Tomasz Żółtak
Grzegorz Golonka
Instytut Badań Edukacyjnych

Czy zgadywanie ma znaczenie? Różnice w oszacowaniach umiejętności z modeli IRT 2PL i 3PL w sytuacji występowania zgadywania

Wprowadzenie

We współcześnie wykorzystywanych narzędziach służących do pomiaru dydaktycznego coraz większą rolę odgrywają zadania zamknięte, w których rola rozwiązującego zadanie sprowadza się do oceny poprawności odpowiedzi przedstawionych mu razem z treścią zadania. Posiadają one niezaprzeczalne zalety z punktu widzenia organizacji procesu oceniania testu, który pozbawiony jest elementów subiektywności i może być przeprowadzony bardzo szybko, a przy zastosowaniu odpowiednich rozwiązań technicznych wręcz automatycznie. Jednocześnie zadania tego typu często są poddawane krytyce. Jedną z najczęściej podnoszonych kwestii jest ich podatność na zgadywanie.

Choć liczba różnych działań (rozwiązań zadania), jakie może podjąć rozwiązujący zadanie, zależy od formatu zadania i schematu jego oceniania, to jednak typowo jest ona niewielka. W przypadku chyba najczęściej wykorzystywanego formatu zadania, w którym należy wybrać dokładnie jedną poprawną odpowiedź spośród kilku zaproponowanych, liczba możliwych rozwiązań jest równa po prostu liczbie zaproponowanych odpowiedzi. Jeśli jest ona niewielka – a typowo jest ich cztery lub pięć – nawet osoba, która dokonywałaby wyboru odpowiedzi w sposób całkowicie losowy (z równym prawdopodobieństwem wyboru każdego z możliwych rozwiązań), ma relatywnie wysokie prawdopodobieństwo wskazania prawidłowej odpowiedzi (odpowiednio 25% lub 20%). Rodzi to dosyć powszechną obawę, że zastosowanie tej formy zadań może prowadzić do wypaczania wyników pomiaru dydaktycznego poprzez przypisywanie części rozwiązujących test posiadanie umiejętności, których w istocie nie opanowali (Twardowska i in., 2011).

Można jednak zadać sobie pytanie, na ile współcześnie wykorzystywane modele psychometryczne pozwalają uwzględnić wpływ ewentualnego występowania zgadywania na wyniki testów. Należy przy tym zauważyć, że występowanie zgadywania wydaje się mieć różne znaczenie w zależności od celu pomiaru. Pomiar sprawdzający, mający na celu określenie, w jakim stopniu zdający spełniają wymagania określonych standardów edukacyjnych, zdaje się tu bowiem stawiać dużo większe wymagania niż pomiar różnicujący, w którym podstawowym celem jest określenie uszeregowania zdających ze względu na poziom natężenia badanej cechy (por. Niemierko, 2009). Jeśli bowiem przyjmiemy, że wszyscy rozwiązujący test „mają takie same szanse" na zgadywanie, to wydaje

się, że jego występowanie nie powinno w istotny sposób zaburzać naszych przewidywań odnośnie ich uporządkowania. Niemniej warto zadać pytanie, jakie znaczenie dla wyników pomiaru ma w takiej sytuacji wybór modelu statystycznego, którego będziemy używać do analizy wyników testu i wyliczenia oszacowań wartości mierzonej cechy.

W niniejszym artykule podejmujemy właśnie kwestię wpływu zgadywania na wyniki pomiaru różnicującego. Konkretnie, używając metod symulacyjnych, zajmujemy się tu porównaniem własności dwóch modeli IRT, szeroko wykorzystywanych do skalowania wyników testów składających się z zadań zamkniętych, z których jeden pozwala modelować występowanie zgadywania, a drugi nie. Podstawowe pytanie, jakie stawiamy, to na ile wybór jednego z nich wpływa na uzyskiwane oszacowania (punktowe) wartości mierzonej cechy. W dalszej kolejności rozpatrujemy również wpływ na błędy standardowe takich oszacowań.

W pierwszej części artykułu porównane zostaną dwa rozpatrywane modele, w drugiej omówione zostaną założenia i procedura przeprowadzonego badania symulacyjnego, a w trzeciej uzyskane wyniki.

Zgadywanie w modelach IRT

Spośród szerokiej klasy modeli IRT w niniejszym opracowaniu zajmować będziemy się własnościami tylko dwóch modeli, odpowiednich do modelowania odpowiedzi na zadania oceniane w sposób binarny (0 lub 1), w szczególności zadania zamknięte. Te dwa modele to: dwuparametryczny model logistyczny (2PL) i trzyparametryczny model logistyczny (3PL). W obu modelach zakłada się, że odpowiedzi udzielane na zadania zawarte w teście są obserwowalnymi wskaźnikami cechy (umiejętności), co do której zakłada się, że jest ona opisywana zmienną ciągłą i sama w sobie (w postaci zmiennej ciągłej) nie daje się bezpośrednio zaobserwować. Na potrzeby estymacji parametrów modelu konieczne jest jeszcze przyjęcie założenia co do rozkładu tej cechy w badanej populacji. Co do zasady przyjmuje się tu rozkład normalny standaryzowany i taką sytuację będziemy dalej rozpatrywać. Jednocześnie zakłada się, że związek pomiędzy mierzoną cechą a odpowiedzią na każde z zadań ma charakter probabilistyczny, uwarunkowany poziomem cechy oraz parametrami zadania. W modelu 2PL związek ten opisywany jest funkcją:

$$P(X_i = 1|\theta) = 1 - \frac{1}{1 + \exp\left[a_i(\theta - b_i)\right]}$$

gdzie:

θ - wartość mierzonej cechy,

α_i - parametr dyskryminacji *i*-tego zadania,

b_i - parametr trudności *i*-tego zadania.

Z kolei w modelu 3PL związek pomiędzy wartością mierzonej cechy opisywane jest wzorem:

$$P(X_i = 1|\theta) = c_i + (1 - c_i)\left(1 - \frac{1}{1 + \exp\left[a_i(\theta - b_i)\right]}\right)$$

gdzie:

θ - wartość mierzonej cechy,

α_i - parametr dyskryminacji *i*-tego zadania,

b_i - parametr trudności *i*-tego zadania,

c_i - parametr (pseudo)zgadywania *i*-tego zadania.

Graficzne reprezentacje tych wzorów można zobaczyć na rysunku 1. Łatwo zauważyć, że różnica pomiędzy tymi dwoma modelami polega na sposobie, w jaki odwzorowują one prawdopodobieństwo udzielenia poprawnej odpowiedzi przez osoby o relatywnie (w stosunku do trudności zadania) niskim natężeniu badanej cechy. W modelu 2PL zakłada się, że prawdopodobieństwo to w miarę zmniejszania się natężenia cechy stopniowo maleje aż do wartości bliskich zeru. W modelu 3PL zakłada się z kolei, że bez względu na to, jak niskie nie byłoby natężenie badanej cechy, prawdopodobieństwo udzielenia na zadanie poprawnej odpowiedzi nigdy nie spadnie poniżej poziomu wyznaczanego przez wartość parametru *c* (na rysunku 1. wynosi ona 0,2). Jedną z najbardziej oczywistych interpretacji, jaką możemy przypisać niezerowej wartości tego parametru, jest występowanie zgadywania.

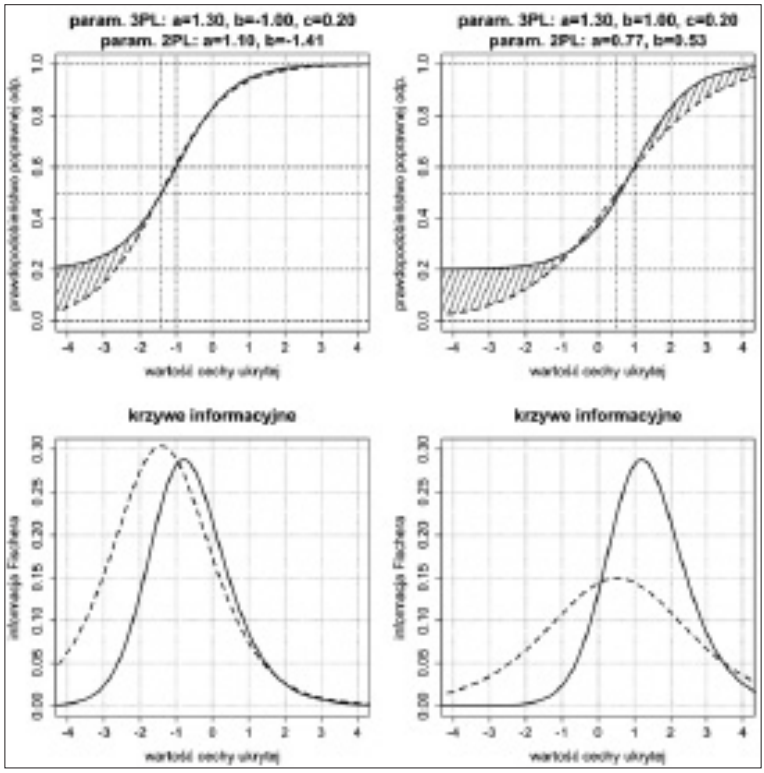
Warto przy tym zaznaczyć, że w praktyce wartość tego parametru, estymowana na podstawie danych, często wyraźnie odbiega od odwrotności liczby odpowiedzi, jakie zdający miał do wyboru. Uświadamia nam to, że zachowania zdających postawionych przed bardzo trudnym dla nich zadaniem, nawet jeśli zawierają w sobie elementy zgadywania, to jednak nie są „w pełni losowe". Dla podkreślenia tego faktu część autorów postuluje określanie go mianem parametru pseudozgadywania (Kondratek i Pokropek, 2013).

Zadajmy sobie teraz pytanie, jakie konsekwencje będzie mieć, jeśli zadanie, z wyraźnie większą od zera wartością parametru zgadywania, którego własności opisuje model 3PL spróbujemy opisać przy pomocy modelu 2PL. Na rysunku 1. liniami ciągłymi przedstawione zostały krzywe charakterystyczne dwóch zadań, z których każde charakteryzuje się wartością parametru dyskryminacji równą 1,3, wartością parametru zgadywania równą 0,2 i wartościami parametru trudności równymi odpowiednio -1 (lewy górny panel) i 1 (prawy górny panel). Na rysunek naniesione zostały też linie przerywane, obrazujące przebieg krzywych charakterystycznych z modeli 2PL, których wartości parametrów zostały dobrane w ten sposób, aby zminimalizować wariancję różnic pomiędzy prawdziwym prawdopodobieństwem udzielenia poprawnej odpowiedzi (wynikającym z modelu 3PL) a analogicznym prawdopodobieństwem przewidywanym na podstawie modelu 2PL. Przyjęto przy tym założenie, że w badanej populacji cecha ma rozkład normalny standaryzowany, skutkiem czego w procesie optymalizacji większa waga przykładana jest do różnic w przebiegu krzywych charakterystycznych w pobliżu punktu 0 na osi X (tj. tam, gdzie występuje dużo osób) niż do różnic w punktach leżących dalej po bokach (gdzie, zgodnie z przyjętym założeniem, jest dużo mniej badanych osób).

Możemy tu zauważyć, że parametry dyskryminacji i trudności uzyskane z modeli 2PL zdecydowanie odbiegają od prawdziwych – z modeli 3PL – co wobec wysokiej wartości parametru zgadywania (0,2) nie powinno specjalnie dziwić.

Warto przy tym zwrócić uwagę, że dyskryminacja jest w stosunku do wartości z modelu 3PL ściągnięta w kierunku zera, a trudność zadań wyraźnie zaniżona. Jednocześnie jednak, przynajmniej w przypadku łatwiejszego z dwóch rozpatrywanych zadań, krzywa charakterystyczna z modelu 2PL dobrze przybliża przebieg prawdziwej krzywej charakterystycznej w bardzo szerokim zakresie wartości mierzonej cechy. Oczywiście po lewej stronie wykresu pojawiają się duże rozbieżności, jednak występują one poniżej wartości cechy równej -2 odchylenia standardowe, a więc dotyczą one tylko niewielkiego odsetka badanej populacji (~2,5%).

Trudniejsze do odwzorowania modelem 2PL okazuje się zadanie trudne. Ze względu na konieczność zadbania o lepsze odwzorowanie przebiegu w dolnej części krzywej, wyraźne różnice pojawiły się tu również w zakresie wysokiego natężenia cechy. Niemniej nawet tutaj w dosyć szerokim zakresie wartości około ±1 odchylenia standardowego od średniej (co obejmuje blisko 70% badanej populacji) różnice w przebiegu krzywych są bardzo niewielkie i nie przekraczają 3 punktów procentowych.



Rysunek 1. Na górnych panelach: krzywe charakterystyczne dwóch zadań 3PL i najlepiej przybliżające je (wg kryterium minimalizacji kwadratów różnic w populacji) krzywe charakterystyczne zadań 2PL (zakreskowane pole wyznacza różnice w przebiegu krzywych). Na dolnych panelach: funkcje informacyjne odpowiadające funkcjom charakterystycznym z górnych paneli

W dalszej kolejności warto zadać pytanie, jaki wpływ mogą mieć takie różnice na oszacowania wartości mierzonej cechy wyliczane na podstawie modelu i ich błędy standardowe. Wiadomo, że w modelu 2PL, w sytuacji gdy wszyscy zdający rozwiązują ten sam zestaw zadań, suma wartości parametrów dyskryminacji zadań, które zostały rozwiązane poprawnie, wyznacza uporządkowanie osób ze względu na natężenie mierzonej cechy (Birnbaum, 1968). Można więc powiedzieć, że waga, jaka przywiązywana jest do zadania przy wyznaczaniu przewidywanej wartości mierzonej cechy, powiązana jest z jego jakością pomiarową, w modelu 2PL utożsamianą właśnie z wartością parametru dyskryminacji[1]. W przypadku modelu 3PL nie da się już przedstawić podobnie prostego wzoru, na podstawie którego możliwe byłoby wyznaczenie uporządkowania badanych, jednak i w tym przypadku można powiedzieć, że zadanie daje tym więcej informacji o wartości mierzonej cechy, im ma ono lepsze własności pomiarowe. Te zaś związane są z wysoką wartością parametru dyskryminacji i niską wartością parametru zgadywania (Birnbaum, 1968).

W związku z tym niższą dyskryminację z modeli 2PL w przykładzie przytoczonym na rysunku 1. możemy interpretować jako rodzaj korekty, przekazujący informacje o słabych własnościach pomiarowych zadania w sytuacji, gdy części tej informacji nie może przekazać parametr zgadywania, nieobecny w modelu 2PL. Jednocześnie różnice w wartościach parametrów trudności nie będą mieć żadnego wpływu na oszacowania wartości mierzonej cechy, jako że parametry trudności nie odgrywają w tym żadnej roli. Pozwala to wysunąć przypuszczenie, że oszacowania mierzonej cechy uzyskane z modelu 2PL mogą być dosyć zbliżone do tych uzyskiwanych z modelu 3PL nawet w sytuacji, gdy model 2PL ewidentnie nie opisuje dobrze sposobu funkcjonowania niektórych zadań zawartych w teście (nie uwzględnia zgadywania).

Jeśli chodzi o spodziewane różnice w odniesieniu do błędów standardowych oszacowań mierzonej cechy, można odwołać się do krzywych informacyjnych zadań, które wyrysowano na dolnych panelach rysunku 1. Im wyższa wartość krzywej informacyjnej, tym mniejszy będzie błąd standardowy oszacowania dla osób o danej wartości cechy (Kondratek i Pokropek, 2013). W tym przypadku różnice pomiędzy dwoma modelami okazują się bardzo wyraźne. Model 2PL ma wyraźną tendencję do niedoszacowywania błędów standardowych w zakresie niskich wartości mierzonej cechy (tj. w zakresie, w którym prawdopodobieństwo poprawnej odpowiedzi zależy głównie od wartości parametru zgadywania), jak również, choć tu różnice są niewielkie, w zakresie bardzo wysokich wartości cechy. Z kolei w zakresie średnich i względnie wysokich wartości cechy można spodziewać się, że błędy standardowe szacowane z modelu 2PL będą zawyżone.

Opis badania symulacyjnego

W celu bardziej rygorystycznego zweryfikowaniu zależności będących przedmiotem naszego zainteresowania przeprowadzone zostało symulacyjne badanie własności modeli. Głównym przedmiotem zainteresowania była w nim siła związku liniowego pomiędzy oszacowaniami mierzonej cechy uzyskanymi

[1] Warto zwrócić uwagę, że jak długo wszyscy zdający rozwiązują te same zadania, przy wyznaczaniu przewidywanej wartości mierzonej cechy w ogóle nie bierze się pod uwagę trudności zadań.

z modeli 2PL i 3PL a prawdziwymi wartościami badanej cechy[2] oraz pomiędzy oszacowaniami z tych dwóch modeli. Analizie poddane zostały również związki pomiędzy wielkością błędów standardowych wyliczonych na podstawie modeli 2PL i 3PL oraz to, jak różnice pomiędzy błędami standardowymi z obu modeli powiązane są z wartością mierzonej cechy.

Analizie poddano zależność pomiędzy siłą opisanych powyżej związków, a wybranymi własnościami testów. Uwzględnione zostały:

1. Długość testu. Analizowano dwa warianty: (1) test składający się z 10 zadań i (2) test składający się z 20 zadań.
2. Liczbę zadań, w których występowało zgadywanie. Dla każdego z wariantów długości testu uwzględniono wszystkie możliwe sytuacje, tj. od 0 (zgadywanie nie wystąpiło w żadnym zadaniu) do liczby zadań w teście (zgadywanie występuje we wszystkich zadaniach).
3. Poziom zgadywania. Analizowano dwa warianty: (1) zgadywanie równe 0,15, (2) zgadywanie równe 0,25.
4. Liczbę obserwacji („rozwiązujących" test), na podstawie których prowadzona jest estymacja. Analizowano dwa warianty: (1) 500 obserwacji, (2) 5000 obserwacji.

Spośród kilku możliwych metod wyliczania oszacowań wartości mierzonej cechy na podstawie wyestymowanych wcześniej parametrów modelu (zob. Kondratek i Pokropek, 2013) wybrana została metoda Expected A'Posteriori (EAP), jako obecnie najbardziej rozpowszechniona. Można się przy tym spodziewać, że inne metody wyliczania oszacowań charakteryzują się, jeśli chodzi o analizowane tu zależności, bardzo podobnymi własnościami. Do estymacji parametrów zadań wykorzystano metodę Marginal Maximum Likelihood (MML). Wszystkie obliczenia przeprowadzone zostały w programie R w wersji 3.1.0, z wykorzystaniem biblioteki mirt w wersji 1.3.

Podstawowym założeniem symulacji było zastosowanie do generowania danych mechanizmów, w których występował czynnik zgadywania, a potem poddanie tak wygenerowanych danych analizie przy użyciu modelu uwzględniającego parametr zgadywania (model 3PL) oraz modelu, który zakładał niewystępowanie zgadywania (model 2PL) i porównanie uzyskanych wyników. Badanie symulacyjne przebiegało w ten sposób, że w każdej iteracji:

1. Najpierw wybierana była liczba obserwacji, którym przypisywane były wartości mierzonej cechy, wylosowane z rozkładu normalnego standaryzowanego.
2. Następnie wybierana była liczba zadań, w których występowało zgadywanie i poziom zgadywania.
3. Wartości parametrów dyskryminacji i trudności losowane były niezależnie od siebie. Trudność z rozkładu normalnego standaryzowanego, dyskryminacja z rozkładu logarytmiczno-normalnego o wartości oczekiwanej 1.3 i odchyleniu standardowym 0,23[3].

[2] Badanie symulacyjne, w którym prowadzący je samodzielnie generują dane, pozwala na badanie takich zależności, które nie mogą być analizowane na podstawie wyników rzeczywistych testów, z racji niemożliwości przypisania zdającym prawdziwych wartości badanej cechy.

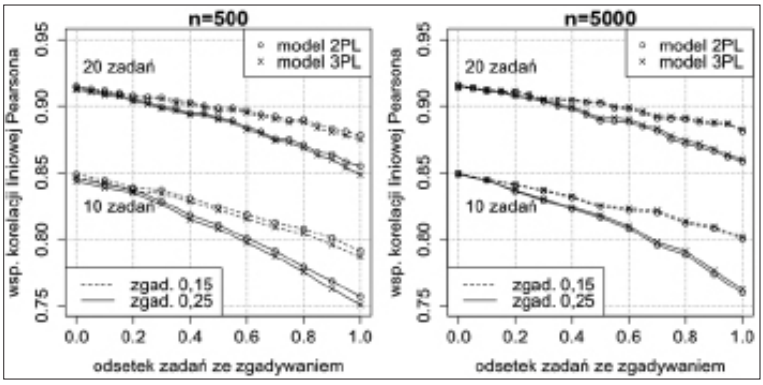
[3] Rozkład normalny powstający przez jego zlogarytmizowanie ma wartość oczekiwaną ~0,26 i odchylenie standardowe ~0,17.

4. Niezależnie od wartości parametrów dyskryminacji i trudności losowano, w których zadaniach występować będzie zgadywanie.
5. Kolejnym krokiem było „rozwiązywanie" testu, w czasie którego poszczególnym obserwacjom przypisywana była punktacja za poszczególne zadania, wyliczana zgodnie z założeniami modelu 3PL i wartościami parametrów zadań ustalonymi w poprzednich krokach.
6. Na podstawie uzyskanej punktacji estymowany był model 2PL oraz model 3PL, a na ich podstawie oszacowania EAP wartości mierzonej cechy dla poszczególnych obserwacji.
7. Na zakończenie każdej iteracji wyliczane i zapisywane były wartości korelacji opisującej analizowane zależności.

Symulację powtarzano tak długo, aby dla każdej kombinacji długości testu i poziomu zgadywania przypadało co najmniej 1500 iteracji dla liczby obserwacji równej 5000 i co najmniej 2000 iteracji dla liczby obserwacji równej 500. Analizie poddane zostały wyniki z łącznie 14 970 iteracji.

Wyniki badania symulacyjnego

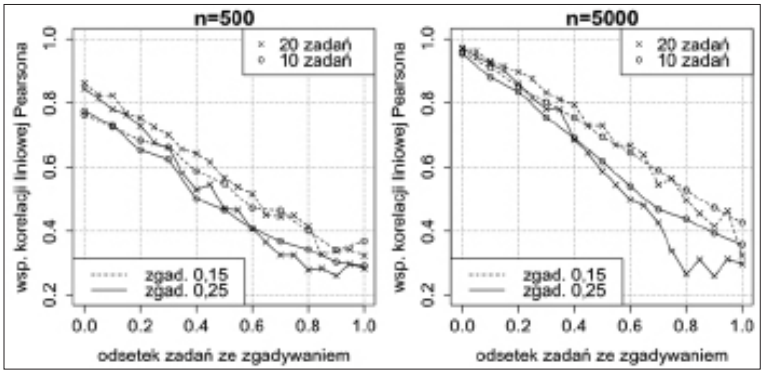
Na rysunku 2. pokazane zostały średnie wartości współczynnika korelacji liniowej Pearsona pomiędzy prawdziwymi wartościami mierzonej cechy a jej oszacowaniami z modeli 2PL i 3PL. Już na początku można zauważyć, że rodzaj użytego modelu pozostaje niemal bez wpływu na wyniki i to nawet w sytuacji, gdy zgadywanie występuje we wszystkich zadania w teście i to na wysokim poziomie (0,25). Za ciekawe można uznać, że w sytuacji, gdy dostępnych jest relatywnie niewiele jednostek obserwacji (500), oszacowania z modeli 2PL okazują się wręcz minimalnie, lepiej odwzorowywać prawdziwe wartości badanej cechy, jednak sytuacja ulega odwróceniu, gdy wnioskowanie odnosi się do większej grupy (5000 obserwacji). Trzeba jednak podkreślić, że różnice te są bardzo niewielkie i nie mają właściwie żadnego praktycznego znaczenia.



Rysunek 2. Średnie wartości współczynników korelacji liniowej pomiędzy prawdziwymi wartościami cechy ukrytej a oszacowaniami z modeli IRT 2PL i 3PL w zależności od liczby obserwacji, liczby zadań, z których składał się test i natężenia zgadywania

Tak niewielkie różnice wynikają z bardzo dużego podobieństwa oszacowań z obu modeli. W ramach blisko 15 tysięcy przeanalizowanych iteracji najniższa odnotowana wartość współczynnika korelacji liniowej pomiędzy oszacowaniami z modelu 2PL i z modelu 3PL wyniosła 0,896, ale nawet w najbardziej niekorzystnym spośród rozpatrywanych scenariuszy (krótki test, zgadywanie na poziomie 0,25 we wszystkich zadaniach) średnia wartość tej korelacji wyniosła 0,986, a w bardziej sprzyjających warunkach była tylko większa (z racji niewielkiej zmienności wyników tych nie prezentujemy na wykresie).

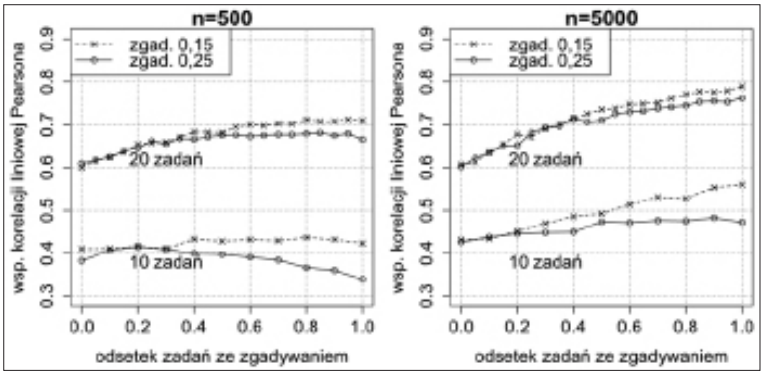
Wracając do czynników wpływających na jakość odwzorowania prawdziwych wartości mierzonej cechy w oszacowaniach wyliczanych na podstawie modelu, można stwierdzić, że występowanie zgadywania oddziałuje tu wyraźnie negatywnie. Im większe natężenie zgadywania, zarówno w kategoriach liczby zadań, w których występuje, jak też wartości zgadywania, tym mniejsza siła związku pomiędzy wartościami przewidywanymi a prawdziwymi. Jak już wspomniano, fakt uwzględnienia zgadywania w modelu używanym do estymacji nie przynosi tu właściwie żadnej poprawy w stosunku do modelu 2PL, nawet gdy ten drugi jest w sposób oczywisty nieadekwatny. Bardzo silny związek liczby zadań w teście i jego rzetelności (siły związku oszacowań i wartości prawdziwych) nie powinien dziwić. Warto przy tym odnotować, że przy dłuższym teście różnice w natężeniu zgadywania mają nieco mniejszy wpływ na uzyskiwane wyniki.



Rysunek 3. Średnie wartości współczynników korelacji liniowej pomiędzy wyestymowanymi błędami standardowymi oszacowań cechy ukrytej dla poszczególnych obserwacji z modelu IRT 2PL i z modelu IRT 3PL w zależności od liczby obserwacji, liczby zadań, z których składał się test i natężenia zgadywania

Jeśli chodzi o porównanie wartości błędów standardowych oszacowań mierzonej cechy, pomiędzy rozpatrywanymi w naszym badaniu scenariuszami mamy do czynienia z bardzo dużymi różnicami w sile związku. Wyniki w tym zakresie podsumowuje wykres 3. O ile dla dużej liczby obserwacji i niewystępowania zgadywania błędy standardowe wyliczone na podstawie modelu 2PL pozostają bardzo silnie powiązane z błędami z modelu 3PL, o tyle siła związku bardzo szybko spada wraz ze wzrostem liczby zadań, w których występuje zgadywanie. Nieco mniejsze znaczenie wydaje się mieć poziom zgadywania, a długość testu pozostaje niemal bez wpływu na wyniki. W najbardziej niekorzystnych spośród rozpatrywanych scenariuszy, charakteryzujących się bardzo wysokim

natężeniem zgadywania, zależność liniowa pomiędzy błędami standardowymi z modeli 2PL i 3PL jest już dosyć niewielka, ze średnimi wartościami współczynnika korelacji na poziomie 0,3-0,4. Jednocześnie warto odnotować, że średnie różnice pomiędzy błędami standardowymi z modeli 2PL i 3PL nie odbiegają systematycznie od zera.



Rysunek 4. Średnie wartości współczynników korelacji liniowej pomiędzy cechą ukrytą a różnicą pomiędzy wyestymowanymi błędami standardowymi oszacowań cechy ukrytej dla poszczególnych obserwacji z modelu IRT 2PL i z modelu IRT 3PL (tj. $cor[\hat{\theta}, bs(\hat{\theta}_{2PL}) - bs(\hat{\theta}_{3PL})]$) w zależności od liczby obserwacji, liczby zadań, z których składał się test i natężenia zgadywania

Na podstawie przeprowadzonego we wcześniejszej części artykułu porównania formalnych własności modeli 2PL i 3PL sformułowany został wniosek, że model 2PL powinien mieć tendencję do niedoszacowywania wielkości błędów standardowych w zakresie niskich wartości mierzonej cechy i tendencję do przeszacowywania tych błędów w zakresie średnich i umiarkowanie wysokich wartości tej cechy. Gdyby tak było, oznaczałoby to występowanie pozytywnej korelacji prawdziwych wartości mierzonej cechy i różnicy między błędem standardowym z modelu 2PL a błędem standardowym z modelu 3PL. Jak pokazuje wykres 4., przypuszczenia te znajdują częściowe potwierdzenie w wynikach symulacji. Kiedy testowaniu poddawana jest duża grupa osób, wyraźnie widać, że wraz ze wzrostem natężenia zgadywania wzrasta też siła korelacji, a więc różnice pomiędzy błędami standardowymi z modeli 2PL i 3PL mają tendencję do układania się zgodnie z naszymi przewidywaniami. Jednocześnie jednak zmniejszenie liczby zadań wyraźnie osłabia tę zależność. Z kolei w sytuacji, gdy testowaniu poddawana jest tylko niewielka grupa osób, nie obserwuje się wzrostu siły takiej zależności wraz ze wzrostem natężenia zgadywania.

Wnioski

W artykule podjęty został problem, na ile stosowanie modelu 2PL do skalowania wyników testów złożonych z zadań zamkniętych może prowadzić do uzyskania wiarygodnych oszacowań wartości badanej cechy w sytuacji, gdy wiadomo, że model ten jest nieadekwatny ze względu na obserwowane w danych przejawy występowania zgadywania. Co jest nieco zaskakujące, właściwie

we wszystkich rozważanych sytuacjach, nawet w sytuacji niezwykle wysokiego natężenia zgadywania w procesie generowania danych, jakość oszacowań punktowych uzyskiwanych z modeli 2PL nie ustępowała jakości oszacowań z modeli 3PL, uwzględniających występowanie zgadywania. W istocie oszacowania z obu modeli były do siebie bardzo zbliżone, a w większości rozpatrywanych sytuacji niemal ze sobą tożsame.

Oczywiście uzyskanych tu wyników nie należy przeceniać. Rozpatrywana kwestia jest bowiem jedynie wąskim wycinkiem zastosowań analizowanych modeli. W sytuacji, gdy model okazuje się nie pasować do danych, nie możemy ufać w szczególności np. oszacowanym parametrom zadań. Model 3PL (i inne bardziej złożone modele) bez wątpienia jest niezwykle użyteczny w kontekście diagnostyki własności psychometrycznych zarówno poszczególnych zadań, jak i całych arkuszy testowych.

Jednocześnie w sytuacji, gdy podstawowym albo wręcz jedynym celem skalowania jest uzyskanie punktowych oszacowań natężenia cechy wśród badanych osób w celu wykorzystania tych oszacowań w dalszych analizach, kwestia wyboru pomiędzy modelem 2PL a 3PL okazuje się mieć zupełnie marginalne znaczenie i to bez względu na ewentualne różnice w stopniu dopasowania modeli do danych. Użycie modelu 2PL może mieć przy tym pewne zalety praktyczne – zdecydowanie krótszy czas estymacji w przypadku dużych danych, a w przypadku bardzo ograniczonej liczby obserwacji większą stabilność wyników.

Należy przy tym podkreślić, że uzyskane wyniki odnoszą się do własności oszacowań punktowych. Występowanie znacznych różnic w oszacowaniach błędów standardowych estymatorów EAP pomiędzy modelami 2PL a 3PL (tym większych, im większe natężenie zgadywania w procesie generującym dane) sugeruje, że uzyskane przez nas wyniki nie mogą być generalizowane na analizy z użyciem tzw. Plausible Values (zob. Kondratek i Pokropek, 2013). Określenie własności estymatorów tego rodzaju wymagać będzie podjęcia dodatkowych badań w przyszłości.

Bibliografia

1. Birnbaum A. *Some latent trait models and their use in inferring an examinee's ability* [w:] Lord F. M., Novick M. R. (red.), „Statistical theories of mental test scores". Addison–Wesley, Reading, MA 1968.
2. Kondratek B., Pokropek A., *IRT i pomiar edukacyjny* [w:] „Edukacja" 2013, nr 4 (124).
3. Niemierko B., *Diagnostyka edukacyjna*, Wydawnictwa Naukowe PWN, Warszawa 2009.
4. Twardowska A., Grajkowski W., Chrzanowski M., Ostrowska B., Spalik K., *Dlaczego warto zamykać zadania?* [w:] Niemierko B., Szmigel M.K. (red.), *Ewaluacja w edukacji: koncepcje, metody, perspektywy*, gRUPA TOMAMI, Kraków 2011.