

Dorota Węziak

Szkoła Główna Handlowa w Warszawie

ZASTOSOWANIE SKALOWANIA RASCHA DO OCENY TRAFNOŚCI SKALI POMIAROWEJ

1. Wstęp

W badaniach społecznych do analizowania zjawisk złożonych i bezpośrednio nieobserwowalnych wykorzystuje się skale pomiarowe złożone, zbudowane najczęściej z kilku, kilkunastu lub kilkudziesięciu skal prostych określanych również mianem stwierdzeń [10, s. 20; 11, s. 48; 12, s. 89], twierdzeń [9, s. 477-480], wypowiedzi [3, s. 412], aspektów [10, s. 20], kwestii [10, s. 20] bądź pozycji skali [8, s. 128-129; 10, s. 20]. Są to na ogół krótkie zdania oznajmujące [8, s. 127], z którymi respondent może się zgadzać lub nie, a z których każda ma na celu objęcie innego aspektu badanego zjawiska. Jeśli poszczególne pozycje skali mają różną intensywność, to tworzona przez nie skala złożona ma charakter skali skumulowanych ocen. Oznacza to, że pomiar odbywa się poprzez zsumowanie odpowiedzi każdego z respondentów na pozycje skali, a uzyskana suma określa miejsce respondenta na konstruowanej podziałce [10, s. 20].

Ze względu na to, że badana właściwość empiryczna nie jest bezpośrednio obserwowalna, może być zmierzona tylko za pomocą obserwowalnych zmiennych wskaźnikowych. Bardzo ważne jest zatem sprawdzenie, czy rzeczywiście pozycje skali pomiarowej są powiązane z mierzoną kategorią ukrytą, co przekłada się na trafność narzędzia pomiarowego.

Stosowany w badaniach społecznych kwestionariusz ankietowy jest zwykle jednakowy dla wszystkich badanych, więc również skala, za pomocą której zmierzona zostanie badana właściwość empiryczna, będzie miała taką samą treść dla wszystkich respondentów. Jest to ważne z punktu widzenia standaryzacji wyników. Należałoby jednak zapytać, czy zawsze uzasadnione. Bardzo ważną kwestią jest sprawdzenie, czy rzeczywiście narzędzie zadziałało w identyczny sposób wobec

wszystkich respondentów, jako że różnice w rozumieniu poszczególnych pozycji skali mogą wpłynąć na trafność narzędzia pomiarowego.

Do weryfikacji, na ile koncepcja autora narzędzia przełożyła się na percepcję respondenta, zastosowano porządkowe skalowanie Rascha. Należy jednak pamiętać, że ta metoda może być stosowana tylko do skal skumulowanych ocen zbudowanych z pozycji o zróżnicowanej intensywności.

2. Skalowanie Rascha – podstawowe informacje o metodzie

Skalowanie Rascha to zespół wielu metod skalowania mających na celu zrozumienie zebranych danych poprzez identyfikację anomalii w nich występujących. Nie jest to technika modelowania danych. Wręcz przeciwnie, w tej metodzie sprawdza się dopasowanie danych do modelu optymalnego wynikającego z założeń metody. Gdy dane nie pasują do modelu, nie szuka się modelu „lepszego”, ale wnikliwie analizuje się zebrane dane w celu znalezienia źródła niedopasowania, wychodząc z założenia, że to nie model jest źródłem problemu, a dane. Jest to podejście zdecydowanie różniące się od powszechnie przyjmowanego, dlatego też stało się jedną z przyczyn ogromnej krytyki, z jaką spotkał się G. Rasch, ogłaszając swoje dokonania. U podstaw zrozumienia metody leży uzmysłwienie sobie tego, że to model służy do opisania zebranych danych, nie zaś odwrotnie, a występujące anomalie w dopasowaniu nie są wynikiem złej konstrukcji modelu, tylko pewnych zakłóceń powstałych w procesie pomiaru z konstrukcją narzędzia pomiarowego włącznie.

W skalowaniu Rascha (bez względu na typ) badana cecha ukryta – opisywana przez zespół obserwowanych zmiennych wskaźnikowych – ma charakter ciągły i jest przedstawiana na abstrakcyjnym kontinuum o granicach w $-\infty$ i $+\infty$. Wszystkie zmienne wskaźnikowe uzyskują oszacowania intensywności (D_i), a wszyscy badani respondenci oszacowania swoich pozycji na skali (B_n) odpowiadające poziomowi badanej właściwości empirycznej, który prezentują. Oba typy oszacowań B_n i D_i mierzone są logitem i leżą na jednej osi metrycznej odpowiadającej kontinuum badanej cechy ukrytej. Czy dane odpowiadają założeniom metody¹, a zatem czy pasują do modelu, oceniane jest za pomocą statystyk dopasowania.

Model w dychotomicznym skalowaniu Rascha (*dichotomous Rasch model*) przybiera postać:

$$\pi_{ni} = \frac{e^{\beta_n - \delta_i}}{1 + e^{\beta_n - \delta_i}}, \quad (1)$$

¹ Porównanie intensywności dwóch zmiennych wskaźnikowych nie zależy ani od tego, którzy konkretnie respondenci udzielili na nie odpowiedzi, ani od intensywności innych zmiennych wskaźnikowych wykorzystywanych w skali. Pozycja dwóch respondentów nie zależy od tego, na które konkretnie pytania udzielili odpowiedzi, a także od pozycji innych badanych respondentów.

gdzie: π_{ni} – prawdopodobieństwo wybrania przez n -tego respondenta kategorii odpowiedzi 1 (*tak*) dla i -tej pozycji skali,

β_n – ocena pozycji n -tego respondenta,

δ_i – ocena intensywności i -tej pozycji skali.

Dychotomiczne skalowanie Rascha wykorzystuje się do badania skal skumulowanych ocen składających się z baterii pozycji o odpowiedziach dychotomicznych typu tak/nie, zgadzam się/nie zgadzam się, występuje/nie występuje itp.

W przypadku baterii pozycji mierzonych na skalach porządkowych, przy założeniu, że ich skale odpowiedzi są identyczne, należy stosować skalowanie porządkowe (*rating scale model*). Model skalowania porządkowego definiuje się następująco:

$$\pi_{nix} = \frac{e^{\beta_n - (\delta_i + \tau_x)}}{1 + e^{\beta_n - (\delta_i + \tau_x)}}, \quad (2)$$

gdzie: π_{ni} – prawdopodobieństwo wybrania przez n -tego respondenta kategorii odpowiedzi x dla i -tej pozycji skali

β_n – ocena pozycji n -tego respondenta na skali,

δ_i – ocena intensywności i -tej pozycji skali,

τ_x – x -ty parametr struktury skali (x -ty próg).

Koncepcja oceny dopasowania danych do modelu opiera się na porównaniu dwóch macierzy: w przypadku skalowania dychotomicznego zero-jedynkowej macierzy empirycznej i obliczonej z wykorzystaniem wzoru (1) macierzy teoretycznej, w przypadku skalowania porządkowego – macierzy zaobserwowanych odpowiedzi i macierzy oczekiwanych odpowiedzi zdefiniowanych następująco:

$$E_{ni} = \sum_{x=0}^n x \cdot p_{nix}, \quad (3)$$

gdzie: E_{ni} – wartość oczekiwana odpowiedzi n -tego respondenta na i -tą pozycję skali,

x – kod kategorii odpowiedzi przy założeniu, że odpowiedzi kodowane są od 0,

p_{nix} – prawdopodobieństwo wybrania przez n -tego respondenta kategorii odpowiedzi x na i -tą pozycję skali zdefiniowane tak jak we wzorze (2).

Stosowane w skalowaniach Rascha statystyki dopasowania dzielą się na kilka grup. Korzystać można z mierników wystandaryzowanych i niewystandaryzowanych, ważonych i nieważonych, uwzględniających podział respondentów na jednorodne pod względem sumy uzyskanych punktów grupy i bez takiego podziału. Pomimo bogactwa statystyk, w literaturze toczy się spór, które z nich są lepsze. Obecnie wydaje się, że przewagę zyskały mierniki w formie wystandaryzowanej i obliczane dla wszystkich respondentów bez podziału na grupy, choć może to być skutek wykorzystywanego oprogramowania. W tej chwili w literaturze przeważają publikacje, w których do obliczeń wykorzystany został program

WINSTEP, korzystający właśnie z mierników wystandaryzowanych ogólnych. Jednak od niedawna funkcjonuje na rynku inny program do skalowania Rascha – RUMM2020. Ten program do oceny dopasowania wykorzystuje mierniki niewystandaryzowane, obliczane po wcześniejszym przydzieleniu respondentów do odpowiednio zdefiniowanych grup na zasadzie podobieństwa wyniku ogólnego. Od momentu pojawienia się tego programu liczba publikacji wykorzystujących mierniki niewystandaryzowane wzrasta. Rośnie również liczba opracowań porównujących zalety i wady poszczególnych miar. Zainteresowany czytelnik znajdzie najważniejsze miary wraz z wzorami w załączniku do niniejszego artykułu.

Badając stopień (nie-)dopasowania, sprawdzić można sposób funkcjonowania poszczególnych pozycji skali, sposób udzielania odpowiedzi przez poszczególnych respondentów, a także sposób działania poszczególnych kategorii odpowiedzi.

Uzyskawszy wartość wybranej statystyki dopasowania wskazującą na brak zgodności z modelem, należy najpierw spróbować zidentyfikować przyczynę takiego stanu rzeczy, a dopiero potem rozważać usunięcie. Wynika to z tego, że przyczyn braku dopasowania może być wiele. Na przykład pozycja skali może być wieloznaczna, może mierzyć inną cechę ukrytą. Może również mieć bardzo małą moc dyskryminacyjną lub wręcz przeciwnie – zbyt dużą, co również w przypadku skalowania Rascha zostanie wykryte jako brak dopasowania.

3. Identyfikacja pozycji skali wykazujących różnice w funkcjonowaniu wobec różnych grup respondentów

Pozycje skali o zróżnicowanym funkcjonowaniu wobec wszystkich badanych respondentów określane są jako obciążone pozycje skali (*item bias*) lub stwierdzenia obciążone lub stronnicze [5, s. 100]. W skalowaniu Rascha obciążenie/stronniczość pozycji skali najbardziej ogólnie definiuje się jako posiadanie przez nią dwóch statystycznie różnych oszacowań intensywności D_i w zależności od podgrupy respondentów, dla której dokonuje się szacunku [15, s. 393]. Skalowanie Rascha zakłada, że respondenci o tej samej pozycji na kontinuum badanej właściwości ukrytej (a więc o tym samym parametrze B_n) udzielają, z pewną dokładnością określoną standardowym błędem szacunku, takich samych odpowiedzi na poszczególne pytania ankiety. Innymi słowy, interpretują daną pozycję skali w ten sam sposób. Różnice występujące między oszacowaniami ich pozycji (miejsca) na skali powinny zatem być wynikiem jedynie różnic w intensywności postrzegania przez nich badanej cechy o charakterze ukrytym, a nie wynikiem stosunku do odpowiednich pozycji skali.

Ponieważ w literaturze spotyka się czasem pogląd, że klasyczne skalowanie Rascha (*Rasch model*) jest specjalnym przypadkiem modeli wyniku zadania testowego (*item response theory*) i mimo że celem tego opracowania nie jest wyjaśnianie tej kwestii, w celu uniknięcia nieporozumień pokrótce przypomniane zostaną zasadnicze różnice w definiowaniu obciążenia pozycji skali w tych dwóch grupach modeli.

W przypadku modeli wyniku zadania testowego definicja obciążenia/stronniczości jest szersza. Za obciążone pozycje skali (*biased item*) uznaje się te, które [15, s. 393-394]:

- mają to samo oszacowanie intensywności, ale różne nachylenie krzywej charakterystycznej – taki przypadek w skalowaniu Rascha kwalifikowany jest jako brak dopasowania danej pozycji skali do modelu, a nie obciążenie, i wykrywany jest przez statystyki dopasowania,
- mają to samo oszacowanie intensywności, ale ich krzywe charakterystyczne nie mają lewostronnej asymptoty równej 0 i prawostronnej równej 1; w takiej sytuacji daną pozycję skali w przypadku skalowania Rascha traktuje się jako nie pasującą do modelu i diagnozuje za pomocą statystyk dopasowania.

Ponadto trzeba zauważyć, że metody do wykrywania obciążonych pozycji skali w skalowaniu Rascha działają tylko i wyłącznie wtedy, gdy wcześniej zostało sprawdzone dopasowanie „podejrzanych” pozycji skali do założeń metody za pomocą statystyk dopasowania. Dopiero wtedy można próbować prawidłowo diagnozować pozycje obciążone.

4. Metody wykrywania różnic w funkcjonowaniu pozycji skali – analiza DIF²

W obrębie wszystkich typów skalowań Rascha wyróżnia się trzy podejścia do wykrywania pozycji skali o zróżnicowanym sposobie funkcjonowania wobec różnych grup respondentów.

Pierwsze ujęcie opiera się na osobnym oszacowaniu intensywności pozycji skali dla różnych (zwykle dwóch) grup respondentów, a następnie porównaniu różnic między nimi za pomocą statystyki *t*-Studenta zdefiniowanej w następujący sposób [15, s. 396]:

$$t = \frac{D_{i1} - D_{i2}}{\sqrt{s_{i1}^2 + s_{i2}^2}}, \quad (4)$$

gdzie: D_{i1} – intensywność *i*-tej pozycji skali w pierwszej podgrupie,

D_{i2} – intensywność *i*-tej pozycji skali w drugiej podgrupie,

s_{i1}^2 – standardowy błąd szacunku dla D_{i1} ,

s_{i2}^2 – standardowy błąd szacunku dla D_{i2} .

Celem tego podejścia jest sprawdzenie, czy sposób reakcji na badaną pozycję skali jest zależny – tak jak zakłada skalowanie Rascha – tylko od współdziałania pozycji (miejsca) respondenta na skali mierzącej badaną kategorię ukrytą i intensywności pozycji skali. Jeśli różnica między oszacowaniami intensywności badanej pozycji skali okaże się statystycznie istotna, może to wskazywać na wpływ rozpatrywanej cechy deskryptywnej na postrzeganie pozycji skali i jest to wówczas

² DIF – *differential item functioning* – takie określenie występuje w literaturze polskojęzycznej [5, s. 188].

przesłanka do dokładniejszego zbadania sposobu rozumienia danej pozycji skali przez respondentów.

Statystycznie różne oszacowania intensywności jednej z pozycji skali oznaczają, że gdyby rozpatrywano odpowiedzi dwóch osób o tym samym wyniku ogólnym na skali, a zatem o tym samym oszacowaniu poziomu badanej kategorii ukrytej wyrażonym oszacowaniem B_n , to okazałoby się, że prawdopodobieństwo wybrania danej kategorii odpowiedzi jest dla każdego z nich inne i zależne od grupy, do której osoba przynależy. To przeczyłoby założeniu skalowania Rascha, które mówi, że prawdopodobieństwo udzielania x -tej kategorii odpowiedzi na i -te pytanie przez n -tego respondenta zależy tylko i wyłącznie od postrzegania przez tego respondenta badanej cechy ukrytej i od intensywności pozycji skali.

Metoda oparta na oddzielnej kalibracji intensywności pozycji skali w różnych podgrupach respondentów ograniczona jest do porównań parami. W przypadku konieczności porównywania większej liczby podgrup wykorzystuje się statystykę $\chi^2(UB)$ (ujęcie drugie). Statystyka ta oparta jest na porównaniu różnic między odpowiedziami obserwowanymi a oczekiwanymi i zdefiniowana jest następująco [16, s. 167]:

$$\chi^2(UB)_i = \sum_{j=1}^J \frac{\left(\sum_{n \in j} x_{ni} - \sum_{n \in j} p_{ni} \right)^2}{\sum_{n \in j} w_{ni}}, \quad (5)$$

gdzie: J – liczba porównywanych podgrup,

N_j – liczba osób w każdej podgrupie,

x_{ni} – odpowiedź n -tej osoby na i -tą pozycję skali,

p_{ni} – prawdopodobieństwo po-prawnej odpowiedzi n -tej osoby na i -tą pozycję skali,

w_{ni} – wariancja odpowiedzi.

Wzór (5) znajduje zastosowanie w przypadku klasycznego skalowania Rascha. W celu obliczenia analogicznej statystyki w porządkowym skalowaniu Rascha należy prawdopodobieństwo p_{ni} zastąpić wartością oczekiwaną odpowiedzi zgodnie ze wzorem (3).

W ujęciu trzecim wykrywanie różnie funkcjonujących pozycji skali oparte jest na analizie krzywej charakterystycznej pozycji skali³. Procedura wykorzystuje podstawową zasadę teorii cech ukrytych mówiącą, że respondenci o tej samej pozycji względem mierzonej kategorii ukrytej powinni mieć taką samą wartość oczekiwaną odpowiedzi na daną pozycję skali bez względu na określające ich cechy deskryptywne. A zatem koncepcja wykrywania pozycji obciążonych opiera się w tym podejściu na porównaniu między sobą wartości oczekiwanych odpo-

³ Krzywa charakterystyczna pozycji skali (*item characteristic curve* – ICC) przedstawia zależność między wartością oczekiwaną odpowiedzi udzielanych przez respondentów a wartością mierzonej zmiennej ukrytej.

wiedzi w badanych grupach respondentów, ale w podziale na odpowiednie przedziały. Przedziały te konstruowane są tak, aby zawierały respondentów jak najbardziej jednorodnych pod względem badanej właściwości ukrytej. Krzywe charakterystyczne pozycji skali wykazujące obciążenie wykreślone dla badanych grup respondentów nie będą się pokrywały w niektórych lub we wszystkich przedziałach. Istotność tych rozbieżności oceniona jest za pomocą analizy wariancji przeprowadzonej na wystandaryzowanych resztach, czyli różnicach między oczekiwanymi i obserwowanymi odpowiedziami. Bada się istotność wpływu cech deskryptywnych na różnice w położeniu krzywych, a tym samym w intensywnościach pozycji skali, ale także istotność interakcji pomiędzy charakterystyką opisującą respondentów a przedziałem, do którego zostali zakwalifikowani.

W celu zidentyfikowania pozycji obciążonej najpierw identyfikuje się respondenta ze względu na podejrzaną o powodowanie różnic w rozumieniu danej pozycji skali cechę deskryptywną i ze względu na przynależność do danego przedziału klasowego (*class interval*), a następnie analizuje się wystandaryzowane reszty obliczane według wzoru (6).

$$z_{ni} = \frac{x_{ni} - E[X_{ni}]}{\sqrt{V[X_{ni}]}} \quad (6)$$

gdzie: x_{ni} – kod odpowiedzi,
 $E[X_{ni}]$ – wartość oczekiwana odpowiedzi,
 $V[X_{ni}]$ – wariancja odpowiedzi.

W celu wykrycia występowania obciążenia sprawdza się istotność dwóch efektów: wpływu cechy deskryptywnej na intensywność badanej pozycji skali i wpływu interakcji pomiędzy poziomem badanej cechy deskryptywnej a przynależnością do przedziału klasowego. Wpływ drugiego efektu głównego, jakim jest wpływ podziału na przedziały klasowe, nie jest związany z wykrywaniem obciążeń, ale służyć może do oceny dopasowania pozycji skali do założeń metody. To podejście do wykrywania niedopasowania nie jest szeroko opisane w literaturze, ale wykorzystywane jest w programie RUMM2020.

5. Korekta obciążenia

Wykrywanie obciążonych pozycji skali jest bardzo użyteczne na etapie konstrukcji złożonej skali pomiarowej. Zwykle ich usunięcie pozwala poprawić trafność konstruowanego narzędzia. Jednak w przypadku skal o niewielkiej liczbie pozycji nie zawsze można sobie na to pozwolić. W takiej sytuacji przeprowadza się operację „zrównywania” (*equating, addressing*) [18]. Procedura ta pozwala skorygować obciążenie pozycji bez konieczności ich usuwania z analizy. Przebiega ona w ten sposób, że pozycję obciążoną dzieli się na tyle nowych pozycji skali, ile było grup respondentów wyróżnionych na bazie poziomów cechy deskryptywnej powodującej różnice w rozumieniu pozycji skali. Tym samym zamiast jednej obciążonej pozycji skali uzyskuje się k nowych pozycji (gdzie k odpowiada liczbie grup

respondentów), mających braki danych dla respondentów nie należących do k -tej grupy.

W przypadku skali już działającej również powinno się sprawdzać jej trafność z uwzględnieniem analizy obciążenia. W razie jego wykrycia dokonuje się korekty za pomocą „zrównywania”.

6. Ilustracja

Do analizy wykorzystano skalę mierzącą stosunek do osób starszych złożoną z ośmiu następujących pozycji⁴:

A1. Dzięki doświadczeniu są ciągle potrzebni

A2. Gwarantują zachowanie tradycyjnych wartości w społeczeństwie

A3. Młodsze generacje mogą korzystać z ich obecności, wiedzy, doświadczenia

A4. Społeczeństwo powinno brać pod uwagę prawa starszych

A5. Społeczeństwo powinno brać pod uwagę problemy osób starszych

A6. Osoby starsze są nieproduktywne i tylko stanowią obciążenie dla społeczeństwa

A7. Starsze osoby stanowią przeszkodę dla zmian

A8. Starsze osoby stanowią ciężar dla społeczeństwa

Wykorzystane dane pochodziły z „Badania poglądów na zagadnienia ludnościowe oraz politykę ludnościową PPA2 (Population Policy Attitudes Survey)” przeprowadzonego w IV kwartale 2001 r. na części próby BAEL⁵ przez Instytut Statystyki i Demografii Szkoły Głównej Handlowej w Warszawie przy współpracy z Głównym Urzędem Statystycznym w ramach projektu „Population Policy Acceptance Study. The Viewpoint of Citizens and Policy Actors regarding the Management of Population Related Change DIALOG” – V Program Ramowy Unii Europejskiej.

Respondenci mieli do wyboru jedną z pięciu kategorii odpowiedzi, z których każda była kodowana liczbowo: 1 – zdecydowanie się zgadzam, 2 – zgadzam się, 3 – ani się zgadzam, ani się nie zgadzam, 4 – nie zgadzam się, 5 – zdecydowanie się nie zgadzam. Ze sposobu etykietowania wynikało, że osoby z wyższym wynikiem ogólnym uzyskanym przez zsumowanie odpowiedzi charakteryzowały się bardziej negatywnym stosunkiem do osób starszych niż osoby o niższym wyniku ogólnym. Aby ułatwić interpretację wyników, pytania zostały przekodowane w taki sposób, aby wyższy wynik ogólny odpowiadał bardziej pozytywnemu stosunkowi do osób starszych. W operacji tej uwzględniono także przeciwne zorientowanie stwierdzeń A6, A7 i A8. Pozostawiona również została środkowa kategoria neutralna „ani się zgadzam, ani się nie zgadzam”.

Skala została uznana za odpowiadającą założeniom skalowania Rascha. Do oceny dopasowania poszczególnych stwierdzeń wykorzystano mierniki wystanda-

⁴ W dalszej części tekstu termin „pozycja skali” używany będzie zamiennie z terminem „stwierdzenie”.

⁵ BAEL – badanie aktywności ekonomicznej ludności.

ryzowane oparte na statystyce χ^2 – *SIMS* i *SOMS*. Szczegółowe wyniki prezentuje tab. 1.

Wszystkie pozycje badanej skali okazały się być zgodne założeniami metody – wartości statystyk *SIMS* i *SOMS* nie wykroczyły poza przedział $<-2, 2>$ [2, s. 209].

Tabela 1. Ocena dopasowania poszczególnych stwierdzeń do założeń porządkowego skalowania Rascha

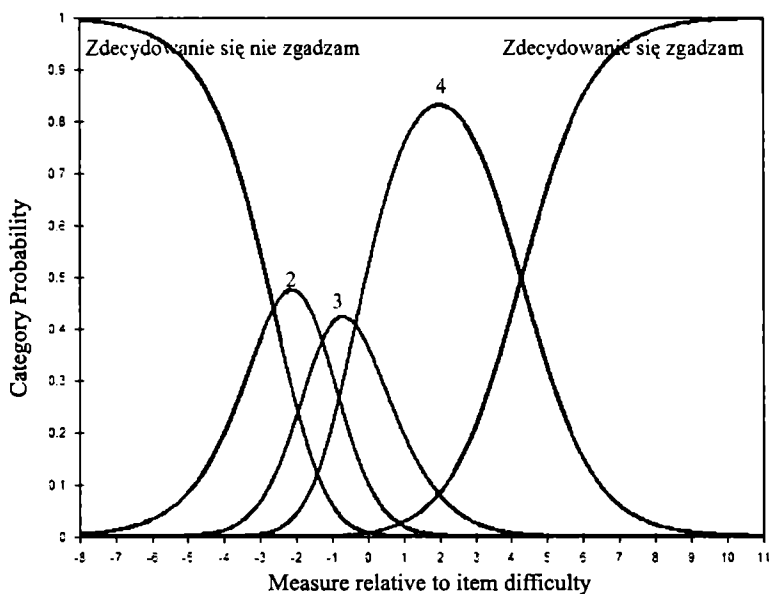
Stwierdzenia	OIPS*	B.S.**	Statystyki dopasowania			
			IMS	SIMS	OMS	SOMS
(r)A7. Starsze osoby stanowią przeszkodę dla zmian	0,32	0,15	0,98	-0,1	0,94	-0,4
(r)A6. Osoby starsze są nieproduktywne i tylko stanowią obciążenie dla społeczeństwa	0,21	0,15	1,14	1,1	1,09	0,7
(r)A8. Starsze osoby stanowią ciężar dla społeczeństwa	0,13	0,15	1,13	1	1,11	0,9
A2. Gwarantują zachowanie tradycyjnych wartości w społeczeństwie	0,08	0,15	0,93	-0,5	0,84	-1,3
A3. Młodsze generacje mogą korzystać z ich obecności, wiedzy, doświadczenia	0,02	0,15	0,82	-1,5	0,81	-1,5
A4. Społeczeństwo powinno brać pod uwagę prawa starszych	-0,17	0,15	0,78	-1,9	0,78	-1,8
A1. Dzięki doświadczeniu są ciągle potrzebni	-0,26	0,15	1,22	1,7	1,14	1,1
A5. Społeczeństwo powinno brać pod uwagę problemy osób starszych	-0,33	0,15	0,93	-0,5	0,92	-0,6
Średnia	0,00	0,15	0,99	-0,1	0,95	-0,4
Odchylenie standardowe	0,22	0,00	0,15	1,22	0,14	1,0

*OIPS – oszacowanie intensywności pozycji skali. **B.S. – błąd szacunku.

Źródło: obliczenia własne.

Również zastosowana skala odpowiedzi okazała się działać prawidłowo (rys. 1). Każda z kategorii odpowiedzi była najbardziej prawdopodobna i nie zaobserwowano zakłóceń w kolejności ich występowania wynikającej z percepcji respondentów. Można było zatem stwierdzić, że rzeczywiście zastosowana skala odpowiedzi rozdzielała kontinuum badanej właściwości empirycznej w sposób zgodny z zamierzeniem autora narzędzia.

Do sprawdzenia, które ze stwierdzeń nie działają w identyczny sposób dla wszystkich grup respondentów, wykorzystane zostało podejście oparte na oddzielnym oszacowaniu intensywności pozycji skali dla dwóch grup respondentów. Rozpatrzono podział według grup wieku – wyróżnione zostały dwie kategorie: respondenci w wieku do 50 lat życia i starsi, a także według płci. Po niezależnym oszacowaniu intensywności pozycji skali dla grup respondentów sprawdzono, czy występujące między nimi różnice są statystycznie istotne. Wyniki zawierają tab. 2 i 3.



1 (zdecydowanie się nie zgadzam)	2	3	4	5 (zdecydowanie się zgadzam)
--	---	---	---	---------------------------------

Rys. 1. Odległości między kategoriami odpowiedzi wynikające z percepcji respondentów
Źródło: opracowanie własne.

Wykonane obliczenia pozwoliły stwierdzić, że na poziomie istotności $\alpha = 0,01$ występują różnice w rozumieniu stwierdzenia A6. *Osoby starsze są nieproduktywne i tylko stanowią obciążenie dla społeczeństwa* między osobami w wieku do 50 lat i starszymi. Podobnie przedstawia się sytuacja w przypadku stwierdzeń: A8. *Starsze osoby stanowią ciężar dla społeczeństwa* oraz A4. *Spółeczeństwo powinno brać pod uwagę prawa starszych*, z tą różnicą, że jest to prawda przy przyjętym poziomie istotności $\alpha = 0,05$. Istnieje zatem podejrzenie, że te trzy stwierdzenia (szare tło w tabeli) wpływają negatywnie na trafność całej skali.

W przypadku stwierdzeń A6 i A8 oszacowanie w grupie osób w wieku do 50 lat okazało się wyższe niż w grupie osób w wieku 50 lat lub więcej. Oznacza to, że po porównaniu osób o tym samym wyniku ogólnym z tych dwóch grup wiekowych, okaże się, że te z „grupy młodszej” będą bardziej zdecydowanie nie zgadzać się⁶ z każdym z tych stwierdzeń niż osoby z „grupy starszej”. A zatem osoby z „grupy starszej” o tym samym stosunku do osób starszych co te z „grupy młodszej” będą się bardziej stanowczo zgadzały (zakładając chwilowo dla uproszczenia rozu-

⁶ Należy pamiętać o przeciwnym zorientowaniu tego stwierdzenia w porównaniu do kierunku całej skali pomiarowej;

mowania, że przeciwieństwem niezgadzenia się jest zgadzanie się) z opiniami, że: (r)A6. *Osoby starsze są nieproduktywne i tylko stanowią obciążenie dla społeczeństwa* oraz (r)A8. *Starsze osoby stanowią ciężar dla społeczeństwa*.

Tabela 2. Porównanie oszacowań intensywności pozycji skali dla dwóch grup wiekowych

Stwierdzenia	D_{11}^*	D_{12}^{**}	t	Krytyczny poziom istotności
A1. <i>Dzięki doświadczeniu są ciągle potrzebni</i>	-0,348	-0,224	-0,3772	0,7064
A2. <i>Gwarantują zachowanie tradycyjnych wartości w społeczeństwie</i>	-0,123	-0,313	0,5314	0,5957
A3. <i>Młodsze generacje mogą korzystać z ich obecności, wiedzy, doświadczenia</i>	-0,234	0,307	-1,7429	0,0830
A4. <i>Spółeczeństwo powinno brać pod uwagę prawa starszych</i>	-0,648	0,094	-2,2336	0,0267*
A5. <i>Spółeczeństwo powinno brać pod uwagę problemy osób starszych</i>	-0,338	0,033	-1,1662	0,2450
(r)A6. <i>Osoby starsze są nieproduktywne i tylko stanowią obciążenie dla społeczeństwa</i>	0,981	0,074	3,0627	0,0025**
(r)A7. <i>Starsze osoby stanowią przeszkodę dla zmian</i>	0,055	-0,062	0,4117	0,6810
(r)A8. <i>Starsze osoby stanowią ciężar dla społeczeństwa</i>	0,654	0,091	1,9787	0,0493*

* D_{11} – oszacowanie intensywności i -tej pozycji skali w grupie osób w wieku do 50 lat.

** D_{12} – oszacowanie intensywności i -tej pozycji skali w grupie osób w wieku 50 lat i więcej.

Źródło: opracowanie własne.

Tabela 3. Porównanie oszacowań intensywności pozycji dla respondentów różnej płci

Stwierdzenia	D_{11}^*	D_{12}^{**}	t	Krytyczny poziom istotności
A1. <i>Dzięki doświadczeniu są ciągle potrzebni</i>	-0,295	-0,158	-0,533	0,297
A2. <i>Gwarantują zachowanie tradycyjnych wartości w społeczeństwie</i>	-0,129	0,085	-0,789	0,216
A3. <i>Młodsze generacje mogą korzystać z ich obecności, wiedzy, doświadczenia</i>	-0,095	-0,502	1,422	0,078
A4. <i>Spółeczeństwo powinno brać pod uwagę prawa starszych</i>	-0,207	-0,55	1,131	0,130
A5. <i>Spółeczeństwo powinno brać pod uwagę problemy osób starszych</i>	-0,212	-0,175	-0,129	0,449
(r)A6. <i>Osoby starsze są nieproduktywne i tylko stanowią obciążenie dla społeczeństwa</i>	0,497	1,117	-2,714	0,004*
(r)A7. <i>Starsze osoby stanowią przeszkodę dla zmian</i>	-0,074	0,023	-0,389	0,349
(r)A8. <i>Starsze osoby stanowią ciężar dla społeczeństwa</i>	0,515	0,159	1,498	0,068

* D_{11} – oszacowanie intensywności i -tej pozycji skali wśród kobiet.

** D_{12} – oszacowanie intensywności i -tej pozycji skali wśród mężczyzn.

Źródło: opracowanie własne.

Również w przypadku stwierdzenia A4. *Spółeczeństwo powinno brać pod uwagę prawa starszych osoby ze „starszej grupy” intensywniej się z nim zgadzały.*

Po oddzielnej kalibracji stwierdzeń z uwzględnieniem podziału według płci tylko jedno ze stwierdzeń okazało się mieć dwa statystycznie różne oszacowania intensywności. Było to stwierdzenie (r)A6. *Osoby starsze są nieproduktywne i tylko stanowią obciążenie dla społeczeństwa*; wyższe oszacowanie wystąpiło w grupie mężczyzn. Sugerowało to, że w porównaniu z kobietami mężczyźni częściej i bardziej intensywnie się z nim nie zgadzali, zawsze przy porównywaniu osób o tych samych wynikach ogólnych.

Różne oszacowania intensywności pozycji skali (stwierdzeń) w badanych grupach respondentów wskazywać mogą na różnice w interpretowaniu stwierdzeń. Należy zbadać, co poszczególne grupy respondentów rozumieją przez poszczególne sformułowania. Przydatne do tego okazać się może przeprowadzenie odpowiednich badań o charakterze jakościowym. Należy również sprawdzić, czy uzyskane wyniki nie są skutkiem przeciwnego zorientowania dwóch z trzech wykazujących obciążenie stwierdzeń.

7. Podsumowanie

W artykule zaprezentowany został sposób sprawdzania trafności narzędzia pomiarowego z zastosowaniem skalowania Rascha. W wyniku przeprowadzonych analiz udało się pogłębić wiedzę na temat sposobu interpretowania przez badanych respondentów poszczególnych pozycji skali tworzących analizowaną złożoną skalę pomiarową. Wykazano, że:

1) występują istotne różnice w rozumieniu stwierdzenia (r)A6. *Osoby starsze są nieproduktywne i tylko stanowią obciążenie dla społeczeństwa* w grupach wyróżnionych zarówno na podstawie wieku, jak i płci;

2) występują istotne różnice w rozumieniu stwierdzeń A4 oraz (r)A8 w dwóch badanych grupach wiekowych.

Ponieważ analiza obciążenia została wykonana tylko jednym z możliwych sposobów, dalsze prace będą miały na celu sprawdzenie, czy podobne rezultaty uzyska się, korzystając z innych metod badania obciążenia.

Ze względu na to, że analizowana skala została już wykorzystana w badaniu właściwym, co ogranicza możliwości jej korekty, przed analizowaniem jej wyników przeprowadzona powinna zostać procedura „zrównywania”. Wskazane jest również przeprowadzenie dokładniejszych badań nad sposobem rozumienia przez respondentów problematycznych pozycji skali, a następnie odpowiednie ich sformułowanie przed ewentualnym ponownym zastosowaniem skali.

Ograniczeniem zaprezentowanego podejścia do badania trafności skali jest to, że znajduje ono zastosowanie tylko w przypadku jednowymiarowych skal skumulowanych ocen. Jednocześnie podkreślić należy, że sposób określania stopnia jednowymiarowości skali w sensie skalowania Rascha pozostaje w dalszym ciągu niesatysfakcjonujący.

Załącznik – miary dopasowania

1. OMS – *outfit mean square* – dla i -tej pozycji skali

$$OMS_i = \frac{\sum_{n=1}^N z_{ni}^2}{N},$$

gdzie: z_{ni} – wystandaryzowana różnica między odpowiedzią obserwowaną a wy-modelowaną n -tego respondenta na i -tą pozycję skali,

N – liczba osób,

OMS ma rozkład χ^2 o N stopniach swobody, wartości oczekiwanej równej 1 i wariancji:

$$s_i^2 = \frac{\sum_{n=1}^N (C_{ni} / W_{ni}^2)}{N^2} - \frac{1}{N},$$

gdzie C_{ni} jest kurtozą X_{ni} .

2. SOMS – *standardized weighted mean square outfit* – to przekształcony według transformacji Wilsona i Hilferty'ego OMS.

$$SOMS_i = \left(OMS_i^{1/3} - 1 \right) \left(\frac{3}{s_i} \right) + \frac{s_i}{3}.$$

SOMS ma rozkład t -Studenta.

3. IMS – *infit mean square* – dla i -tej pozycji skali

$$IMS_i = \frac{\sum_{n=1}^N z_{ni}^2 W_{ni}}{\sum_{n=1}^N W_{ni}},$$

gdzie: z_{ni} – wystandaryzowana różnica między odpowiedzią obserwowaną a wy-modelowaną n -tego respondenta na i -tą pozycję skali,

W_{ni} – wariancja różnic między odpowiedzią oczekiwaną a obserwowaną;

IMS ma rozkład χ^2 o N stopniach swobody, wartości oczekiwanej równej 1 i wariancji:

$$q_i^2 = \frac{\sum_{n=1}^N (C_{ni} - W_{ni}^2)}{\left(\sum_{n=1}^N W_{ni} \right)^2},$$

gdzie C_{ni} jest kurtozą X_{ni} .

4. SIMS – *standardized weighted mean square infit* to przekształcony według transformacji Wilsona i Hilferty’ego IMS. SIMS ma rozkład *t*-Studenta.

$$\text{SIMS}_i = \left(\text{IMS}_i^{1/3} - 1 \right) \left(\frac{3}{q_i} \right) + q_i / 3.$$

W celu obliczenia powyższych miar dla *n*-tego respondenta należy zamienić indeksy: *n* na *i*, *i* na *n*, *N* na *L* (gdzie *L* to liczba pozycji skali).

5. $\chi^2(UB)$

$$\chi^2(UB)_i = \sum_{j=1}^J z_{ji}^2 = \sum_{j=1}^J \frac{\left(\sum_{n \in j}^{N_j} X_{ni} - \sum_{n \in j}^{N_j} E[X_{ni}] \right)^2}{\sum_{n \in j}^{N_j} w_{ni}},$$

gdzie: *J* – liczba porównywanych podgrup,
N_j – liczba osób w każdej podgrupie,
X_{ni} – odpowiedź *n*-tej osoby na *i*-te stwierdzenie,
E[X_{ni}] – wartość oczekiwana udzielonej odpowiedzi *n*-tej osoby na *i*-te stwierdzenie,
w_{ni} – wariancja odpowiedzi.

Literatura

- [1] Andrich D. (1988), *Rasch Models for Measurement*, Newbury Park, California: Sage.
- [2] Bond G.T., Fox M.Ch. (2001), *Applying The Rasch Model: Fundamental Measurement in the Human Sciences*, Lawrence Erlbaum Associates, Publishers, Mahwah, New Jersey, London.
- [3] Churchill G.A. (2002), *Badania marketingowe. Podstawy metodologiczne*, Wyd. Naukowe PWN, Warszawa.
- [4] Green E.K. (1996), *Applications of the Rasch Model to Evaluation of Survey Data Quality*, „New Directions for Evaluation”, no. 70, Jossey-Bass Publishers, Summer.
- [5] Hornowska E. (2001), *Testy psychologiczne. Teoria i praktyka, Wykłady z psychologii*, t. 6, Wydawnictwo Naukowe Scholar, Warszawa.
- [6] Interpreting RUMM2020 PART I – Dichotomous Data, August 2004, <http://www.rummlab.com/>.
- [7] Interpreting RUMM2020 PART II – Polytomous Data, August 2004, <http://www.rummlab.com/>.
- [8] Kaczmarczyk S. (2003), *Badania marketingowe. Metody i techniki*, PWE, Warszawa.
- [9] Nachmias D., Frankfort-Nachmias Ch. (2001), *Metody badawcze w naukach społecznych*, Wydawnictwo Zysk i Spółka, Poznań.
- [10] Ostasiewicz W. (2003), *Istota pomiaru statystycznego*, [w:] *Pomiar statystyczny*, red. W. Ostasiewicz, AE, Wrocław.
- [11] Rószkiewicz M. (2002), *Metody ilościowe w badaniach marketingowych*, Wyd. Naukowe PWN, Warszawa.

- [12] Sagan A. (2004), *Badania marketingowe. Podstawowe kierunki*, AE, Kraków.
- [13] Smith R.M. (1994), *A Comparison of the Power of Rasch Total and Between-item Fit Statistics to Detect Measurement Disturbances*, „Educational and Psychological Measurement”, vol. 54, no. 1, Spring.
- [14] Smith R.M. (1994), *Detecting Item Bias in the Rasch Rating Model*, „Educational and Psychological Measurement”, vol. 54, no. 1, Winter.
- [15] Smith R.M. (2004), *Detecting Item Bias with the Rasch Model*, [w:] *Introduction to Rasch Measurement. Theory, Models and Applications*, red. E.V. Smith, R.M. Smith, Jam Press, Maple Grove, Minnesota.
- [16] Smith R.M. (1991), *Item and Person Analysis with the Rasch Model*, Mesa Press, Chicago.
- [17] Smith R.M., Schumacker R.E., Bush M.J. (1998), *Using Item Squares to Evaluate Fit to the Rasch Model*, „Journal of Outcome Measurement”, 2(1).
- [18] Snider P.D. (2003), *Exploring the Relationships between Individualism and Collectivism and Attitudes towards Counseling among Ethnic Chinese, Australian, and American University Students*, A thesis submitted in partial completion of the requirements for the degree of Doctor of Philosophy, Murdoch University, September,
<http://www.lib.murdoch.edu.au/adt/browse/view/adt-MU20040430.140708>.

ASSESSMENT OF THE VALIDITY OF THE SCALE WITH THE RASCH MODEL

Summary

The aim of the article is to present the Rasch Model and its application to the assessment of the validity of the scale. The article begins with the presentation of the method pointing out the differences between the Rasch Model and the Rasch Rating Scale Model. Then the techniques for detecting item bias with the Rasch Model are presented. The theoretical part of the article is followed by the application part where items of the scale measuring attitude towards the old are verified in terms of the bias.