Anna Hawrot
Instytut Badań Edukacyjnych
Zespół SUEK, Zespół EWD

Wybór ma znaczenie,
czyli problem miar dopasowania pozycji testowych do modelu pomiarowego w tworzeniu testów za pomocą metody IRT

Teoria odpowiedzi na pozycje testowe (*item response theory*, IRT), choć jeszcze w Polsce niezbyt popularna i znacznie słabiej znana niż klasyczna teoria testu (KTT), zaczyna pojawiać się w pomiarze psychologicznym i edukacyjnym (np. test Gw-TWT, Dziedziewicz, Karwowski, 2012; test TUnSS, Kaczan, Rycielski, w tym tomie). Przyczyn niewielkiej popularności IRT w porównaniu z KTT upatrywać można m.in. w jej większej złożoności przy jednoczesnym braku wyczerpującej literatury w języku polskim, jak i konieczności stosowania specjalistycznego, niejednokrotnie kosztownego oprogramowania w celu analizy właściwości testów i składających się na nie zadań. Jej zalety są jednak niezaprzeczalne (por. np. de Ayala, 2009; Kondratek, 2007), stąd też nie jest zaskoczeniem wzrastające nią zainteresowanie. Na osoby korzystające z IRT podczas konstrukcji testów i badania właściwości narzędzi już powstałych czeka szereg decyzji, np. dotyczących metody estymacji parametrów, sposobu normalizacji, wykorzystywanych miar dopasowania całego modelu i poszczególnych zadań. Dla statystyków pracujących w ramach IRT podjęcie ich nie stanowi problemu, jednak ze względu na złożoność stosowanego aparatu matematycznego praktykom może nastręczać trudności.

Jak wyżej wspomniano, analizy w ramach IRT wymagają zastosowania specjalistycznego oprogramowania. Obecne na rynku programy różnią się pod wieloma kluczowymi dla analiz względami, m.in. oferują różne miary dopasowania pozycji, często nie dostarczając jednocześnie ich szerokiego wachlarza. Dotyczy to zwłaszcza programów bezpłatnych, zazwyczaj uboższych niż programy wydawane komercyjnie. Stawia to często badacza w sytuacji bez wyboru, w której skorzystać może tylko z oprogramowania bezpłatnego, o ograniczonej liczbie opcji, w tym o ograniczonej liczbie miar dopasowania pozycji. Tymczasem miary te dostarczają istotnych informacji na temat jakości zadań, niejednokrotnie kluczowych dla dalszych losów pozycji testowych.

Z tego punktu widzenia istotnym więc wydaje się pytanie o to, czy ta sama analiza przeprowadzona w dwóch różnych programach, oferujących różne miary dopasowania pozycji, będzie prowadzić do tych samych wniosków. Czy może to, jakie oprogramowanie (a tym samym, jakie miary dopasowania) zastosujemy, nie ma większego znaczenia? Intuicja podpowiada, iż należy spodziewać się jakichś różnic w wynikach (wszak różne miary różnią się zapewne sposobem ich obliczania), lecz wnioski wyciągnięte na ich podstawie nie powinny się zasadniczo różnić - dotyczą przecież tego samego aspektu. Spróbujmy zweryfikować to przypuszczenie.

By umożliwić czytelnikowi śledzenie przedstawionych analiz, omówienie empirycznego przykładu poprzedzi krótkie wprowadzenie do problematyki testowania dopasowania modeli IRT. Na bieżąco objaśniane także będą kluczowe elementy analiz.

Miary dobroci dopasowania pozycji w modelach IRT

Ocena dobroci dopasowania modelu IRT służy weryfikacji, czy model, mający opisywać pewien fragment rzeczywistości, faktycznie tę rzeczywistość opisuje. Obejmuje ona dwa kroki - weryfikację założeń leżących u podstaw modelu (np. jednowymiarowości) oraz ocenę zbieżności predykcji modelu z danymi empirycznymi (Swaminathan, Hambelton, Rogers, 2007). Miary dobroci dopasowania pozycji, obok miar dobroci dopasowania całego modelu oraz miar dobroci dopasowania na poziomie obserwacji, stanowią element drugiego z wymienionych kroków.

Opracowano wiele miar dobroci dopasowania pozycji w wariantach obejmujących różne modele IRT i różne warianty szacowania parametrów. Przedstawienie ich wszystkich wykracza poza objętość i cele niniejszej pracy, poniżej jednak scharakteryzowano ich logikę.

Podstawowym i dość intuicyjnym sposobem oceny dopasowania pozycji w modelach IRT jest analiza różnic między wartościami oczekiwanymi na podstawie modelu i obserwowanymi w danych. W przypadku pozycji o dychotomicznym formacie odpowiedzi (0-1) na podstawie modelu oblicza się prawdopodobieństwo udzielenia poprawnej odpowiedzi przez osobę o określonym poziomie mierzonej cechy ukrytej (*theta)*. Owo prawdopodobieństwo to nic innego jak proporcja osób o określonym poziomie cechy, które udzieliły prawidłowej odpowiedzi w danej pozycji. Proporcje te porównuje się z obserwowanymi proporcjami poprawnych odpowiedzi i na tej podstawie oblicza reszty będące podstawą miar rozbieżności między modelem a danymi. Niejednokrotnie jednak liczba osób o danym poziomie cechy jest zbyt mała, dlatego dzieli się próbę na kilka podgrup i dla nich oblicza prawdopodobieństwa. Obliczone w ten sposób różnice służyć mogą do wyznaczenia różnego rodzaju miar, np. Infit, Outfit, jak również testów weryfikujących istotność statystyczną występujących różnic, np. χ^2, Q_1, LM (Swaminathan i in., 2007). Podkreślić należy, że metody te nie są równoważne - mają odmienne właściwości, są w niejednakowym stopniu wrażliwe na różne przyczyny niedopasowania zadań, w tych samych warunkach wykazują inną częstotliwość występowania błędu I rodzaju (por. DeMars, 2010). W wielu przypadkach pojawiają się też problemy ze zbyt dużą mocą testów statystycznych w przypadku, gdy próby są bardzo liczne.

Oprócz miar liczbowych, możliwe jest także wizualne porównywanie wykresów krzywych empirycznych i teoretycznych poszczególnych zadań.

Przykład empiryczny

Wykorzystano wyniki Testu Matryc Ravena - wersja Standard, forma Klasyczna (TMS-K; Jaworowska, Szustrowa, 2007) zebrane w toku badań podłużnych realizowanych przez Pracownię Szkolnych Uwarunkowań Efektywności Kształcenia z Instytutu Badań Edukacyjnych[1]. Test składa się z 60 zadań zgrupowanych w 5 równolicznych serii (A-E). Każde zadanie ma postać matrycy z brakującym fragmentem, którą należy uzupełnić jednym spośród ośmiu prezentowanych wycinków. Za każdą prawidłową odpowiedź osoba badana otrzymuje 1 punkt, a wynik końcowy stanowi sumę punktów uzyskanych we wszystkich zadaniach. Test przeznaczony jest do diagnozy intelektu.

Próba liczy 5413 uczniów klas III szkoły podstawowej (2699 dziewcząt, 2714 chłopców) ze 177 losowo wybranych szkół w całej Polsce. Dane zebrano między styczniem a kwietniem 2011 roku. Analizy nie obejmują obserwacji liczących 5 lub więcej braków danych (Jaworowska, Szustrowa, 2007).

Analizy przeprowadzono z wykorzystaniem oprogramowania MIRT Package (Glas, 2010) oraz Acer Conquest 2.0 (Wu, Adams, Wilson i Haldane, 2007). Podają one różne miary dopasowania zadań - odpowiednio test LM (Lagrange Multiplier test, np. Glas, 1988, 2007) oraz Infit, Outfit (np. de Ayala, 2009).

W obu wyżej wymienionych programach przeprowadzono analizę w jednowymiarowym modelu z jednym parametrem (1PL), z szacowaniem parametrów metodą MML (*Marginal Maximum Likelihood estimation,* patrz: np. de Ayala, 2009) i fiksacją średniego poziomu cechy ukrytej na wartości 0. Parametry trudności zadań wraz z ich błędami standardowymi (oszacowane jednakowo przez oba programy z dokładnością do trzeciego miejsca po przecinku) przedstawiono w tabeli 1.

Tabela 1. Trudność zadań oszacowana w modelu 1PL

ZADANIE	Trudność	Bł. st.	ZADANIE	Trudność	Bł. st.
A1	-4,68	0,16	C7	-0,73	0,04
A2	-4,64	0,16	C8	0,11	0,04
A3	-6,18	0,24	C9	-0,60	0,04
A4	-5,53	0,17	C10	1,29	0,04
A5	-5,06	0,13	C11	1,85	0,04
A6	-5,24	0,15	C12	3,41	0,07
A7	-2,90	0,06	D1	-3,93	0,09
A8	-1,98	0,05	D2	-1,51	0,05
A9	-3,46	0,07	D3	-1,19	0,05
A10	-2,31	0,05	D4	-1,15	0,04
A11	-0,30	0,04	D5	-2,02	0,05
A12	0,46	0,04	D6	-0,82	0,04
B1	-5,84	0,19	D7	-0,26	0,04
B2	-4,32	0,10	D8	0,11	0,04
B3	-4,14	0,10	D9	0,71	0,04

[1] Projekt badawczy „Badanie jakości i efektywności edukacji oraz instytucjonalizacja zaplecza badawczego" współfinansowany z funduszy Unii Europejskiej.

ZADANIE	Trudność	Bł. st.	ZADANIE	Trudność	Bł. st.
B4	-3,23	0,07	D10	0,32	0,04
B5	-2,69	0,06	D11	2,10	0,05
B6	-1,26	0,04	D12	3,90	0,08
B7	-0,51	0,04	E1	0,16	0,04
B8	-0,20	0,04	E2	0,88	0,04
B9	-0,57	0,04	E3	0,87	0,04
B10	-1,05	0,04	E4	1,35	0,04
B11	-0,29	0,04	E5	1,31	0,04
B12	1,08	0,04	E6	1,73	0,04
C1	-3,26	0,06	E7	1,86	0,04
C2	-3,03	0,06	E8	2,06	0,05
C3	-1,46	0,04	E9	3,55	0,07
C4	-0,71	0,04	E10	3,15	0,06
C5	-1,42	0,04	E11	3,61	0,07
C6	-0,21	0,04	E12	3,34	0,07

Jak wyżej wspomniano, zastosowane programy podają różne miary dopasowania pozycji testowych. Acer ConQuest 2.0 podaje statystyki Infit i Outfit, natomiast MIRT - test LM. W praktyce obie miary weryfikują dopasowanie empirycznych i teoretycznych krzywych charakterystycznych pozycji. Test LM jest wrażliwy na sytuacje rozbieżności między faktyczną dyskryminacją pozycji a dyskryminacją w modelu. Jego obliczenie wymaga podziału próby na podgrupy o zbliżonych liczebnościach na podstawie wyniku surowego w teście (MIRT wyodrębnia 3 takie podgrupy). W każdej z podgrup obliczone zostają obserwowane oraz oczekiwane na podstawie modelu średnie wyniki danej pozycji, średnie bezwzględnych wartości różnic między nimi oraz wartości statystyki LM dla każdej pozycji wraz z ich istotnością statystyczną. Średnie bezwzględnych wartości różnic pozwalają wnioskować o skali rozbieżności (im jej wartość jest bliższa zeru, tym lepiej), natomiast wzorzec różnic obserwowanych w podpróbach - o kierunku odchyleń (np. zbyt niskie wartości w podgrupie pierwszej i zbyt wysokie w trzeciej sugerują, iż pozycję cechuje dyskryminacja wyższa niż dyskryminacja w modelu; Glas, 1988, 2007).

Miary Outfit i Infit opierają się na kwadratach różnic między wartościami oczekiwanymi i obserwowanymi, przy czym Infit jest wartością ważoną. W efekcie miary te są wrażliwe na różne odstępstwa od modelu. Infit jest wrażliwy na odstępstwa od przewidywanych odpowiedzi osób o poziomie cechy zbliżonej do trudności zadania, natomiast Outfit - na odstępstwa od przewidywanych odpowiedzi osób o poziomie cechy oddalonej od trudności zadania (de Ayala, 2007). Podobnie jak test LM mogą one wskazywać na różnice w faktycznej dyskryminacji pozycji w porównaniu do dyskryminacji w modelu (DeMars, 2010). Wartość idealna obu statystyk, wskazująca na idealne dopasowanie, wynosi 1, przy czym wartości znacznie powyżej 1 wskazują na występowanie w danych większej zmienności niż przewidywana przez model, natomiast wartości poniżej 1 - przeciwnie, wskazują na zmienność w danych mniejszą niż przewiduje model. Niedopasowanie to może być widoczne na wykresach krzywych charakterystycznych pozycji pod postacią odpowiednio zbyt małego lub zbyt dużego kąta nachylenia krzywych empirycznych. Jednocześnie przyjmuje

się, iż pozycje, dla których wartości tych miar nie mieszczą się w przedziale od 0,5 do 1,5, mają niewielką wartość pomiarową (de Alaya, 2007). Zalecane są jednak różne „widełki" w zależności od celu pomiaru np. dla testów wysokiej stawki wymienia się przedział 0,8-1,2, dla testów, których wynik nie decyduje o losach osób je wypełniających: 0,7-1,3, natomiast dla narzędzi badawczych, w których wykorzystywane są skale ocen: 0,6-1,4 (por. np. Bond, Fox, 2001). Oprócz wspomnianych miar, ACER ConQuest oblicza także przedziały ufności pozwalające weryfikować istotność statystyczną różnic między modelem a danymi (nie wszystkie programy je podają). W sytuacji, gdy przedział pokrywa wartości miary, należy przyjąć hipotezę zerową o braku istotnych różnic między modelem a danymi (Wu i in., 2007).

Przyjrzyjmy się zatem miarom dopasowania pozycji raportowanym przez oba programy (dane zawarto w tabeli 2.). Jeśli chodzi o MIRT, test LM okazał się istotny statystycznie w przypadku 49 pozycji. Ze względu na jego zbyt dużą moc w razie prób bardzo licznych skoncentrować należy się na analizie średnich różnic między wartościami oczekiwanymi a obserwowanymi w wyodrębnionych podgrupach, które pozawalają na weryfikację skali niedopasowania (Glas, 2007). Nie przekraczają one wartości 0,07. Wskazuje to na dobre dopasowanie na poziomie zadań (np. Glas, 1988; van Schoor i in., 2006). Jeżeli przyjmiemy bardziej restrykcyjne kryterium oceny dobroci dopasowania, uznając za punkt graniczny wartość różnic powyżej 0,05, niedopasowanie pojawia się w przypadku 6 pozycji: A8, B9, B10, B11, D4 oraz D7. Jednak w przypadku 5 z nich wartość ta przekracza postawioną granicę o 0,01, natomiast jednej – o 0,02 (B9). Skala różnic nie jest więc duża. Krzywe charakterystyczne (ze względu na ograniczoną objętość nie zamieszczono ich w niniejszej pracy), także nie budzą zastrzeżeń. Jednak nadmienić należy, że krzywe empiryczne wyrysowano na podstawie trzech podgrup, w związku z czym ich porównanie z krzywymi teoretycznymi nie daje zbyt bogatych informacji. Mimo to wartości średnich różnic skłaniają do sformułowania wniosku o dobrym dopasowaniu 54 pozycji, natomiast pozostałych 6 na poziomie granicznym. Podsumowanie informacji na temat wykrytego niedopasowania zawarto w tabeli 3.

Tabela 2. Miary dopasowania pozycji raportowane przez programy MIRT i ACER ConQuest

zadanie	Outfit	PU	Infit	PU	LM	grupa 1		grupa 2		grupa 3		średnia różnic	grupa 1 n	grupa 2 n	grupa 3 n
						obs.	ocz.	obs.	ocz.	obs.	ocz.				
A1	1,84	(0,96, 1,04)	1,07	(0,81, 1,19)	27,06*	0,97	0,96	0,98	0,99	0,99	1	0,01	1678	1865	1869
A2	1,67	(0,96, 1,04)	1,06	(0,82, 1,18)	15,9*	0,97	0,96	0,98	0,99	0,99	1	0,01	1679	1864	1869
A3	0,65	(0,96, 1,04)	1,03	(0,59, 1,41)	0,83	0,99	0,99	1	1	1	1	0	1682	1862	1868
A4	1,41	(0,96, 1,04)	1,04	(0,71, 1,29)	9,31*	0,99	0,98	0,99	1	1	1	0	1679	1865	1868
A5	1,67	(0,96, 1,04)	1,04	(0,77, 1,23)	9,73*	0,98	0,97	0,99	0,99	0,99	1	0	1679	1865	1868
A6	1,19	(0,96, 1,04)	1,03	(0,75, 1,25)	1,11	0,98	0,97	0,99	0,99	1	1	0	1681	1862	1869
A7	0,98	(0,96, 1,04)	1	(0,93, 1,07)	4,54	0,82	0,82	0,94	0,95	0,99	0,98	0	1865	1673	1874
A8	1,5	(0,96, 1,04)	1,19	(0,95, 1,05)	144,79*	0,75	0,67	0,84	0,88	0,91	0,95	0,06	1850	1896	1666
A9	0,83	(0,96, 1,04)	0,97	(0,90, 1,10)	5,05	0,87	0,88	0,97	0,97	0,99	0,99	0	1671	1870	1871
A10	1,09	(0,96, 1,04)	1,06	(0,94, 1,06)	10,39*	0,75	0,73	0,89	0,91	0,96	0,96	0,01	1853	1675	1884

zadanie	Outfit	PU	Infit	PU	LM	grupa 1		grupa 2		grupa 3		średnia różnic	grupa 1 n	grupa 2 n	grupa 3 n
						obs.	ocz.	obs.	ocz.	obs.	ocz.				
A11	1,06	(0,96, 1,04)	1,04	(0,98, 1,02)	14,01*	0,34	0,3	0,57	0,59	0,79	0,8	0,02	1770	1933	1709
A12	1,14	(0,96, 1,04)	1,08	(0,98, 1,02)	42,98*	0,23	0,18	0,4	0,41	0,62	0,66	0,03	1742	1919	1751
B1	1,34	(0,96, 1,04)	1,03	(0,66, 1,34)	2,5	0,99	0,99	1	1	1	1	0	1680	1865	1867
B2	1,06	(0,96, 1,04)	1,01	(0,85, 1,15)	3,57	0,95	0,94	0,99	0,99	0,99	0,99	0	1677	1866	1869
B3	0,65	(0,96, 1,04)	0,93	(0,86, 1,14)	18,7*	0,92	0,93	0,99	0,98	1	0,99	0,01	1678	1865	1869
B4	0,88	(0,96, 1,04)	0,97	(0,91, 1,09)	5,01	0,84	0,85	0,96	0,96	0,99	0,99	0,01	1672	1870	1870
B5	0,86	(0,96, 1,04)	0,95	(0,93, 1,07)	19,78*	0,78	0,79	0,94	0,94	0,99	0,97	0,01	1868	1672	1872
B6	1,09	(0,96, 1,04)	1,06	(0,97, 1,03)	15,73*	0,55	0,51	0,76	0,78	0,89	0,9	0,02	1827	1691	1894
B7	1,1	(0,96, 1,04)	1,07	(0,97, 1,03)	35,44*	0,4	0,35	0,62	0,64	0,81	0,83	0,03	1792	1925	1695
B8	0,87	(0,96, 1,04)	0,9	(0,98, 1,02)	110,01*	0,21	0,28	0,57	0,55	0,83	0,77	0,05	1760	1723	1929
B9	0,8	(0,96, 1,04)	0,85	(0,97, 1,03)	194,59*	0,26	0,36	0,68	0,64	0,89	0,83	0,07	1768	1724	1920
B10	0,75	(0,96, 1,04)	0,84	(0,97, 1,03)	250,62*	0,38	0,47	0,77	0,74	0,95	0,89	0,06	1820	1702	1890
B11	0,82	(0,96, 1,04)	0,86	(0,98, 1,02)	187,7*	0,2	0,3	0,61	0,58	0,84	0,78	0,06	1761	1717	1934
B12	0,96	(0,96, 1,04)	0,99	(0,97, 1,03)	5,75	0,1	0,11	0,28	0,28	0,53	0,52	0,01	1918	1729	1765
C1	0,92	(0,96, 1,04)	1	(0,91, 1,09)	0,83	0,86	0,86	0,96	0,96	0,99	0,99	0	1671	1869	1872
C2	1,25	(0,96, 1,04)	1,06	(0,92, 1,08)	21,85*	0,85	0,83	0,94	0,95	0,97	0,98	0,02	1660	1876	1876
C3	1,14	(0,96, 1,04)	1,08	(0,96, 1,04)	53,66*	0,62	0,56	0,79	0,82	0,9	0,93	0,04	1844	1894	1674
C4	1,13	(0,96, 1,04)	1,07	(0,97, 1,03)	25,76*	0,43	0,39	0,66	0,67	0,82	0,85	0,03	1797	1701	1914
C5	0,87	(0,96, 1,04)	0,91	(0,96, 1,04)	30,07*	0,52	0,55	0,82	0,81	0,94	0,92	0,02	1841	1682	1889
C6	1,05	(0,96, 1,04)	1,04	(0,98, 1,02)	17,1*	0,32	0,29	0,57	0,57	0,76	0,78	0,02	1769	1936	1707
C7	0,85	(0,96, 1,04)	0,91	(0,97, 1,03)	74,86*	0,35	0,39	0,67	0,68	0,9	0,85	0,03	1793	1710	1909
C8	1	(0,96, 1,04)	1,02	(0,98, 1,02)	5,76	0,24	0,23	0,5	0,49	0,71	0,73	0,01	1765	1929	1718
C9	0,93	(0,96, 1,04)	0,95	(0,97, 1,03)	6,69*	0,35	0,37	0,67	0,66	0,86	0,84	0,01	1803	1911	1698
C10	1,14	(0,96, 1,04)	1,06	(0,97, 1,03)	43,37*	0,13	0,09	0,24	0,23	0,42	0,47	0,03	1723	1900	1788
C11	1,24	(0,96, 1,04)	1,04	(0,96, 1,04)	14,69*	0,07	0,05	0,15	0,15	0,32	0,34	0,01	1695	1892	1824
C12	4,55	(0,96, 1,04)	1,15	(0,90, 1,10)	297,48*	0,08	0,01	0,04	0,04	0,04	0,11	0,05	1695	1858	1858
D1	0,61	(0,96, 1,04)	0,93	(0,88, 1,12)	45,66*	0,9	0,92	0,99	0,98	1	0,99	0,01	1676	1868	1868
D2	0,85	(0,96, 1,04)	0,9	(0,96, 1,04)	75,4*	0,51	0,57	0,85	0,82	0,95	0,92	0,04	1834	1688	1890
D3	0,83	(0,96, 1,04)	0,9	(0,97, 1,03)	72,29*	0,45	0,5	0,78	0,77	0,93	0,9	0,03	1808	1707	1897
D4	0,76	(0,96, 1,04)	0,84	(0,97, 1,03)	233,72*	0,39	0,49	0,8	0,76	0,95	0,9	0,06	1824	1705	1883
D5	0,75	(0,96, 1,04)	0,85	(0,95, 1,05)	128,78*	0,62	0,68	0,93	0,88	0,97	0,95	0,04	1859	1676	1877
D6	0,85	(0,96, 1,04)	0,89	(0,97, 1,03)	95,29*	0,34	0,42	0,74	0,7	0,89	0,86	0,05	1805	1706	1901
D7	0,94	(0,96, 1,04)	0,96	(0,98, 1,02)	7,33*	0,28	0,3	0,57	0,57	0,79	0,78	0,01	1772	1701	1939
D8	0,93	(0,96, 1,04)	0,94	(0,98, 1,02)	33,87*	0,19	0,23	0,49	0,49	0,77	0,73	0,03	1746	1954	1712
D9	0,92	(0,96, 1,04)	0,94	(0,98, 1,02)	19,97*	0,12	0,14	0,35	0,35	0,63	0,6	0,02	1734	1932	1746
D10	0,89	(0,96, 1,04)	0,91	(0,98, 1,02)	69,27*	0,16	0,2	0,42	0,44	0,75	0,69	0,04	1735	1954	1723
D11	1,63	(0,96, 1,04)	1,14	(0,95, 1,05)	126,57*	0,09	0,04	0,15	0,12	0,21	0,29	0,05	1708	1869	1835
D12	1,89	(0,96, 1,04)	1,07	(0,87, 1,13)	43,9*	0,02	0,01	0,03	0,02	0,05	0,07	0,02	1693	1862	1854
E1	1,01	(0,96, 1,04)	1,02	(0,98, 1,02)	4,82	0,24	0,22	0,48	0,48	0,71	0,72	0,01	1758	1936	1718

zadanie	Outfit	PU	Infit	PU	LM	grupa 1		grupa 2		grupa 3		średnia różnic	grupa 1 n	grupa 2 n	grupa 3 n
						obs.	ocz.	obs.	ocz.	obs.	ocz.				
E2	1,01	(0,96, 1,04)	1,01	(0,97, 1,03)	8,38*	0,15	0,13	0,3	0,31	0,55	0,56	0,01	1733	1896	1783
E3	1,05	(0,96, 1,04)	1,03	(0,97, 1,03)	10,26*	0,15	0,13	0,31	0,32	0,55	0,56	0,01	1724	1908	1780
E4	0,9	(0,96, 1,04)	0,9	(0,97, 1,03)	73,09*	0,07	0,08	0,18	0,22	0,51	0,45	0,04	1705	1902	1805
E5	0,93	(0,96, 1,04)	0,93	(0,97, 1,03)	33,72*	0,07	0,09	0,2	0,23	0,51	0,46	0,03	1707	1902	1803
E6	1,02	(0,96, 1,04)	0,95	(0,96, 1,04)	24,21*	0,07	0,06	0,13	0,17	0,39	0,37	0,02	1701	1887	1824
E7	1,67	(0,96, 1,04)	1,14	(0,96, 1,04)	169,63*	0,13	0,05	0,16	0,15	0,25	0,34	0,06	1714	1872	1826
E8	1,32	(0,96, 1,04)	1,06	(0,95, 1,05)	58,95*	0,08	0,04	0,13	0,13	0,25	0,3	0,03	1699	1873	1840
E9	2,47	(0,96, 1,04)	1,05	(0,89, 1,11)	47,22*	0,03	0,01	0,03	0,03	0,07	0,09	0,02	1690	1864	1857
E10	2,33	(0,96, 1,04)	1,08	(0,91, 1,09)	79,21*	0,05	0,01	0,05	0,05	0,09	0,13	0,03	1688	1869	1854
E11	3,64	(0,96, 1,04)	1,12	(0,89, 1,11)	195,39*	0,05	0,01	0,04	0,03	0,04	0,09	0,03	1692	1862	1857
E12	3,74	(0,96, 1,04)	1,12	(0,90, 1,10)	169,27*	0,06	0,01	0,05	0,04	0,06	0,11	0,04	1689	1867	1853

Legenda: PU - przedział ufności; * - istotne na poziomie 0,05; LM - wartość statystyki testowej LM; obs. - wartości obserwowane; ocz. - wartości oczekiwane na podstawie modelu; średnia różnic - średnia bezwzględnych wartości różnic między wartościami oczekiwanymi i obserwowanymi w wyodrębnionych grupach; n - liczebność wyodrębnionych grup.

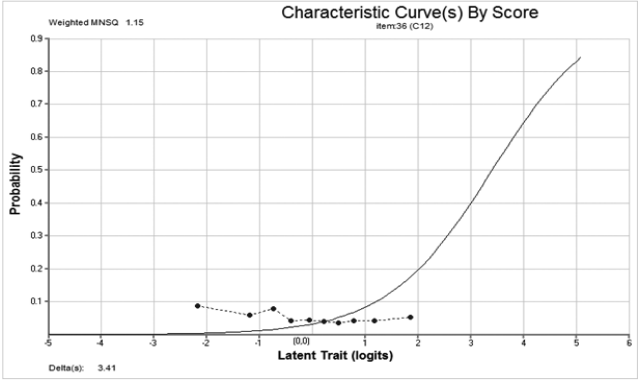
Do jakich wniosków na temat dopasowania prowadzi analiza na podstawie miar podawanych przez ACER ConQuest? Przegląd wartości Infit wydaje się wstępnie potwierdzać wnioski, które wysnulibyśmy na podstawie testu LM. Wprawdzie jedynie dla niespełna 1/2 pozycji (26) przedziały ufności pokrywają wartości Infit, jednak podobnie jak test LM, znaczenie odgrywa tu rozmiar próby. Ponownie konieczne jest więc odwołanie się do „reguł kciuka". Wartości Infit rozciągają się między 0,84 (B10, D4) i 1,19 (A8), dla 48 pozycji zamykając się w przedziale 0,9-1,1. Przyjmując, że poziom inteligencji jest uwarunkowany biologicznie, niezmienny w toku życia i w dużym stopniu decyduje o sukcesie życiowym, założyć można, iż wyniki testu inteligencji, będące istotną informacją np. w diagnozie indywidualnej, mogą być podstawą podejmowania decyzji co do losów jednostki. Należy zatem stawiać mu wysokie wymagania w zakresie właściwości psychometrycznych i odwoływać się do reguł kciuka zalecanych dla testów wysokiej stawki. Zgodnie z nimi za dobrze dopasowane zadania uznaje się te, których Outfit i Infit mieszczą się w przedziale wartości 0,8-1,2 (np. Bond, Fox, 2001). W tym świetle dopasowanie zadań mierzone za pomocą Intfit jest satysfakcjonujące.

Tabela 3. Pozycje wskazane jako niedopasowane przez miary Outfit, Infit i test LM

program	miara	liczba pozycji	pozycje wskazane	
			w obu programach	w jednym z programów
MIRT	LM	6	A8, B10, D4, E7	B9, B11
ACER ConQuest	Outfit	23		A1, A2, A3, A4, A5, B1, B3, C2, C11, C12, D1, D5, D11, D12, E8, E9, E10, E11, E12
	Infit	0	-	-

Sytuacja jednakże zmienia się, gdy przyjrzymy się miarom Outfit. Przyjmują one wartości od 0,61 (D1) do 4,55 (C12). Wartości w granicach 0,8-1,2 osiągnęło 37 pozycji, poniżej 0,8 - 6 pozycji (A3, B3, B10, D1, D4, D5), powyżej 1,2 - 17 pozycji, przy czym wartości powyżej 1,5 - aż 11 (A1, A2, A5, C12, D11, D12, E7, E9-E12). Okazuje się więc, że w przypadku 17 zadań model nie jest w stanie przewidzieć zmienności pojawiającej się w danych, natomiast w przypadku 6 - wzorzec odpowiedzi jest zbyt deterministyczny. Rzut oka na trudności pozycji (tabela 1; można skorzystać także z wyników dla podgrup podawanych przez MIRT - tabela 2.) pozwala stwierdzić, że problematyczne zadania to przede wszystkim zadania skrajnie łatwe (np. A1-A5) i skrajnie trudne (np. C12, D12, E8-E12), choć prawidłowość ta nie zawsze się powtarza (np. A6). W skrajnych sytuacjach empiryczna krzywa charakterystyczna okazuje się być zorientowana poziomo - ma to miejsce w przypadku trzech pozycji: C12, E11, E12, które charakteryzują się najwyższymi wartościami Outfit (odpowiednio: 4,55; 3,64; 3,74) przy akceptowalnych wartościach Infit (odpowiednio: 1,15; 1,12; 1,12). Przypomnijmy, że średnie bezwzględnych wartości różnic dla tych pozycji (raportowana przez MIRT) wyniosły odpowiednio 0,05; 0,04 i 0,03, wskazując na dobre dopasowanie. Wykres dla pozycji C12 (jako przykładowy) przedstawiono na rysunku 1. Krzywą teoretyczną zaznaczoną linią ciągłą, natomiast empiryczną - linią przerywaną.

Pozostałe „problematyczne" pozycje wykazują mniejsze rozbieżności między krzywą empiryczną a teoretyczną; nie zamieszczono ich ze względu na ograniczoną objętość pracy.



Rysunek 1. Empiryczna i teoretyczna krzywa charakterystyczna zadania C12

Podsumowując wyniki analiz z wykorzystaniem miar Outfit i Infit, na podstawie bardzo wysokich wartości miary Outfit należy wyłączyć z testu trzy pozycje (C12, E11, E12) ze względu na cechującą je zbyt dużą zmienność, nieprzewidywalność wzorca odpowiedzi osób o poziomie cechy oddalonej od trudności zadań. W przypadku 8 kolejnych (o wartościach Outfit powyżej 1,5) należy rozważyć podobne działanie. Podsumowanie informacji na temat wykrytego niedopasowania zawarto w tabeli 3. Nadmienić należy jednak, iż przedstawiona analiza dotyczy tylko jednego aspektu weryfikacji dopasowania modelu IRT - aspektu dopasowania pozycji.

Podsumowanie

Wnioski wyciągnięte na podstawie analizy z zastosowaniem Infit i Outfit stoją w sprzeczności z wnioskami wysnutymi na podstawie testu LM podawanego przez MIRT. Test LM wskazał na pewne problemy z dopasowaniem 6 zadań. Cztery z nich (A8, B10, D4, E7) przekroczyły jednocześnie akceptowalne granice miary Outfit. Jednakże test ten wykazał dobre dopasowanie 19 zadań, z którymi problemy sygnalizowała miara Outfit. Wśród zadań tych znalazło się 10 (z łącznie 11 pozycji) o wartościach Outfit powyżej 1,5, a więc takich, których wyłączenie z testu należy rozważyć. Pozostałe zadania wskazane przez test LM (B9 i B11), okazały się mieścić w granicach 0,8-1,2 zarówno dla Outfit, jak i Infit. Warto przypomnieć, że zadanie B9 wskazane zostało przez test LM jako najbardziej niedopasowane (wartość średniej różnic: 0,07).

Choć rezultaty te zasługują na odrębną, pogłębioną analizę przyczyn, wstępnie upatrywać ich można zarówno we właściwościach podawanych przez oprogramowanie miar, jak i kwestiach technicznych, takich jak np. liczba podgrup wyodrębnianych w ramach analiz z zastosowaniem testu LM (ta jednak nie podlega w MIRT modyfikacji użytkownika).

Wnioski

Przedstawiona powyżej analiza jednoznacznie falsyfikuje przedstawioną na początku pracy intuicję dotyczącą zbieżności wniosków wysnuwanych na podstawie różnych miar dopasowania pozycji. Okazuje się, iż sama dostępność oprogramowania, poprzez limitowanie dostępu do różnych miar dopasowania zadań, może być czynnikiem wpływającym na wnioski wysuwane przez badacza. Oczywiście, nie zawsze musi mieć to miejsce. Wynik ten uwidacznia jednak niebagatelne znaczenie uważnej i przemyślanej analizy danych raportowanych przez oprogramowanie. Nie może mieć ona „automatycznego" charakteru, zwykłego rzutu okiem na podane statystyki i weryfikacji, czy mieszczą się w cytowanych w literaturze „widełkach". Kluczowa jest znajomość właściwości stosowanych miar, gdyż są one wrażliwe na różne aspekty niedopasowania i wykazują różne właściwości (np. poziom błędów I rodzaju) w zależności np. od liczby pozycji w teście czy rozmiaru próby. Zastosowanie tej, która jest niewrażliwa na występujące niedopasowanie lub w naszych warunkach wykazuje np. podwyższenie częstości występowania błędu I rodzaju, może prowadzić do fałszywych wniosków. Jedynym lekarstwem na ten problem wydaje się być kompetencja badacza, jego wiedza na temat zalet i ograniczeń wykorzystywanych statystyk, która pomoże ocenić wiarygodność rezultatów, jak i obszar, którego dotyczą (tj. aspekt występowania bądź braku niedopasowania). Rozsądnym rozwiązaniem wydaje się być też stosowanie kilku miar jednocześnie, jednak nie zawsze jest to możliwe. Najlepsze nawet oprogramowanie nie zastąpi jednak zdrowego rozsądku badacza - wszak statystyki idealne nie istnieją, a ostateczna decyzja dotycząca dalszych losów pozycji należy zawsze do niego.

Bibliografia:

1. de Ayala, R.J. (2009). *Theory and practice of Item Response Theory*. New York, NY: Guilford Press.
2. Bond, T.G., Fox, Ch.M. (2001). *Applying the Rasch Model: Fundamental measurement in the human sciences.* Mahwah, NY: Lawrence Erlbaum.
3. DeMars, Ch. (2010). *Item response theory.* New York, NY: Oxford University Press.
4. Dziedziewicz, D., Karwowski, M. (2012). *Graficzno-werbalny Test Wyobraźni Twórczej (Gw-TWT): podręcznik tymczasowy*. Warszawa: APS.
5. Glas, C.A. (1988). *The derivation of some tests for the Rasch model from the multinominal distribution*. Psychometrika, Vol. 53 (4), 525-546.
6. Glas, C.A. (2007). *Testing generalized Rasch models* [in:] M.C. von Davier, C.H. Carstensen (Eds.). *Multivariate and mixture distribution Rasch models. Extensions and applications.* Springer.
7. Glas, C.A. (2010). *Preliminary manual of the software program Multidimentional Response Theory (MIRT)*. Pobrano z: http://www.utwente.nl/gw/omd/afdeling/temp_test/mirt-manual.pdf.
8. Jaworowska, A., Szustrowa, T. (2007). *Test Matryc Ravena w wersji Standard. Formy: Klasyczna, Równoległa, Plus. Polskie standaryzacje.* Warszawa: Pracownia Testów PTP.
9. Kaczan, R., Rycielski, P. (w tym tomie). *Diagnoza umiejętności dzieci 5-, 6- i 7-letnich za pomocą Testu Umiejętności na Starcie Szkolnym TUnSS*. W tym tomie.
10. Kondratek, B. (2007). *Teoria odpowiadania na pozycje testowe oraz klasyczna teoria testów. Porównanie w kontekście modelowania statystycznego sytuacji eksperymentalnej badania testem*. Biuletyn Badawczy Egzamin: Klasyczna i probabilistyczna teoria testu, Vol. *9*, 76-194.
11. Swaminathan, H., Hambelton, R.K., Rogers, H.J. (2007). *Assessing item fit of item response theory models* [w:] C.R. Rao, S. Sinharay (red.) *Handbook of Statistics Volume 26: Psychometrics.* Amsterdam: Elsevier.
12. van Schoor, N.M., Knol, D.L., Glas, C.A., Ostelo, R.W., Leplege, A., Cooper, C., Johnell, O., Lips, P. (2006). *Development of the Qualeffo-31, an osteoporosis-specific quality-of-life questionnaire. Osteoporosis International, Vol.17*, 543-551.
13. Wu, M.L., Adams, R.J., Wilson, M.R., Haldane, S.A. (2007). *ACER Conquest version 2.0: generalized item response modeling software.* Camberwell, VIC: ACER Press.