej liczby wykresów.

Symulacje dla modelu HRM-SDT

W 2013 r. w Instytucie Badań Edukacyjnych rozpoczęto prace mające na celu zbadanie porównywalności oceniania między okręgowymi komisjami egzaminacyjnymi (OKE) i oszacowanie efektu egzaminatora w odniesieniu do egzaminów maturalnych z języka polskiego (poziom podstawowy) i matematyki (poziom podstawowy i rozszerzony). Do badań wybrano losowo 232 egzaminatorów z całego kraju (po 29 osób z każdej OKE) dla obydwu przedmiotów oraz po 897 prac maturalnych każdego typu z lat 2011 i 2012. Przy tak dużej liczbie prac i oceniających wykorzystanie pełnego schematu przydziału prac do egzaminatorów, w którym każdy ocenia każdą pracę, byłoby ogromnie czasochłonne i kosztowne. W związku z tym użyto niepełnego schematu, w którym każda praca z języka polskiego była oceniona ośmiokrotnie, a z matematyki czterokrotnie. W konsekwencji wymagało to użycia odpowiedniego modelu analiz, uwzględniającego założone braki danych.

Opisywany wyżej model HRM-SDT został przez jego autorów przetestowany na danych pochodzących od 2350 zdających test językowy, którzy pisali dwa eseje punktowane na skali 1–5. Każdy esej był oceniany przez 2 z 54 oceniających, przy czym pierwszy esej oceniało 34 egzaminatorów, drugi – 33, a 13 oceniających przyznawało oceny za obydwa eseje, lecz dla różnych uczniów (DeCarlo i in., 2011). Był to więc niepełny schemat przydziału zadań do oceniających, podobnie jak w prowadzonym przez IBE badaniu. Autorzy modelu przytaczają również wyniki wcześniejszych symulacji, które według nich świadczą o przydatności modelu i dobrym odtwarzaniu parametrów użytych do tych symulacji, zarówno w pełnych, jak i niepełnych schematach przydziału prac do oceniających. Wskazują oni jednak na to, że zaledwie dwa zadania użyte do estymacji mogą dawać słabo oszacowane parametry na drugim poziomie modelu (trudność i dyskryminacja zadań). Estymowali oni model z użyciem częściowego podejścia Bayesowskiego, mianowicie *posterior mode estimation* (PME), choć można tego dokonać zarówno metodą największej wiarygodności (m*aximum likelihood estimation*, MLE), jak i pełnej analizy Bayesowskiej z użyciem metody *Markov chain Monte Carlo* (MCMC; DeCarlo i in., 2011; Patz i in., 2002).

W danych maturalnych użytych w prowadzonym przez IBE badaniu mamy do czynienia z więcej niż dwoma zadaniami w każdym z użytych testów. W przypadku matematyki jest to między 9 i 12 zadań (w zależności od testu), a dla języka polskiego każde wypracowanie oceniane jest na sześciu skalach (kryteriach). Dzięki temu oszacowania parametrów zadań na drugim poziomie modelu HRM-SDT powinny być dość dobre. Również większa liczba ocen pojedynczej pracy (4 dla matematyki, 8 dla języka polskiego) powinna zaowocować dobrym dopasowaniem modelu do danych.

Ze względu na ponad czterokrotnie większą liczbę egzaminatorów (232), a także znacząco mniejszą liczbę prac (897 dla każdego testu), w porównaniu do sytuacji opisywanej przez autorów modelu, warto przeprowadzić dodatkowe symulacje, aby sprawdzić przydatność modelu do analizy danych zebranych na podstawie maturalnych arkuszy egzaminacyjnych.

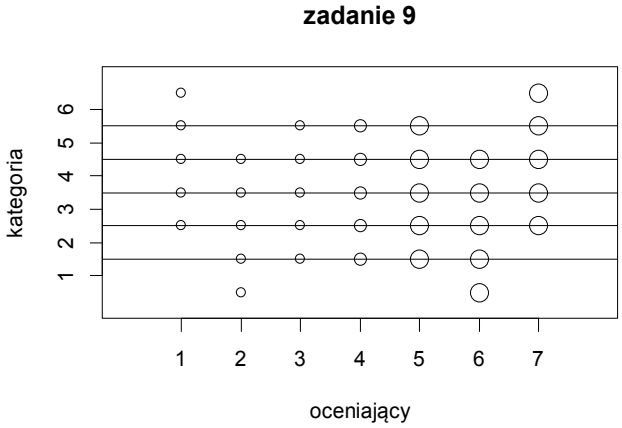
Jedną z części procedury oceniania podczas egzaminu maturalnego jest podwójna ocena minimum 10% losowo wybranych prac. Elastyczność wybranego modelu powinna umożliwić analizę efektu egzaminatora na podstawie tak uzyskanych danych. Schemat przydziału prac do oceniających podczas podwójnego oceniania w trakcie sesji egzaminacyjnych również jest schematem niepełnym, podobnie jak w opisywanych badaniach, choć jest on mniej skomplikowany ze względu na dwukrotną ocenę jednej pracy.

Estymacji modeli HRM-SDT w przeprowadzonych na potrzeby tego artykułu symulacjach dokonano pełną metodą Bayesowską z użyciem MCMC za pomocą pakietu statystycznego R (biblioteka rjags) i programu JAGS 3.4.0. Każdorazowo estymowano jeden łańcuch liczący 1000 iteracji, z czego 500 posłużyło za *burn-in*, a pozostałe 500 do oszacowania wartości parametrów. Przyjęte rozkłady a priori dla poszczególnych parametrów modeli przedstawiono w Tabeli 2. Dla progów (c_{kjr}) wybrano rozkład normalny ze średnią odpowiadającą położeniu „idealnego" progu dla danej kategorii i odchyleniem standardowym równym 1, przy czym zgodnie z wymogami GRM, wartości dla kolejnych progów posortowano rosnąco. Dla precyzji (d_{kr}) wybrano rozkład normalny ze średnią równą 3,5 (taką średnią otrzymali autorzy modelu w swoich analizach) i odchyleniem standardowym 1. Dodatkowo wartości precyzji ograniczono tak, aby były one większe lub równe 0,01. Symulacje przeprowadzono w dwóch etapach.

W pierwszej kolejności wygenerowano dane dla pełnego schematu przydziału prac do egzaminatorów (każdy egzaminator ocenia każdą pracę) z użyciem danych z egzaminu maturalnego na poziomie podstawowym z matematyki z 2011 r. Było to 10 zadań, o liczbie kategorii 3–6 (odpowiednio wg kolejności zadań: 3, 3, 3, 3, 3, 3, 3, 5, 6, 5). W celu wygenerowania danych potrzebne były „prawdziwe" kategorie oceny (η_{nk}) dla każdego ucznia w każdym zadaniu z pracy, a także precyzja (d_{kr}) i progi (c_{kjr}) dla każdego egzaminatora w każdym zadaniu. „Prawdziwe" kategorie otrzymano poprzez zastosowanie modelu do rzeczywistych ocen 868 prac maturalnych użytych w badaniu (nie dla wszystkich prac udało się połączyć oceny z egzaminu). Jako rzeczywiste oceny mogły zostać użyte bezpośrednio wyniki egzaminu (pomimo ich obciążenia efektem egzaminatora), ale dzięki takiemu zabiegowi otrzymany rozkład kategorii powinien być bardziej zbliżony do rozkładu „prawdziwych" ocen. Precyzję i progi dla poszczególnych egzaminatorów ustalono arbitralnie tak, aby zasymulować oceniających różniących się precyzją i łagodnością/surowością. Wygenerowano oceny siedmiu

Tabela 2
Parametry rozkładów a priori dla parametrów modelu HRM-SDT użytych w symulacjach

Parametr	Rozkład	*M*	*SD*	Uwagi
a_k	normalny	1	1	Wartości ograniczono do $a_k \geq 0{,}01$.
b_{kj}	normalny	0	1	Wartości sortowano rosnąco dla kolejnych kategorii.
d_{kr}	normalny	3,5	1	Wartości ograniczono do $d_{kr} \geq 0{,}01$.
c_{kjr}	normalny	*j*-0,5	1	Wartości sortowano rosnąco dla kolejnych kategorii.



Rysunek 2. Wizualizacja parametrów modelu HRM-SDT użytych do wygenerowania danych symulacyjnych na przykładzie zadania z 6 kategoriami i 7 hipotetycznymi egzaminatorami.

Okręgi oznaczają położenie progów poszczególnych egzaminatorów, a poziome linie progi „idealne". Wielkość okręgów odpowiada precyzji oceniających przeskalowanych do zakresu 1–2.

egzaminatorów, z precyzją równą odpowiednio: 1,5; 1,5; 1,5; 3,5; 6,5; 6,5; 6,5. Wartości progów ustalono w taki sposób, że we wszystkich zadaniach egzaminatorzy nr 1 i nr 7 każdy z progów mieli przesunięty o +1 w stosunku do progów „idealnych", egzaminatorzy nr 2 i nr 6 o -1, a pozostałym (nr 3–5) ustalono progi „idealne". Wykres wygenerowanych parametrów dla przykładowego zadania o sześciu kategoriach znajduje się na Rysunku 2. Wielkość okręgów wyznaczających położenie progów na wykresie zależy od precyzji egzaminatora – wartości zostały przeskalowane do zakresu 1–2, gdzie najmniejszy okrąg na wykresie odpowiada przeskalowanej wartości $d_{kr} = 1$, a największy przeskalowanej wartości $d_{kr} = 2$.

Użyte przy generowaniu wartości zostały bardzo dobrze odtworzone przez model. Oszacowane „prawdziwe" kategorie ocen (η_{nk}) w ponad 99% przypadków zgadzały się z tymi, które posłużyły do wygenerowania danych. Różnice pomiędzy progami oszacowanymi w modelu i zastosowanymi do wygenerowania danych były minimalne; średnia różnic wyniosła 0, odchylenie standardowe 0,08, a ich rozkład był normalny. W przypadku precyzji odzyskane parametry nieco odbiegały od pierwotnych: średnia różnica wyniosła 0,17, odchylenie standardowe 0,41, a rozkład był wyraźnie prawoskośny. Powodem skośności rozkładu były niższe oszacowania najbardziej precyzyjnych egzaminatorów (nr 5–7, założone $d_{kr} = 6{,}5$). Szczególnie w przypadku egzaminatorów nr 6 i 7, którzy charakteryzowali się dodatkowo łagodnością lub surowością, różnice były największe i sięgały 1,57, choć średnio było to 0,51. Dla egzaminatora nr 5 średnia różnica wyniosła 0,35, a maksymalna 0,93. Dla oceniających nr 1–4 różnice w precyzji były minimalne, średnio została ona oszacowana o 0,04 niżej niż założona. Tabela 3 przedstawia podsumowanie rozkładów różnic między oszacowanymi parametrami c_{kjr} i d_{kr} a ich wartościami wykorzystanymi do wygenerowania danych.

Drugim sposobem weryfikacji przydatności modelu było sprawdzenie, jak odtwarza on parametry przy analizie danych zebranych wg rozbudowanego, niepełnego schematu przydziału prac zastosowanego w badaniu IBE. Analizom poddano również

Tabela 3
Parametry rozkładów różnic precyzji i progów oszacowanych w modelu HRM-SDT i użytych do wygenerowania danych (symulacja dla pełnego schematu przydziału prac do egzaminatorów)

Parametr	Min	*q*1	*Me*	*q*3	Max	*M*	*SD*
d_{kr}	-0,33	-0,10	0,02	0,33	1,57	0,17	0,41
c_{kjr}	-0,20	-0,04	0,00	0,04	0,50	0,00	0,08

dane z egzaminu z matematyki, lecz tym razem użyto zadań zarówno z poziomu podstawowego, jak i rozszerzonego z obydwu badanych lat (2011 i 2012). Łącznie te cztery testy obejmują 42 zadania oceniane na skali o długości 3–7 (odpowiednio liczba zadań: 13, 4, 16, 4 i 5). W tym przypadku również model estymowany był dwukrotnie – parametry precyzji, progi i „prawdziwe" oceny otrzymane z pierwszej estymacji zostały użyte do wygenerowania danych do drugiego oszacowania. Pozwoliło to na sprawdzenie oszacowań całej gamy występujących w rzeczywistości efektów egzaminatora bez konieczności arbitralnego przypisywania wartości parametrów d_{kr} i c_{kjr} dla 232 egzaminatorów.

Wygenerowane na potrzeby symulacji oceny egzaminatorów były w 85% zgodne z rzeczywistymi ocenami egzaminatorów zebranymi w badaniach – egzaminator przyznaje oceny z pewnym prawdopodobieństwem, zatem takie rozbieżności są dopuszczalne. Oceny „prawdziwe" (η_{nk}) z obydwu estymacji były natomiast zgodne w 94%, co można uznać za bardzo dobry wynik. Różnice między progami z obydwu estymacji były nieznaczne: średnia wyniosła 0, odchylenie standardowe 0,31, a rozkład był normalny. Różnice w przypadku precyzji były większe niż podczas pierwszej symulacji, lecz średnia wyniosła 0, odchylenie standardowe 0,4, a rozkład nie wykazywał skośności. W Tabeli 4 zamieszczono podsumowanie rozkładów różnic precyzji i progów z obydwu estymacji.

Jak wskazują wyniki opisanych w tej części artykułu symulacji, duża liczba egzaminatorów i skomplikowany, niepełny schemat przydziału prac nie wpływają na stabilność oszacowań parametrów. Pozwala to na stwierdzenie przydatności modelu HRM-SDT do analizy danych zebranych na podstawie arkuszy maturalnych z matematyki w prowadzonym przez IBE badaniu porównywalności oceniania i efektu egzaminatora. Jak wspomniano wcześniej, model ten powinien równie dobrze posłużyć do analizy danych uzyskanych podczas tzw. podwójnego oceniania w trakcie sesji egzaminacyjnej. Ze względu na odmienny schemat przydziału prac i mniejszą liczbę ocen pojedynczej pracy podczas egzaminów należałoby przeprowadzić dodatkowe symulacje.

Tabela 4
Parametry rozkładów różnic precyzji i progów z dwóch estymacji modelu HRM-SDT dla schematu przydziału prac do egzaminatorów użytego w badaniu IBE

Parametr	Min	*q*1	*Me*	*q*3	Max	*M*	*SD*
d_{kr}	-2,88	-0,10	0,02	0,17	2,01	0,00	0,40
c_{kjr}	-1,85	-0,10	0,00	0,10	1,67	0,00	0,31

Podsumowanie

Przegląd aktualnej literatury w zakresie efektu egzaminatora (lub szerzej: oceniającego) ukazuje duże zróżnicowanie modeli pozwalających na analizę tego zagadnienia. Mamy do czynienia zarówno z modelami opartymi na analizie wariancji, jak i na podejściu IRT. Modele znacznie różnią się między sobą i często pozwalają tylko na szacowanie pewnej grupy efektów, jak np. tylko precyzji czy tylko łagodności/surowości. Najbardziej wszechstronne możliwości daje *hierarchical rater model with signal detection theory* (HRM-SDT; DeCarlo i in., 2011). W modelu tym, stosującym podejście IRT, mamy do czynienia z odseparowaniem efektów egzaminatora od efektów zadania. Dzięki temu jesteśmy w stanie oszacować precyzję, łagodność/surowość czy efekty z grupy zniekształcenia skali dla każdego z oceniających we wszystkich zadaniach osobno.

Symulacje wykonane na danych z arkuszy maturalnych wskazują na przydatność modelu HRM-SDT do analiz tego typu. Pomimo użytego w badaniu IBE złożonego, niepełnego schematu przydziału prac do egzaminatorów i dużej liczby egzaminatorów (232), parametry odtworzone przez model niewiele odbiegały od tych zastosowanych do wygenerowania danych symulacyjnych. Zaproponowane w artykule ogólne miary łagodności/surowości i zniekształcenia skali dla zadania pozwalają na agregację zarówno na poziomie zadań, jak i egzaminatorów. Jest to istotne, gdy chcemy dokonać porównania oceniających i wypowiadać się np. o jakości systemu egzaminacyjnego. Jednym z głównych założeń prowadzonego przez IBE badania była ocena porównywalności oceniania okręgowych komisji egzaminacyjnych w zakresie egzaminów maturalnych z matematyki i języka polskiego. Wybrany model wraz z zaproponowanymi miarami powinien dobrze służyć temu celowi.

Efekt oceniającego jest zjawiskiem występującym nie tylko w pomiarze edukacyjnym, lecz wszędzie tam, gdzie dokonuje się oceny jakiejś cechy. Uniwersalność modelu HRM-SDT powinna zapewnić satysfakcjonujące wyniki jego stosowania również w takich dziedzinach, jak psychologia, socjologia czy marketing. Szczególnie przydatny może się on okazać w psychologii organizacji i zarządzania do analizy oceny pracowników czy analizy ocen respondentów w badaniach socjologicznych lub badaniach rynku.

Literatura

DeCarlo, L. T., Kim, Y. i Johnson, M. S. (2011). A Hierarchical rater model for constructed responses, with a signal detection rater model. *Journal of Educational Measurement*, *48*(3), 333–356.

Dolata, R., Putkiewicz, E. i Wiłkomirska, A. (2004). *Reforma egzaminu maturalnego – oceny i rekomendacje*. Warszawa: Instytut Spraw Publicznych.

Dubiecka, A., Szaleniec, H. i Węziak, D. (2006). Efekt egzaminatora w egzaminach zewnętrznych. W: B. Niemierko i M. K. Szmigel (red.), *O wyższą jakość egzaminów szkolnych* (98–115). Kraków: Polskie Towarzystwo Diagnostyki Edukacyjnej.

Glas, C. A. (2012). Generalizability theory and item response theory. W: T. J. Eggen i B. P. Veldkamp (red.), *Psychometrics in practice at RCEC* (1–13). Enschede: Ipskamp Drukkers.

Patz, R. J. i Junker, B. W. (1999). Applications and rxtensions of MCMC in IRT: multiple Item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, *24*(4), 342–366.

Patz, R. J., Junker, B. W., Johnson, M. S. i Mariano, L. T. (2002). The Hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics*, *27*(4), 341–384.

Saal, F. E., Downey, R. G. i Lahey, M. A. (1980). Rating the ratings: assessing the psychometric quality of rating data. *Psychological Bulletin*, *88*(2), 413–428.

Scheerens, J., Glas, C. A. i Thomas, S. M. (2003). *Educational Evaluation, Assessment and Monitoring: A Systemic Approach*. Lisse: Swets & Zeitlinger.

Scullen, S. E., Mount, M. K. i Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology*, *85*(6), 956–970.

Szmigel, M. K. i Szaleniec, H. (2001). Z prac nad porównywalnością wyników oceniania zewnętrznego. W: K. Wenta, *Pomiar edukacyjny jako kompetencje pedagogiczne.* Szczecin: Wydawnictwo Naukowe Uniwersytetu Szczecińskiego.

Wolfe, E. W. (2004). Identifying rater effects using latent trait models. *Psychology Science*, *46*(1), 35–51.