ASK, 2006, nr 15, strony 107–131

Dorota Węziak
Szkoła Główna Handlowa w Warszawie

OCENA JAKOŚCI SKALI ZE SZCZEGÓLNYM UWZGLĘDNIENIEM JEJ RZETELNOŚCI I TRAFNOŚCI ZA POMOCĄ SKALOWANIA RASCHA

Celem niniejszego artykułu jest prezentacja mało znanej do tej pory w Polsce metody analizy statystycznej, jaką jest skalowanie Rascha, a szczególnie sposobu jej wykorzystania do oceny rzetelności i trafności kumulatywnej skali pomiarowej. Miary rzetelności i trafności uzyskane w wyniku skalowania Rascha stanowią alternatywę do powszechnie do tej pory stosowanych w Polsce metod oceny rzetelności i trafności, o których piszą między innymi David Magnusson, Jerzy Brzeziński, Jarosław Górniak, Andrzej Machowski, Adam Sagan, zwłaszcza że te pierwsze stosowane mogą być tylko do skal o pozycjach równoległych, zaś miary uzyskane w wyniku skalowania Rascha wykorzystuje się do oceny skal o pozycjach o zróżnicowanej intensywności. W części pierwszej przedstawiono podstawy metodologiczne skalowania Rascha, natomiast część druga służy praktycznemu zastosowaniu tej techniki. Podkreślić należy, że ponieważ rodzina modeli Rascha jest bardzo liczna, tekst ten skupia się szczegółowo na skalowaniu porządkowym, które jest przydatne do badania cech latentnych mierzonych na skalach porządkowych o zróżnicowanej intensywności (skalach kumulatywnych).
Główne pojęcia: model Rascha, skalowanie Rascha, porządkowe skalowanie Rascha, skala porządkowa, rzetelność skali pomiarowej, trafność skali pomiarowej.

Uwagi do Autora lub prośby o nadbitki prosimy kierować: Instytut Statystyki i Demografii, Szkoła Główna Handlowa, 02–554 Warszawa, Al. Niepodległości 162, e-mail: dorota.weziak@sgh.waw.pl

Wprowadzenie

Przeprowadzając badanie za pomocą kwestionariusza ankietowego każdy badacz staje wobec problemów związanych z doborem narzędzia pomiarowego adekwatnego do postawionego problemu badawczego, a także z doborem odpowiednich jednostek badania. Poprzez adekwatne narzędzie rozumie się po pierwsze taki instrument, który będzie dostarczał wiarygodnych wyników, a więc rzetelnych (o takiej samej precyzji z pomiaru na pomiar) i trafnych (dotyczących wyłącznie badanej właściwości). Po drugie będzie zrozumiały dla respondentów, czyli będą oni potrafili go stosować zgodnie z zamierzeniem autora. Dobór jednostek badania jest równie istotny, ale to zagadnienie nie znajduje się w obrębie tematyki tego opracowania i dlatego nie będzie szerzej omawiane. W rozwiązaniu problemów dotyczących adekwatności narzędzia pomiarowego może być użyteczne skalowanie Rascha[1].

Skalowanie Rascha to metoda służąca skonstruowaniu skali interwałowej z danych zero-jedynkowych lub porządkowych, przy założeniu, że stwierdzenia tworzące skalę pomiarową mają zróżnicowaną intensywność. W 1960 roku George Rasch stworzył metodę dla danych zero-jedynkowych typu tak/nie, występowanie/brak występowania, zgadzam się/nie zgadzam się, dobrze/źle (*dychotomiczne skalowanie Rascha*). Obecnie obok wielowymiarowego skalowania Rascha istnieje cała rodzina jednowymiarowych modeli. Na przykład dla danych porządkowych możliwe jest stosowanie następujących technik skalowania:

1) skalowanie porządkowe *(rating scale model)*, gdy wszystkie pozycje skali[2] mają identyczne skale odpowiedzi;
2) skalowanie częściowe (*partial credit model*), w przypadku różnolicznych lub różnie brzmiących kategorii odpowiedzi dla całej jednowymiarowej skali złożonej;
3) skalowanie rangowe *(rank model)*, w przypadku stosowania rangowej skali odpowiedzi;
4) skalowanie wieloaspektowe *(many-facet Rasch model)*, gdy analiza jest wielowymiarowa w tym sensie, że badane są obiekty, ich cechy i oceniający je „sędziowie".

[1] Polski odpowiednik nazwy analizy, która w literaturze anglojęzycznej znana jest jako *Rasch Measurement* lub *Rasch Model*, pochodzi z Górniak (2000: 67).

[2] Przez pozycje skali rozumie się stwierdzenia, które poddawane są ocenie przez respondentów.

W literaturze czasem skalowanie wieloaspektowe traktowane jest jako szczególny rodzaj skalowania częściowego. Poza tym w obrębie skalowania wieloaspektowego wyróżnia się także trzy podejścia hybrydowe (Myford i Wolfe 2004: 503).

W swojej klasycznej jednowymiarowej postaci skalowanie Rascha może być użyte m.in. do ustalenia: (1) czy wszyscy respondenci w ten sam sposób postrzegają skalę, (2) czy skala jest odpowiednio dopasowana do grupy respondentów podlegających badaniu, czyli czy nie jest zbyt trudna lub zbyt łatwa, (3) czy pozycje skali pokrywają w sposób regularny kontinuum badanej właściwości empirycznej – co ma swoje bezpośrednie przełożenie na trafność teoretyczną instrumentu pomiarowego, (4) czy skala jest rzetelna i trafna, (5) czy liczba kategorii odpowiedzi jest optymalna. W tym opracowaniu skupiono się na pierwszych czterech aspektach.

Skalowanie Rascha jest metodą probabilistyczną, a jakość jej wyników można przetestować statystycznie. Najbardziej ogólnie polega ona na tym, że opierając się na przedstawionych poniżej założeniach tworzony jest teoretyczny model, a następnie sprawdzane jest, w jakim stopniu pasują do niego zebrane w wyniku badania dane.

Założenia leżące u podstaw skalowania dychotomicznego są następujące (Bond i Fox 2001: xix):

1) jednowymiarowość (badamy na raz tylko jedną cechę będącą obiektem skalowania),
2) większe prawdopodobieństwo faktu, iż wszyscy respondenci odpowiedzą poprawnie na łatwiejsze pozycje skali niż na trudniejsze,
3) większe prawdopodobieństwo prawidłowego zaliczenia pozycji skali przez osobę o większych zdolnościach.

W przypadku skalowania porządkowego pierwsze założenie pozostaje takie samo, natomiast w przypadku kolejnych dwóch mówi się, że osoby z wysokim wynikiem ogólnym skali są bardziej skłonne do zgadzania się z poszczególnymi stwierdzeniami niż osoby o niższym wyniku ogólnym. A także, że jest bardziej prawdopodobne, iż respondenci odpowiedzą twierdząco na pytania łatwiejsze do zaakceptowania (o mniejszej intensywności) i odwrotnie, taka odpowiedź jest mniej prawdopodobna w przypadku pytań o większej intensywności.

Skalowanie Rascha dla skali porządkowej

Skalowanie porządkowe należy stosować w przypadku baterii stwierdzeń mierzonych na skalach porządkowych, przy założeniu, że ich skale odpowiedzi są identyczne.

Analizując dane za pomocą skalowania Rascha bazujemy na dwóch typach informacji: oszacowania trudności czy też intensywności każdej pozycji (*item difficulty*) D_i i oszacowania pozycji każdego respondenta (*person ability, person position*) B_n. Oszacowania te mierzone są logitem[3], a z faktu, że metoda ma charakter probabilistyczny wynika, że każdemu towarzyszy wielkość błędu z nim związana. Metody wykorzystywane do szacowania parametrów D_i i B_n mają charakter iteracyjny i opierają się najczęściej na metodzie największej wiarygodności[4].

Ogólna postać teoretycznego modelu przedstawionego w postaci funkcji wykładniczej dla skali porządkowej jest następująca:

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

gdzie: π_{nix} – prawdopodobieństwo wybrania przez *n*-tego respondenta kategorii odpowiedzi *x* dla *i*-tej pozycji skali; β_n – ocena pozycji *n*-tego respondenta, δ_i – ocena intensywności *i*-tej pozycji skali, τ_x – *x*-ty parametr struktury skali (*x*-ty próg).

Równanie (1) można również przedstawić w formie logitowej:

$$\ln\left(\frac{\pi_{nix}}{\pi_{ni(x-1)}}\right) = \beta_n - (\delta_i + \tau_x) \tag{2}$$

---

[3] Logit to skrót od *log-odds unit*, jest to więc logarytm naturalny szansy rozumianej jako iloraz prawdopodobieństwa sukcesu *p* do prawdopodobieństwa porażki *1-p* i taka definicja będzie obowiązująca w całym tekście.

[4] Najczęściej używane są: warunkowa metoda największej wiarygodności *(CMLE)*, bezwarunkowa metoda największej wiarygodności *(UCON)* oraz krańcowa metoda największej wiarygodności *(MMLE)*, głównie ze względu na dostępność oprogramowania. Każda z tych metod ma swoich zwolenników i przeciwników, ale biorąc pod uwagę dokładność i trafność uzyskiwanych oszacowań można traktować te metody jako równoważne. Należy jednak zwrócić uwagę na problemy z porównywalnością oszacowań uzyskanych różnymi metodami. Szczegółowy opis tych metod można znaleźć w Linacre (2004: 25–72).

gdzie: $\pi_{ni(x-1)}$ – prawdopodobieństwo wybrania przez *n*-tego respondenta kategorii odpowiedzi *x-1* dla i-tej pozycji skali, pozostałe oznaczenia takie jak we wzorze (1).

Postać modelu (2) podkreśla, że logarytmując iloraz prawdopodobieństwa sukcesu przez prawdopodobieństwo porażki (*odds*) otrzymuje się kombinację liniową latentnych parametrów, co oznacza, że oceny respondentów i pozycji skali są bezpośrednio porównywalne. Jest to cecha charakterystyczna wszystkich typów skalowań Rascha.

Ocena dopasowania

Za pomocą skalowania Rascha sprawdzić można, w jakim stopniu zebrane dane odpowiadają zakładanemu modelowi teoretycznemu, a zatem idea oceny dopasowania opiera się na porównaniu dwóch macierzy: zero-jedynkowej macierzy empirycznej i obliczonej macierzy teoretycznej.

Do oceny jakości danej skali służą wskaźniki dopasowania *(fit statistics) INFIT* i *OUTFIT*. Wskaźniki te wykorzystuje się bez względu na rodzaj zastosowanej techniki skalowania.

Wszystkie wskaźniki *OUTFIT* oparte są na sumie wystandaryzowanych, a następnie podniesionych do kwadratu różnic między odpowiednimi odpowiedziami obserwowanymi i modelowanymi i są czułe na odpowiedzi nietypowe (tzn. odpowiadające dużym różnicom między oszacowaniami intensywności pozycji skali D_i, a oszacowaniami pozycji respondentów B_n). Ich konstrukcja jest koncepcyjnie zbliżona do konstrukcji m.in. odchylenia standardowego reszt w regresji liniowej, z tą różnicą, że po pierwsze, w regresji badamy dopasowanie modelu do danych, a w przypadku skalowania Rascha odwrotnie – danych do modelu i po drugie, w przypadku skalowania Rascha reszty są w postaci wystandaryzowanej. Podobna koncepcja konstrukcji miar dopasowania występuje także w analizach opartych na porównywaniu wartości obserwowanych i oczekiwanych.

Wśród wskaźników *OUTFIT* wyróżnia się *Outfit Mean Square (OMS)* i *Standardized Weighted Mean Square Outfit (SOMS).*

Ponieważ wskaźniki *OUTFIT* są czułe na nietypowe, ekstremalne odpowiedzi, w celu zmniejszenia ich wpływu wprowadzono ważenie. Ważoną odmianą wskaźników *OUTFIT* są wskaźniki *INFIT*. Wśród wskaźników *INFIT* wyróżnia się *Infit Mean Square (IMS)* i *Standardized Weighted Mean Square Intfit (SIMS).* Czytelnicy zainteresowani sposobem obliczania wskaźników dopasowania znajdą odpowiednie wzory w załączniku.

Statystyki *OMS* i *IMS* obrazują rozmiar zniekształceń i wykorzystywane są do identyfikowania i oznaczania respondentów oraz pozycji nietypowych (*misfit*). Jak dotąd nie zostały opracowane jednoznaczne wskazówki dotyczące wartości tych statystyk, przy których respondentów bądź stwierdzenia należy uznać za nietypowe. Trevor G. Bond i Christine M. Fox (2001) podjęli próbę podsumowania dotychczasowych ustaleń w tej kwestii i przedstawili propozycję wartości granicznych dla *IMS* i *OMS* w zależności od wielkości próby, co przedstawia tabela 1.

Tabela 1. Wielkości IMS i OMS, dla których przyjmuje się za nietypowe pozycje bądź respondentów w zależności od wielkości próby

Liczebność próby	*IMS, OMS*
Mniej niż 500	większe od 1,3
Od 500 do 1000	większe od 1,2
Powyżej 1000	większe od 1,1

Źródło: Bond i Fox 2001: 209.

Natomiast John M. Linacre (2002) przedstawił wpływ poszczególnych wartości statystyk *IMS* i *OMS* na jakość pomiaru (tabela 2).

Tabela 2. Wpływ wielkości IMS i OMS na jakość pomiaru

IMS, OMS	Znaczenie dla pomiaru
powyżej 2,0	Zniekształca lub obniża jakość pomiaru. Może być wynikiem tylko jednej lub dwóch obserwacji
od 1,5 do 2,0	Nie obniża jakości pomiaru, ale bezużyteczna przy konstrukcji skali
od 0,5 do 1,5	Efektywna przy konstrukcji skali
poniżej 0,5	Mniej przydatna przy konstrukcji skali, ale nie obniża jakości pomiaru. Może powodować zawyżone miary rzetelności.

Źródło: Linacre, John M. 2002. *What do Infit and Outfit, Mean-square and Standardized mean*?. *„Rasch Measurement Transactions", 16:2*. http://www.rasch.org/rmt/rmt162f.htm

Statystyki *SIMS* i *SOMS* pozwalają ustalić podobnie jak ich niewystandaryzowane odpowiedniki *IMS* i *OMS*, czy dane są dopasowane do modelu. Akceptowalne wartości statystyk *SOMS* i *SIMS* zawierają się w przedziale <-2; 2> (Bond i Fox 2001: 209), przy czym wartości mniejsze od 0 wskazują na zmienność

mniejszą niż ta wynikająca z modelu (zbyt małe zróżnicowanie odpowiedzi), a wartości dodatnie na zmienność większą niż wynikająca z modelu (zbyt duże zróżnicowanie). Przypadek drugi jest o wiele groźniejszy dla jakości pomiaru niż przypadek pierwszy.

Ocena rzetelności i trafności

Jakość kwestionariusza ankietowego oceniona jest między innymi na podstawie jego rzetelności i trafności. Rzetelność odnosi się do stabilności instrumentu pomiarowego, a tym samym do wielkości błędu z nim związanego – błędu, który powstaje w sposób losowy, w kolejnych pomiarach dokonywanych za pomocą tego samego narzędzia. Jest to stosunek wariancji prawdziwego wyniku do wariancji całkowitej. Wariancja prawdziwego wyniku rozumiana jest jako różnica między wariancją całkowitą, a wariancją błędów. Taki sposób rozumienia rzetelności stosowany jest również w skalowaniu Rascha.

O rzetelności utworzonego narzędzia informują współczynnik rzetelności respondentów (*Person Reliability Index* – R_P) i współczynnik rzetelności pozycji (*Item Reliability Index, Item Separation Reliability* – R_I). Oba współczynniki rzetelności przyjmują wartości od 0 do 1 i wyrażają procentowy udział wariancji wynikającej z modelu do całkowitej wariancji.

Współczynniki rzetelności *R* można również przekształcić w indeksy rozłączności (pozycji G_P i respondentów G_I) (*Separation Index*) – które oceniają dyspersję intensywności pozycji i stopnia akceptowalności respondentów na kontinuum badanej właściwości latentnej. Indeks rozłączności może przyjmować wartości od 0 do + ∞ i mierzony jest na poziomie interwałowym. Im wyższa wartość indeksu *G*, tym badane stwierdzenia lub respondenci są bardziej rozproszeni na skonstruowanej skali, czyli pokrywają szerszy zakres badanej właściwości, a także ich kolejność pod względem stopnia ich akceptowalności jest bardziej wiarygodna. Zalecenia co do minimalnej do zaakceptowania wartości indeksu rozłączności różnią się w zależności od różnych autorów. Kathy E. Green i Catherine G. Frantom (2000) zalecają, aby indeks rozłączności wynosił przynajmniej 1, co odpowiada rzetelności na poziomie 0,5, natomiast Bond i Fox (2001) piszą o wartości krytycznej na poziomie 2, co odpowiada rzetelności na poziomie 0,8.

Miarą opartą na wskaźniku rozłączności *G* jest *Warstwa (Stratum),* która jest bardzo użyteczna w określeniu wynikającej z odpowiedzi respondentów liczby grup pozycji różniących się między sobą pod względem intensywności (oddzie-

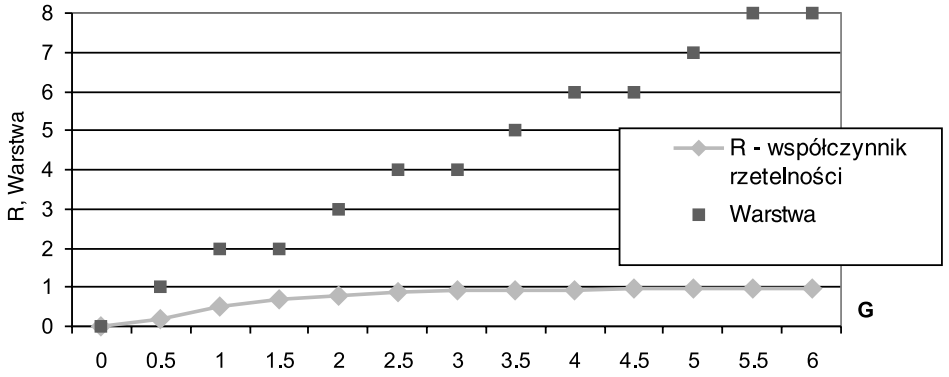
lonych od siebie o przynajmniej trzy standardowe błędy pomiaru) lub grup respondentów różniących się ze względu na badaną cechę latentną.

Podsumowując, należy podkreślić, że współczynnik rzetelności *R*, indeks rozłączności *G* i *Warstwa* mogą być obliczane zarówno dla osób, jak i dla stwierdzeń.

Można zauważyć, że współczynnik rzetelności respondentów R_P jest odpowiednikiem współczynnika α Cronbacha. Przy porównywaniu ich wartości należy jednak pamiętać o „poprawce" ze względu na liczbę stwierdzeń tworzących skalę (*k/(k-1)*), którą zawiera formuła obliczeniowa współczynnika α, a nie posiada współczynnik rzetelności respondentów R_P.

Zależność między wskaźnikiem rozłączności a współczynnikiem rzetelności i warstwą przedstawia wykres 1.

Wykres 1. Zależność między wskaźnikiem rozłączności a współczynnikiem rzetelności i warstwą



Zgodność wewnętrzną skali można sprawdzić dzięki statystykom dopasowania. Gdy wartość statystyk *SIMS* i *SOMS* dla poszczególnych stwierdzeń przekracza ±2 występuje podejrzenie, że dana pozycja skali wprowadza zbyt duże zakłócenia w pomiarze i należy rozważyć jej usunięcie. Zanim jednak taka decyzja zostanie podjęta, zalecane jest sprawdzenie, czy taka sytuacja nie jest wynikiem np. niejednoznacznego sformułowania stwierdzenia, niejasnej instrukcji dołączonej do pytania, a także ponowne oszacowanie parametrów B_n i D_i po usunięciu respondentów, wobec których występuje podejrzenie, że niezbyt uważnie i rzetelnie udzielali odpowiedzi.

Należy podkreślić, że nie ma jednej ustalonej kolejności postępowania. Niektórzy autorzy zalecają, aby w pierwszej kolejności przyjrzeć się poszczególnym

pozycjom skali i zdecydować o ich ewentualnym usunięciu lub pozostawieniu, a potem przeanalizować dane dotyczące respondentów, natomiast inni zalecają równoczesną pracę nad oboma typami wyników.

W przypadku prac nad konstrukcją skali wydaje się uzasadnione usunięcie respondentów wprowadzających zakłócenia do pomiaru. Natomiast jeśli stosuje się skalowanie Rascha do pomiaru badanej cechy, tacy respondenci nie powinni być eliminowani, choć należy w takim przypadku zaznaczyć, że ich pomiar charakteryzuje się ograniczoną wiarygodnością.

Do oceny trafności teoretycznej skali, a więc do sprawdzenia, czy narzędzie mierzy to pojęcie, do którego pomiaru zostało zaprojektowane, również używane są statystyki dopasowania. Gdy dana pozycja nie jest dostatecznie dobrze dopasowana, tzn. statystyki dopasowania wykraczają poza przyjęte normy, istnieje podejrzenie, że wprowadza zbyt wiele zakłóceń do pomiaru badanej cechy lub mierzy inną właściwość, a co za tym idzie – obniża trafność narzędzia.

Do oceny trafności treściowej[5] skali posłużyć mogą oszacowania intensywności poszczególnych stwierdzeń ją tworzących. Dzięki nim możliwe jest uszeregowanie pozycji od najłatwiejszej do najtrudniejszej, przeanalizowanie, czy uzyskane uszeregowanie jest zgodne z założeniami konstruktu, czy poszczególne stwierdzenia nie dublują się lub czy nie występują luki na kontinuum badanego zjawiska, a także czy rozrzut pozycji wzdłuż kontinuum jest wystarczająco duży.

Zbiór oszacowań pozycji respondentów tworzy wyniki tychże osób na badanej skali, tak więc możliwe jest sprawdzenie trafności kryterialnej[6] (zarówno aspektu zbieżnego, jak i rozbieżnego) narzędzia, na przykład poprzez obliczenie odpowiedniej miary korelacji z wybranymi wcześniej kryteriami zewnętrznymi.

Zalety i wady skalowania Rascha

Skalowanie Rascha pozwala na konstrukcję skali o charakterze interwałowym, ale również dostarcza narzędzi do oceny trafności i rzetelności stosowanych w kwestionariuszach skal pomiarowych. Oprócz tego bywa wykorzystywane do sprawdzenia, czy wszyscy badani w ten sam sposób rozumieją poszcze-

[5] Trafność treściowa definiowana jest jako zakres, w jakim pozycje skali mierzącej daną właściwość empiryczną są reprezentatywną próbą zachowań odnoszących się właśnie do tej właściwości – definicja podana za Brzezińskim (2003: 519).

[6] Trafność kryterialna rozumiana jest jako stopień, w jakim wyniki skali są powiązane z zewnętrznym kryterium o ustalonej wcześniej trafności.

gólne pozycje skali, a także poszczególne kategorie odpowiedzi, czy dane narzędzie nie jest dla nich zbyt trudne lub zbyt łatwe, a także czy liczba kategorii odpowiedzi jest optymalna.

Oprócz licznych zalet skalowania Rascha nie można jednak nie wspomnieć, o pewnych wadach czy też niedogodnościach, związanych z tą techniką analizy jakości skali.

Po pierwsze należy pamiętać, że technika analizy danych, o której mowa w tym opracowaniu, zakłada jednowymiarowość mierzonego zjawiska, a także wymaga, aby skala podlegająca analizie składała się ze stwierdzeń o różnej intensywności. Po drugie, jeśli zebrane w wyniku badania dane nie spełniają założeń modelu teoretycznego, proces konstrukcji narzędzia pomiarowego trzeba rozpoczynać od początku. Po trzecie, istnieje możliwość uzyskania takich danych, które będą dobrze pasować do modelu, ale nie będą mieć żadnego sensu merytorycznego, dlatego też konieczna jest dokładna analiza i interpretacja wyników. Po czwarte należy pamiętać, że skalowanie Rascha radzi sobie z brakami danych, ale ich wpływ na uzyskiwane rezultaty nie jest jeszcze do końca znany. Po piąte trzeba mieć świadomość, że dla respondentów, którzy uzyskali maksymalne i minimalne oszacowania, nie można obliczyć statystyk dopasowania. No i wreszcie, brak jest jednoznacznych wytycznych, określających stopień dopasowania danych do modelu, choć w dalszym ciągu trwają prace związane z tym zagadnieniem.

Ilustracja

Do ilustracji wykorzystano dane pochodzące z badania „Badanie poglądów na zagadnienia ludnościowe oraz politykę ludnościową PPA2 (Population Policy Attitudes Survey)" przeprowadzonego w IV kwartale 2001 roku 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". Badanie objęło osoby w wieku 18-64 lat z gospodarstw domowych w mieszkaniach wylosowanych do próby BAEL (Badanie aktywności ekonomicznej ludności). Wywiady przeprowadzono z 4244 respondentami z 2027 gospodarstw domowych. Losowanie próby wykonano z użyciem losowania dwustopniowego. W miastach jednostkami losowania pierwszego stopnia były rejony statystyczne, a na wsi – obwody spisowe. Jednostkami losowania drugiego stopnia były mieszkania. Uzyskana

próba była reprezentatywna ze względu na miejsce zamieszkania, wiek i płeć respondentów (Kotowska i in. 2003).

Skalowanie Rascha zastosowano do oceny skali mierzącej stosunek do osób starszych. Respondentów poproszono o wyrażenie opinii na osiem następujących stwierdzeń:

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

Respondenci mieli do wyboru jedną z pięciu kategorii odpowiedzi, z których każda była etykietowana (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 poprzez zsumowanie odpowiedzi charakteryzowały się bardziej negatywnym stosunkiem w stosunku do osób starszych niż osoby o niższym wyniku ogólnym. W celu ułatwienia interpretacji wyników dokonano przekodowania wyników w sposób przeciwny. W rezultacie wyższy wynik ogólny odpowiadał bardziej pozytywnemu stosunkowi. Jednocześnie, aby to rozumowanie było poprawne, pozostawiono bez zmian kodowanie stwierdzeń *A6*, *A7* i *A8*.

Skalowanie Rascha wykonano za pomocą studenckiej wersji programu WINSTEP, która narzuca ograniczenie liczby analizowanych respondentów do 200, a także wymaga kodowania odpowiedzi począwszy od 0, co zostało uwzględnione i obowiązuje w dalszej części tekstu.

Przed analizą wkładu poszczególnych stwierdzeń sprawdzono, w jakim stopniu dane spełniają założenia modelu. Tabela 3 przedstawia ogólne informacje dotyczące dopasowania, odnoszące się do badanej grupy respondentów.

Średni wynik ogólny uzyskany przez respondentów wyniósł 24,8, podczas gdy wynik minimalny był równy 0, a maksymalny 32. Wszystkie wyniki powstały na bazie zsumowanych odpowiedzi każdej z badanych osób. Przeciętne oszacowanie pozycji respondentów mierzone logitem wyniosło 2,63, co przy uwzględnieniu informacji, że w skalowaniu Rascha przeciętne oszacowanie pozycji skali wynosi 0 (por. tabela 4), pozwala uznać, że przeciętnie biorąc respon-

denci bardzo łatwo zgadzali się z analizowanymi stwierdzeniami. W następnym kroku prowadzi to do konkluzji, że dla tej grupy respondentów skala nie była optymalnie dopasowana pod względem zróżnicowania intensywności stwierdzeń i sugeruje jej rozszerzenie o stwierdzenia, z którymi trudniej się zgodzić, po uprzednim przeprowadzeniu analizy poszczególnych stwierdzeń indywidualnie.

Tabela 3. Ogólne statystyki dla 180 respondentów

<table>
<tr><td rowspan="3"></td><td rowspan="3">WYNIK OGÓLNY</td><td rowspan="3">LICZEBNOŚĆ</td><td rowspan="3">OSZACOWANIE POZYCJI RESPONDENTA</td><td rowspan="3">BŁĄD STANDARDOWY</td><td colspan="4">STATYSTYKI DOPASOWANIA</td></tr>
<tr><td colspan="2">INFIT</td><td colspan="2">OUTFIT</td></tr>
<tr><td>IMS</td><td>SIMS</td><td>OMS</td><td>SOMS</td></tr>
<tr><td>ŚREDNIA</td><td>24,8</td><td>8,0</td><td>2,63</td><td>0,75</td><td>0,95</td><td>-0,4</td><td>0,95</td><td>-0,4</td></tr>
<tr><td colspan="4">Obserwowane: G_p = 1,59 R_p = 0,72</td><td colspan="5">STANDARDOWY BŁĄD ŚREDNIEJ 0,12 WARSTWA 2,453</td></tr>
<tr><td colspan="4">Wynikające z modelu: G_p = 1,90 R_p = 0,78</td><td colspan="5">Minimalny wynik uzyskało 19 respondentów
Braki danych : 1 respondent</td></tr>
<tr><td colspan="5">α CRONBACHA dla danych surowych = 0,8506</td><td colspan="4">średnia interkorelacja między pozycjami = 0,4198</td></tr>
</table>

Przeciętne wartości *IMS* i *OMS* były na poziomie równym bądź bliskim 1, co odpowiada wartości oczekiwanej tych statystyk i świadczy o tym, że grupa respondentów nie powinna być uznana za nietypową. Natomiast statystyki *SIMS* i *SOMS* wyniosły -0,4, co również nie odbiegało znacząco od wartości oczekiwanych tych statystyk, zaś wartości ujemne świadczyły o zbyt małym średnim zróżnicowaniu odpowiedzi respondentów w porównaniu do tego wynikającego z założeń metody.

Współczynnik α Cronbacha, który jest jedną z najczęściej wykorzystywanych miar do oceny rzetelności w badaniach społecznych wyniósł 0,8506 przy średniej interkorelacji między pozycjami na poziomie 0,4198. Górniak (2000) sugeruje, że powinno się dążyć do uzyskiwania skal o rzetelności przynajmniej 0,7. W świetle tych wytycznych analizowana skala okazała się rzetelna na satysfakcjonującym poziomie.

Uzyskany w wyniku porządkowego skalowania Rascha współczynnik rzetelności respondentów R_P był na poziomie 0,72, odpowiadający mu indeks rozłączności G_P wyniósł 1,59, a *Warstwa* 2,53. Zgodnie z zaleceniami Green i Frantom (2000) był on na akceptowalnym poziomie. *Warstwa* na poziomie wyższym od 2 sugerowała możliwość wyodrębnienia za pomocą analizowanej skali przynajmniej dwóch grup respondentów różniących się pod względem stosunku do osób starszych.

Wartość współczynnika rzetelności respondentów R_P po uwzględnieniu liczebności skali złożonej, tj. po przemnożeniu przez 8/7, wyniosła 0,823, co jest wielkością zbliżoną do wartości współczynnika α Cronbacha. Oba wskaźniki nie są identyczne, ponieważ trzeba pamiętać, że choć obie miary zakładają jednowymiarowość badanej skali, to współczynnik α mierzy jej rzetelność na podstawie danych porządkowych, zaś współczynnik rzetelności respondentów – danych interwałowych.

Analogiczna analiza została przeprowadzona dla ośmiu przedstawionych wcześniej pozycji skali – wyniki przedstawia tabela 4. Jako że z założeń metody wynika, iż średnia z oszacowań pozycji skali wynosi 0, a implikacje wynikające z porównania tej wartości ze średnim oszacowaniem pozycji respondentów przedstawiono powyżej, skupiono się na ocenie dopasowania skali traktowanej jak całość.

Tabela 4. Ogólne statystyki dla 8 pozycji skali

	WYNIK OGÓLNY	LICZEBNOŚĆ	OSZACOWANIE INTENSYWNOŚCI	BŁĄD STANDARDOWY	INFIT		OUTFIT	
					IMS	SIMS	OMS	SOMS
ŚREDNIA	558,8	179,8	0,00	0,15	0,99	-0,1	0,95	-0,4
Obserwowane:		G_I = 4,9	R_I = 0,96	STANDARDOWY BŁĄD ŚREDNIEJ			0,62	
Wynikające z modelu:		G_I = 5,69	R_I = 0,97	Warstwa 1,61				

Wartości *IMS* i *OMS* były na poziomie 0,99 i 0,95, a więc bardzo bliskim wartości oczekiwanej tych statystyk, co świadczyło o tym, że skala nie powinna być uznana za nietypową, a w dalszym kroku pozwoliło założyć, że badana grupa stwierdzeń mierzy badany konstrukt, a więc stosunek do osób starszych. Powyższy wniosek potwierdziły również wartości statystyk *SIMS* i *SOMS,* które wyniosły -0,1 i -0,4. W tym przypadku wartości ujemne również świadczą o niewielkim, ale jednak zbyt małym zróżnicowaniu odpowiedzi udzielanych na przeciętną pozycję skali niż to wynikające z modelu. Respondenci zbyt łatwo i zbyt często zgadzali się z badanymi stwierdzeniami.

Indeks rozłączności pozycji G_I wyniósł 4,9[7], zaś *Warstwa* ukształtowała się na poziomie 1,61. Te wartości potwierdziły dobre rozłożenie pozycji skali na kontinuum badanego motywu.

[7] Wyższe wartości indeksu rozłączności dla pozycji niż dla osób są zjawiskiem typowym. Wynika to z faktu, że sposób obliczania indeksu jest uzależniony od liczebności, im większa liczebność, tym wyższa wartość indeksu. A że prawie zawsze jest tak, że liczba badanych osób przewyższa znacznie liczbę pozycji, na które udzielają oni odpowiedzi, to indeks rozłączności pozycji G_I będzie wyższy od indeksu rozłączności respondentów G_P.

Analizy przeprowadzone dla całej grupy badanych respondentów, jak i dla stwierdzeń tworzących skalę dały zadowalające wyniki i pozwoliły w następnym kroku przyjrzeć się indywidualnie poszczególnym pozycjom i poszczególnym respondentom.

Tabela 5 przedstawia pozycje skali uszeregowane według rosnących wartości oszacowań ich intensywności.

Tabela 5. Charakterystyki pozycji skali

POZYCJE SKALI	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 – szacowanie intensywności pozycji skali; B.S. – błąd szacunku

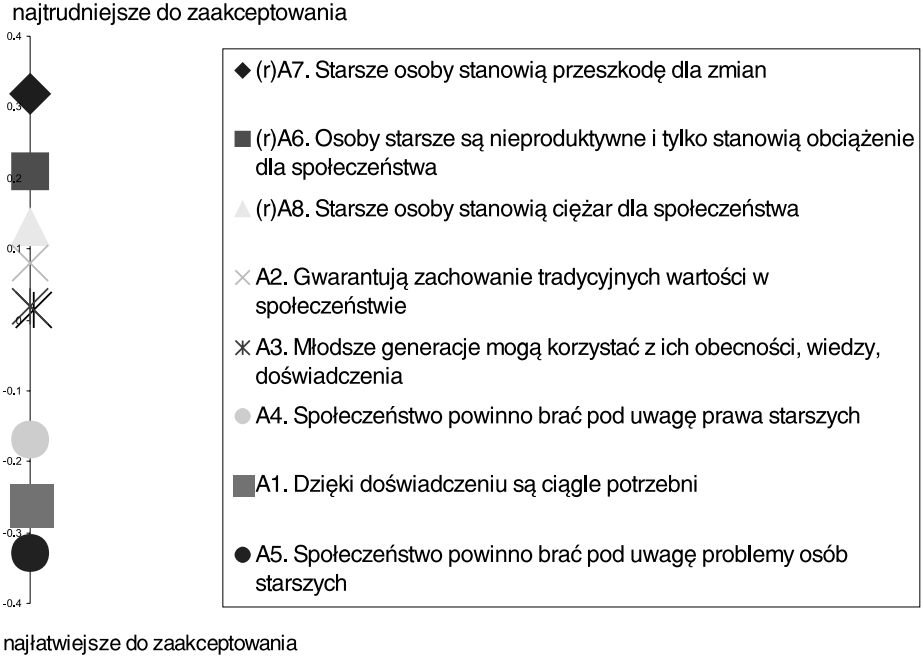
Uszeregowanie badanych stwierdzeń według malejących wartości oszacowań ich intensywności należy rozumieć następująco: respondenci najtrudniej zgadzali się z rozumianym w sposób przeciwny stwierdzeniem *(r)A7. Starsze osoby stanowią przeszkodę dla zmian,* następnie z również o przeciwnym znaczeniu stwierdzeniami *(r)A6. Osoby starsze są nieproduktywne i tylko stanowią obciążenie dla społeczeństwa* i *(r)A8. Starsze osoby stanowią ciężar dla społeczeństwa.* Najłatwiejsze do zaakceptowania były stwierdzenia *A1. Dzięki do-*

świadczeniu są ciągle potrzebni i *A5. Społeczeństwo powinno brać pod uwagę problemy osób starszych.*

Każdemu z oszacowań intensywności pozycji skali odpowiadał standardowy błąd szacunku również mierzony logitem, który dla wszystkich badanych stwierdzeń wyniósł 0,15.

Żadnego z badanych stwierdzeń nie można było uznać za niepasujące do skali, co pokazały statystyki *IMS* i *OMS* zawierające się w przedziałach odpowiednio <0,78; 1,22> i <0,78; 1,14>. Powyższy wniosek okazał się poprawny również w przypadku statystyk *SIMS* i *SOMS*, z których żadna nie przekroczyła zalecanego poziomu <-2, 2>. Pozwoliło to uznać, że skala mierząca stosunek do osób starszych charakteryzowała się trafnością teoretyczną.

Wykres 2. Mapa pozycji



W analogiczny sposób, jak ten przedstawiony powyżej, można dokonać analizy odpowiedzi poszczególnych respondentów. Tabela 6 zawiera charakterystyki wybranych respondentów. Oszacowania pozycji respondentów zawierały się w przedziale <-1,04; 7,53>, natomiast błędy standardowe tych szacunków wynosiły od 0,42 do 1,86. Najbardziej chętnych do akceptowania stwierdzeń było 19

respondentów, którzy uzyskali maksymalny wynik ogólny równy 32, wartość logitu dla ich pozycji, czyli ich poziom akceptowalności stwierdzeń tworzących skalę, został oszacowany na poziomie 7,53. Dla tych „ekstremalnych" 19 respondentów nie są obliczane statystyki dopasowania.

Natomiast najczęściej i najbardziej kategorycznie nie zgadzali się ze stwierdzeniami dwaj respondenci, których poziom braku akceptowalności był największy i oszacowany na poziomie -1,04 logita (wynik ogólny 14). Trzeba zauważyć, że w badanej grupie respondentów nie było żadnego, który uzyskałby minimalny wynik ogólny na poziomie 0 (0 x 8).

Dla 15 respondentów statystyki *SIMS* i *SOMS* przekroczyły dopuszczalny poziom 2, zaś dla 48 były niższe od -2. Istniało zatem podejrzenie, że respondenci o *SIMS* przekraczającej wartość 2 udzielali odpowiedzi w sposób zbyt losowy, nieprzewidywalny. Sugeruje się w takim przypadku przeanalizowanie odpowiedzi, których udzielili, jak również sprawdzenie warunków, w jakich odbywało się badanie i w końcu ewentualne usunięcie takich respondentów z analizy. W przypadku podjęcia najbardziej radykalnego kroku, czyli usunięcia respondenta z analizy, należy ponownie oszacować wszystkie parametry modelu.

Tabela 6. Charakterystyki wybranych respondentów

WYNIK OGÓLNY	OSZACOWANIE POZYCJI RESPONDENTA	BŁAD STANDARDOWY	STATYSTYKI DOPASOWANIA			
			INFIT		OUTFIT	
			IMS	SIMS	OMS	SOMS
14	-1,04	0,42	0,34	-2,1	0,34	-2,1
16	-0,69	0,43	0,03	-4,7	0,03	-4,7
17	-0,50	0,44	0,97	0,1	1,01	0,2
18	-0,30	0,46	1,29	0,7	1,32	0,8
19	-0,07	0,49	0,35	-1,6	0,37	-1,5
20	0,18	0,52	1,51	1	1,41	0,8
21	0,45	0,58	0,48	-0,9	0,45	-0,9
22	0,86	0,65	3,25	2,3	3,21	2,3
23	1,35	0,75	7,29	3,8	7,54	3,8
24	1,99	0,83	0,01	-2,5	0,01	-2,5
25	2,68	0,81	0,64	-0,4	0,66	-0,3
26	3,28	0,75	0,72	-0,5	0,69	-0,6
27	3,81	0,71	1,88	2,5	1,91	2,5
28	4,29	0,70	0,95	-0,2	0,95	-0,2
29	4,80	0,73	1,14	0,6	1,18	0,7
30	5,38	0,82	0,95	0	0,91	0
31	6,23	1,07	1,08	0,4	1,36	0,7
32	7,53	1,86	-	-	-	-
ŚREDNIA	3,1	0,85	0,95	-0,4	0,95	-0,4
ODCH. STD.	2,11	0,34	1,08	1,9	1,10	1,9

Wśród 48 respondentów, dla których *SIMS* był niższy niż -2, 51,1% stanowili mężczyźni, zaś 48,9% kobiety, natomiast ich strukturę wieku przedstawia tabela 7.

Podobne kroki należy podjąć w przypadku osób, dla których statystyka *SIMS* była mniejsza od -2, ponieważ odpowiedzi takich respondentów były zbyt mało zróżnicowane i zbyt mało niezależne od siebie. Sposób odpowiadania tych respondentów był zdeterminowany przez jakiś czynnik zewnętrzny. Być może był to wynik tematyki badania, która miała bardzo delikatny charakter. Można podejrzewać, że ci respondenci udzielali odpowiedzi społecznie pożądanych i oczekiwanych, a nie zgodnych z ich własnymi odczuciami.

Tabela 7. Struktura wieku respondentów nietypowych

GRUPA WIEKOWA	GRUPA WSZYSTKICH 200 BADANYCH RESPONDENTÓW	GRUPA 48 RESPONDENTÓW (SIMS < -2)
Do 20 lat	6,5%	9,5%
20 – 30	28,5%	21,8%
30 – 40	17,5%	25,6%
40 – 50	27,5%	30,8%
50 – 60	15,0%	10,5%
60 lat i więcej	5,0%	1,8%

Najbardziej charakterystyczną cechą skalowania Rascha jest możliwość porównywania między sobą nie tylko respondentów, czy też pozycji, ale dokonanie porównania globalnego, czyli zarówno pozycji skali, jak i respondentów jednocześnie. Mapa pozycji i respondentów, której przykład znajduje się na wykresie 3, jest bardzo dobrym przykładem prezentacji tych wyników.

Mapa pozycji i respondentów przedstawia rozkład pozycji respondentów – po lewej stronie pionowej przerywanej linii – i rozkład oszacowań pozycji skali (z uwzględnieniem progów) – po stronie prawej – względem siebie. Każdy symbol „#" oznacza 4 respondentów. Symbol „M" oznacza średnią, „S" odchylenie standardowe, zaś „T" symbolizuje dwa odchylenia standardowe. Im wyżej wzdłuż pionowej linii znaleźli się respondenci, tym chętniej zgadzali się ze stwierdzeniami i silniej się zgadzali (większa intensywność).

[8] Liczba subpozycji jest zawsze równa liczbie progów, zaś 5 kategoriom odpowiedzi odpowiadają 4 progi.

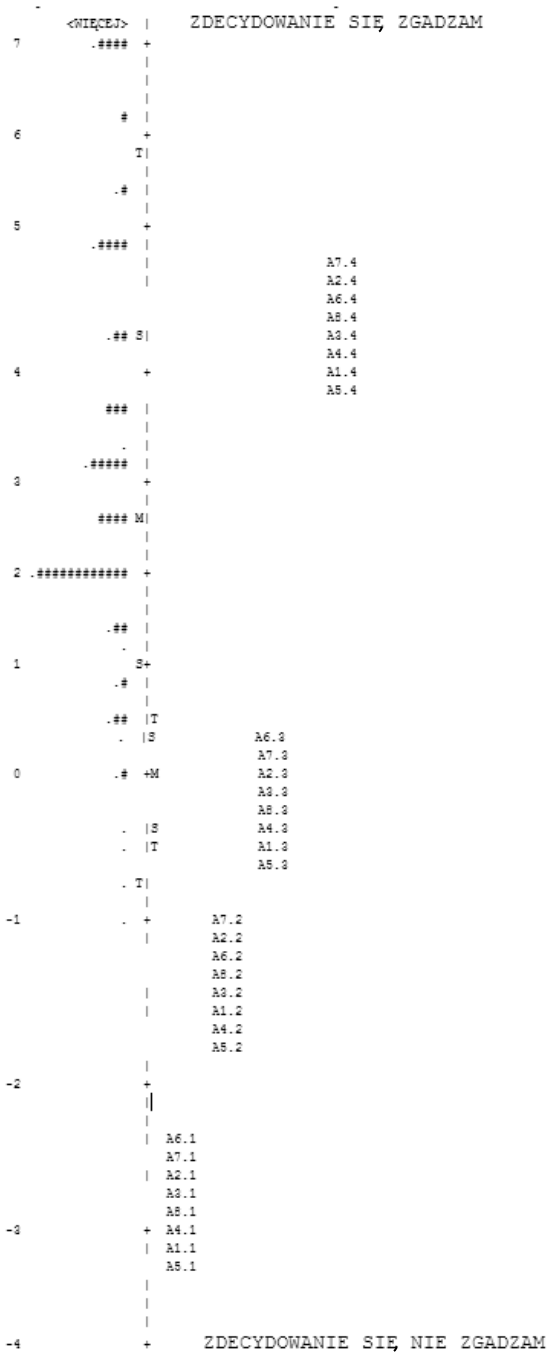
Konstrukcja tego wykresu opierała się na przyjęciu, że każde z ośmiu interesujących nas stwierdzeń składa się z 4 subpozycji[8] o dychotomicznej skali odpowiedzi: 1 – jeżeli respondent wybierze wyższą z dwóch sąsiadujących kategorii odpowiedzi i 0 – jeśli wybierze niższą. Próg pierwszy odpowiadał na wykresie symbolowi „.1" poprzedzonemu symbolem stwierdzenia np. A1, tak więc symbol A1.1 oznaczał, że respondenci usytuowani na wykresie na tym samym poziomie (mający tę samą współrzędną na osi rzędnych), co pierwszy próg dla stwierdzenia A1, będą mieli dokładnie 50% szans wybrania odpowiedzi „nie zgadzam się" wobec odpowiedzi „zdecydowanie się nie zgadzam". Dla wszystkich respondentów mających wyższe oszacowanie pozycji niż oszacowanie akceptowalności A1.1, to prawdopodobieństwo będzie większe od 0,5. Podobne rozumowanie polegające na porównaniu ze sobą usytuowania respondentów i kategorii odpowiedzi na poszczególne pozycje skali przeprowadza się dla pozostałych przypadków.

W przypadku respondentów możemy mówić, że rozkład ich pozycji wyrażonych logitem charakteryzowała niewielka dodatnia asymetria (współczynnik asymetrii na poziomie 0,535), zaś rozstęp mierzony logitem wynosił 8,57. Rozstęp charakteryzujący oszacowania pozycji skali wyniósł około 7,5 logita. Jednak oba rozkłady nie leżały równolegle względem siebie – rozkład pozycji respondentów był zdecydowanie wyżej usytuowany niż rozkład oszacowań pozycji skali, to samo dotyczy średnich wartości oszacowań – co świadczyło o tym, że badana skala nie była dobrze dostosowana do badanej grupy respondentów. Pozycje skali były zbyt łatwe do zaakceptowania.

Z wykresu 3 wynika, że 40 respondentów miało tak bardzo pozytywny stosunek do starszych, że zbudowana skala nie była w stanie dokładnie go zmierzyć, była zbyt słaba. Dodatkowo kategorie odpowiedzi „zdecydowanie się nie zgadzam" i „nie zgadzam się" dla wszystkich stwierdzeń leżały poniżej oszacowań akceptowalności dla większości respondentów, co znaczyło, że wnosiły one wartość poznawczą do zmierzenia badanej potrzeby tylko dla znikomej liczby respondentów. Zwraca również uwagę duża luka między oszacowaniami dla 3 i 4 progu. Sugerować ona może potrzebę większego zróżnicowania intensywności kategorii odpowiedzi odpowiadających pozytywnemu stosunkowi do starszych. Być może zamiast dwóch pozytywnych kategorii odpowiedzi powinny być trzy.

W celu wyrównania obu rozkładów względem siebie można zrezygnować z bardzo ogólnego badania stosunku wobec osób starszych na rzecz badania intensywności stosunku pozytywnego lub dodać więcej stwierdzeń trudniejszych do zaakceptowania.

Wykres 3. Mapa pozycji i respondentów



Źródło: Wydruk programu Winstep Ministep

Kwestią niezwykle istotną zwłaszcza z punktu widzenia trafności treściowej jest zastanowienie się, czy kolejność, w jakiej ustawione są według oszacowań stopnia akceptowalności pozycje skali, ma sens. Czyli, czy na przykład łatwiej jest zgodzić się ze stwierdzeniem *A2. Gwarantują zachowanie tradycyjnych wartości w społeczeństwie* niż *A1. Dzięki doświadczeniu są ciągle potrzebni*. A także czy osoba charakteryzująca się pozytywnym stosunkiem do osób starszych nie zgodzi się ze stwierdzeniem *(r)A7. Starsze osoby stanowią przeszkodę dla zmian,* podczas gdy osoba o bardziej negatywnej postawie nie zgodzi się ze stwierdzeniem *A5. Społeczeństwo powinno brać pod uwagę problemy osób starszych.* Krótka analiza treści stwierdzeń i wyników porządkowego skalowania Rascha pozwala na powyższe pytania odpowiedzieć twierdząco.

Analiza skali oceniającej stosunek do osób starszych wykonana za pomocą porządkowego skalowania Rascha dała satysfakcjonujące wyniki. Sposób uszeregowania pozycji według ich stopnia intensywności pozwolił uznać trafność treściową skali za odpowiednią. Natomiast o dobrej trafności teoretycznej narzędzia świadczyły statystyki dopasowania *SIMS, SOMS, IMS* i *OMS*, które w żadnym przypadku nie przekroczyły wartości uznanych za graniczne.

Rzetelność skali mierzona współczynnikiem rzetelności R_P i indeksem rozłączności G_P była na satysfakcjonującym poziomie, chociaż w przypadku tej drugiej wyniki uzyskane za pomocą skalowania porządkowego nie były takie same, jak w przypadku klasycznego miernika, za jaki uważa się współczynnik α Cronbacha (pomimo liczbowo zbliżonych wartości). Do pełnego obrazu zabrakło jednak weryfikacji skali kategorii odpowiedzi, ale ta analiza wykracza poza zakres tego opracowania.

Analizowanej skali zarzucić można jednak, że obejmuje zbyt wąski zakres badanej właściwości, a więc stosunku do osób starszych, a także że składa się ze stwierdzeń, których zawartość treściowa narzuca respondentom określony typ odpowiedzi – odpowiedzi pożądanej i poprawnej społecznie. W celu ulepszenia narzędzia należałoby dodać stwierdzenia o wyższej intensywności, a więc z którymi trudniej będzie się zgadzać.

Uwagi końcowe

Opracowanie powstało na bazie wątpliwości, jakie wzbudza powszechne przyjmowanie założenia, że można traktować dane mierzone na skalach porządkowych o przynajmniej pięciu kategoriach odpowiedzi jako mające charakter interwałowy, sztucznie zakładając równe odległości między kategoriami. Takie

podejście ma zarówno swoich zwolenników, jak i przeciwników, popierających swoje poglądy licznymi badaniami (por. Górniak 2000: 312–313), natomiast celem tego opracowania było przedstawienie metody tworzenia jednowymiarowej skali złożonej bez przyjmowania sztucznych założeń.

Na podstawie powyższego opracowania można stwierdzić, że skalowanie porządkowe nie jest złotym środkiem, będącym w stanie rozwiązać wszelkie problemy związane z tworzeniem i oceną jakości skali. Niemniej jednak może pomóc w lepszym przyjrzeniu się temu zagadnieniu.

Dalszych prac oraz dodatkowego komentarza wymagają następujące problemy:

1. Zaproponowanie polskiego nazewnictwa. Ta praca stanowi dopiero skromny początek. Być może w miarę upowszechniania się tej metodologii w Polsce uda się wypracować wspólne i bardziej adekwatne określenia.
2. Rozległość tematyki. Skalowanie Rascha obejmuje wiele technik, zaś to opracowanie ograniczono tylko do skalowania porządkowego i wykorzystania go do oceny jakości skali.
3. Obszar zastosowań. Skalowanie Rascha było i jest stosowane głównie w edukacji do oceny trudności testów umiejętności i ich porównywalności, w sporcie, np. do porównania uczciwości sędziów oceniających zawody w łyżwiarstwie figurowym, w naukach medycznych do diagnostyki chorób. Zastosowanie skalowania porządkowego do pomiaru postaw lub motywów zachowań nie jest jeszcze szeroko opisane w literaturze.
4. Ocena dopasowania. Brak jednoznacznych wytycznych dotyczących oceny dopasowania i związany z tym brak ścisłych wskazówek, pozwalających określić trafność narzędzia.

Literatura

Bond, Trevor G. i Christine M. Fox. 2001. *Applying The Rasch Model. Fundamental Measurement in the Human Sciences*. New Jersey: Lawrence Erlbaum Associates, Publishers.

Brzeziński, Jerzy. 2003. *Metodologia badań psychologicznych*. Warszawa: Wydawnictwo PWN.

Górniak, Jarosław. 2000. *My i nasze pieniądze*. Kraków: Aureus.

Green, Kathy E. 1996. *Applications of the Rasch Model to Evaluation of Survey Data Quality*. *„New Directions for Evaluation"* 70: 81–91.

Green, Kathy E. i Catherine G. Frantom. 2002. *Survey development and validation with the Rasch Model.* A paper presented at the International Conference on Questionnaire Development, Evaluation, and Testing. http://www.jpsm.umd.edu/qdet/final_pdf_papers/green.pdf.

Kotowska, Irena E. i in. 2003. *Polityka ludnościowa – cele, rozwiązania, opinie.* Raport z badania finansowanego przez KBN, 5H02B 020 20. Warszawa.

Linacre, John M. 2004. *Estimation methods for Rasch measures.* W: E. V. Smith i R. M. Smith (red.), *Introduction to Rasch Measurement. Theory, Models and Application.* Maple Grove, Minesota: Jam Press, s. 25–71.

Linacre, John M. [b.d.] *Guidelines for Rating Scales*. MESA Research Note nr 2. http://www.rasch.org/rn2.htm.

Linacre, John M. 2002. *What do Infit and Outfit, Mean-square and Standardized mean*? *„Rasch Measurement Transactions", 16:2*. http://www.rasch.org/rmt/rmt162f.htm.

Myforde, Carol M. i Edward W. Wolfe. 2004. *Detecting and measuring rater effects using Many-Facet Rasch Measurement:* Part I. W: E. V. Smith i R. M. Smith (red.), *Introduction to Rasch Measurement. Theory, Models and Application.* Maple Grove, Minesota: Jam Press, s. 460–517.

Stone, Mark H. 2004. *Substantive Scale Construction*. W: E. V. Smith i R. M. Smith (red.), *Introduction to Rasch Measurement. Theory, Models and Application.* Maple Grove, Minesota: Jam Press, s. 201–225.

Wright, Ben D. *Fundamental measurement for psychology*. MESA Memo 64. .

Smith, Everett V. i Richard M. Smith (red.). 2004. *Introduction to Rasch Measurement. Theory, Models and Applications*. Maple Grove, Minesota: Jam Press.

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 wymodelowaną *n*-tego respondenta na *i*-tą pozycję skali, *N* – liczba osób;

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

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

gdzie C_{ni} jest kurtozą X_{ni}.

2. *SOMS – Standardized Weighted Mean Square Outfit* – to przekształcony według transformacji Wilsona i Hilferty'ego *OMS*.

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

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 wymodelowaną *n* – tego respondenta na *i*-tą pozycję skali, W_{ni} – wariancja różnic miedzy odpowiedzią oczekiwaną a obserwowaną;

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

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

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.

$$SIMS_i = \left(IMS_i^{1/3} - 1\right)\left(3/q_i\right) + q_i/3$$

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. Współczynnik rzetelności pozycji R_I definiuje wzór:

$$R_I = \frac{SA_I^2}{SD_I^2}$$

gdzie: SD^2_I – całkowita wariancja oszacowań intensywności pozycji, SA_I^2 – wariancja prawdziwych oszacowań intensywności pozycji: $SA_I^2 = SD_I^2 - MSE_I$, a średniokwadratowy błąd szacunku MSE_I obliczany jest zgodnie z formułą:

$$MSE_I = \frac{\sum_{i=1}^{I} S_i^2}{I}$$

gdzie S_i jest standardowym błędem szacunku dla każdego z oszacowań intensywności pozycji skali.

6. Indeks rozłączności pozycji

$$G_I = \frac{SA_I}{SE_I} = \sqrt{\frac{R_I}{1 - R_I}}$$

gdzie: SA_I – odchylenie standardowe prawdziwych oszacowań intensywności pozycji, SE_I – przeciętny błąd pomiaru (SE_I jest pierwiastkiem kwadratowym z MSE_I).

7. Warstwa

$$Warstwa = \frac{4G_I + 1}{3}$$

Współczynnik rzetelności *R*, indeks rozłączności *G* i *Warstwa* mogą być obliczane również dla osób. Operacja przebiega w ten sam sposób przy jednoczesnej zamianie subskryptów dla pozycji (*I*) subskryptami dla osób (*P*) w formułach z punktów 5, 6 i 7.

ASSESSMENT OF THE QUALITY OF THE SCALE WITH THE RASCH MODEL: SCALE RELIABILITY AND VALIDITY

The aim of the article is to present Rasch model and its application to the assessment of scale validity and reliability. The method is hardly known in Poland. The problem of good quality data is important as such a data is a prerequisite of reliable results of each statistical analysis. The Rasch model provides a different approach to the assessment of the scale quality from that having been used in Poland and described among others by David Magnusson, Jerzy Brzezinski, Jaroslaw Gorniak, Andrzej Machowski, Adam Sagan. The text consists of two parts – theoretical one and practical one. The former includes methodology of the Rasch measurement whereas the latter presents its application. Although there is a big family of the Rasch measurement models this text focuses mainly on the rating scale model. This method is used to analyze rating scale data obtained from the battery of items with different intensities designed to measure one latent trait.

Key words: Rasch model, Rasch measurement, rating scale model, rating scale, scale reliability, scale validity.