Maciej Górkiewicz

Kognitywistyczne spojrzenie na skale psychometryczne: kiedy trudne jest łatwiejsze od łatwego...

Wprowadzenie

W ramach klasycznej teorii IRT (*Item Response Theory*) pomiar jest traktowany w zasadzie tak jak pomiar fizykalny. Niemniej w pomiarze kwestionariuszowym rola człowieka nie ogranicza się jedynie do dokonania odczytu i zapisu pośredniego wyniku wytworzonego przez fizykalny przyrząd pomiarowy. Wynik pośredni, a więc nie tylko ostateczny zapis wyniku, kształtuje człowiek w toku różnorodnych procesów psychicznych, począwszy od wytworzenia w sobie określonego wyobrażenia o pomiarze, nastawienia do proponowanego mu badania, zrozumienia treści itemów i instrukcji wypełniania kwestionariusza oraz formułowania odpowiedzi na pytania kwestionariusza. W związku z tym wiele procedur projektowania kwestionariusza, a następnie wykorzystywania go w praktyce, mniej lub bardziej otwarcie odwołuje się do psychologii człowieka, w tym także do teorii kognitywistycznych, w zasadzie nie wychodząc jednak poza ramy wyobrażeń tworzących podłoże dla procedur IRT. Jest wiele dowodów [np. Collins 2003; Schmidt, Le, Ilies 2003; Bjorner, Ware, Kosinski 2003], że taki sposób włączenia podejścia kognitywistycznego do badań psychometrycznych ma ogólnie bardzo pozytywne konsekwencje praktyczne. W niniejszej pracy uwaga jest jednak skoncentrowana tylko na tym, jak podejście kognistywistyczne podaje w wątpliwość, jak osłabia podstawowe założenie podejścia typu Rasch [Fischer, Molenaar 1995].

Dla odpowiednio zdefiniowanej populacji da się tak dobrać itemy skali, że zaistnieje wystarczająco pełny *consensus* co do liniowego uporządkowania wszystkich itemów wg aspektu mierzonego za pomocą tej skali. W rozważaniach zostaną wykorzystane dwa przykłady. Pierwszy, być może niezbyt poważny, to powszechnie znane zasady dziecięcej gry w kamień – nożyce – papier – studnię. Drugi to uporządkowanie preferencyjne itemów skali funkcjonowania fizycznego (PF, od ang.: *physical functioning*), wchodzącej w skład powszechnie stosowanego kwestionariusza badania zdrowotnych aspektów jakości życia, SF-36 [Martin, Kosinski, Bjorner, Ware, MacLean, Li 2007]. Pierwszy przykład dotyczy codziennych sytuacji, w których rozsądni ludzie o ustalonych binarnych preferencjach aprobują brak przechodniości tych preferencji. Drugi przykład jest bardziej złożony. Twórcy kwestionariusza SF-36 nie negują pożytków stosowania IRT, w tym także podejścia stylu Rasch: przeciwnie, swój pozytywny sto-

sunek zaznaczają już w tytule pracy [Martin i in. 2007]. Niemniej w treści owej pracy pokazują, że różne procedury typu Rasch, zastosowane do dokładnie tych samych danych, prowadzą do odmiennych uporządkowań liniowych. Mogą sobie na to pozwolić, ponieważ w innej pracy Ware, Kosinski i Dewey [2000], cytują tysiące publikacji (prowadzących pośrednio do dziesiątków tysięcy) potwierdzających pomyślne zastosowanie wyników pomiaru za pomocą SF-36. Ponadto wiadomo, że pozornie nieznaczne modyfikacje odpowiedzi uzyskanych od tych samych respondentów mogą być źródłem poważnych trudności obliczeniowych [Gosh 1995].

Wydaje się, że pozostając ściśle w ramach ujęcia IRT, a ogólnie na gruncie nieczłowieczych klasycznych logik, czyli idealizowanego zdrowego rozsądku, jesteśmy bezradni w obliczu obserwowanych odstępstw od liniowych uporządkowań preferencji. W zasadzie pozostaje nam tylko badanie, za pomocą mniej lub bardziej wyrafinowanych metod, czy i na ile odpowiedzi są losowe [Slovic 1995; Bereby-Meyer, Meyer, Flasher 2002]. Ujęcie kognitywistyczne dopuszcza odstępstwa od liniowych uporządkowań preferencji, traktuje je jako coś zwyczajnego, niekoniecznie w sensie: *errare humanum est*, niekoniecznie jako objaw niewłaściwego nastawienia respondenta do samego badania lub do celu badań. W szczególności każdy item może wywoływać u respondenta wspomnienia innych sytuacji [Szałek 2004; Chan, McDermott 2006], wyobrażenie innego kontekstu, a tym samym – powodować zmianę aspektu i kryteriów oceny [Neisser 1994].

W niniejszej pracy układy preferencji są modelowane za pomocą uporządkowań nieliniowych. Model uporządkowania cyklicznego („błędnego koła") odwołuje się do pojęcia bliskich, bezpośrednich skojarzeń. Zgodnie z tym modelem respondent, szukając odpowiedzi, bierze pod uwagę tylko jeden lokalny fragment uporządkowania swoich preferencji. Znając model i punkt wyjściowy, możemy przewidywać decyzje respondenta. Model uporządkowań częściowo liniowych (ram z rozgałęzieniami) pokazuje tylko, które rozgałęzienia (spośród wszystkich możliwych) są w praktyce wybierane przez decydenta, nie dając wskazówek, dlaczego właśnie te rozgałęzienia są preferowane, a inne nie, a także które rozgałęzienie zostanie wybrane w określonej sytuacji [Clemen 2001].

Gra w kamień – nożyce – papier – studnię

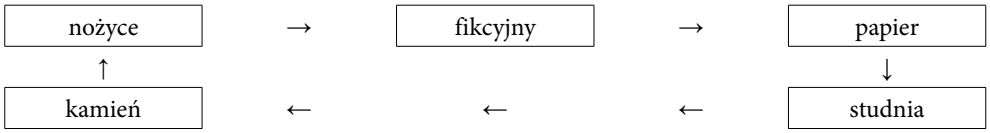
Jest to gra tylko dla dwóch osób. W każdej rundzie uczestnicy chowają ręce za siebie, układają palce w wyobrażenie jednej z czterech figur: kamienia, nożyc, papieru lub studni, i na sygnał pokazują sobie, co który wybrał. Dla każdej możliwej pary niejednakowych figur istnieje ustalona z góry stała relacja słabszy–silniejszy:

Tablela 1. Relacje binarne słabszy–silniejszy

Słabszy	Silniejszy	Wyjaśnienie: dlaczego?
kamień	papier	Kamień zniknie, bo go papier owinie.

kamień	studnia	Kamień zniknie, bo go studnia pochłonie.
papier	nożyce	Papier zniknie, bo go nożyce pokroją.
studnia	papier	Studnia zniknie, bo ją papier nakryje.
nożyce	kamień	Nożyce znikną, bo je kamień rozbije.
nożyce	studnia	Nożyce znikną, bo je studnia pochłonie.

Po dodaniu jednej figury fikcyjnej powyższy układ relacji można przedstawić równoważnie za pomocą jednego uporządkowania cyklicznego, przy założeniu, że osoba, która wybrała określoną figurę, w dalszych porównaniach bierze pod uwagę zawsze dwie najbliższe relacje z lewej i dwie najbliższe relacje z prawej strony:



Rycina 1. Uprządkowanie cykliczne według relacji: silniejszy od

Jak można łatwo zauważyć, przy takim podejściu, niezależnie od dokonanego wyboru figury, zawsze otrzymuje się uporządkowanie liniowe z wybraną figurą pośrodku:

studnia	→	kamień	→	nożyce	→	fikcyjny	→	papier;
kamień	→	nożyce	→	fikcyjny	→	papier	→	studnia;
nożyce	→	fikcyjny	→	papier	→	studnia	→	kamień;
fikcyjny	→	papier	→	studnia	→	kamień	→	nożyce;
papier	→	studnia	→	kamień	→	nożyce	→	fikcyjny.

Dodanie elementu fikcyjnego powoduje, że przy każdym wyborze nie mamy dylematu, czy drugi z kolei element leży po prawej, czy po lewej stronie: zawsze są dokładnie dwa po lewej i dwa po prawej. Poza tym, co nas nie interesuje, dodanie elementu fikcyjnego czyni grę bardziej sprawiedliwą (każdy wybór wygrywa z dwoma wyborami przeciwnika i przegrywa z dwoma innymi wyborami), ale chyba mniej ciekawą z psychologicznego punktu widzenia.

Skala funkcjonowania fizycznego (PF) w kwestionariuszu SF-36

Skala oceny funkcjonowania fizycznego (PF) w kwestionariuszu SF-36 zawiera krótkie opisy 10 czynności. Respondent ma po kolei oceniać każdą czynność oddzielnie, za każdym razem odpowiadając sobie na pytanie: czy w związku z moim obecnym

stanem zdrowia wykonanie tej czynności jest dla mnie trudne (ocena: 1), łatwe (ocena: 3), pośrednio trudne/łatwe (ocena: 2). Wynikową ocenę oblicza się jako sumę wszystkich dziesięciu ocen, a następnie, co nas nie interesuje, odpowiednio standaryzuje się [Ware i in. 2000].

Tablela 2. SF-36: itemy oceny funkcjonowania fizycznego PF (*physical functioning*)

Typ	Item	Opis
Wysiłek	PF_a	wysiłek intensywny, bieganie, podnoszenie ciężarów, ...
	PF_b	wysiłek umiarkowany, odkurzanie, przesuwanie stołu, golf, ...
Schody	PF_d	wejście na kilka pięter
	PF_e	wejście na jedno piętro
Chodzenie	PF_g	przejście ponad jednego kilometra
	PF_h	przejście kilkuset metrów
	PF_i	przejście 100 metrów
Inne	PF_f	schylanie się, skłony
	PF_c	noszenie zakupów spożywczych
	PF_j	mycie i ubieranie się (samodzielne)
* W oryginalnym kwestionariuszu itemy uporządkowane alfabetycznie, od PF_a do PF_j [Martin, Kosinski, Bjorner, Ware, MacLean, Li 2007].		

W tabeli 2 itemy skali pogrupowano w dwie pary itemów, wysiłek i schody, oraz w dwie trójki, chodzenie i inne. Co do obu par itemów oraz trójki „chodzenie", milcząco zakłada się, że są one tak dobrze dobrane i opisane, że uporządkowanie ich wg stopnia trudności można uznać za obiektywne i oczywiste w tym sensie, iż zadeklarowanie przez respondenta odmiennego uporządkowania świadczy o braku dobrej woli lub pomyłce, co może wynikać z niewłaściwego nastawienia lub zaburzeń funkcji poznawczych respondenta. Istnienie i postać właściwego, powszechnie uznawanego co najmniej w pojedynczej badanej populacji, uporządkowania wszystkich 10 itemów skali pozostaje natomiast sprawą otwartą. Marle Martin i in. [2007] badali uporządkowanie liniowe 10 itemów skali PF wg stopnia ich trudności za pomocą znanej procedury Mastersa [1982], opartej na estymowaniu standaryzowanych progów decyzyjnych wyboru pomiędzy oceną 1 a 2, co raczej dotyczy respondentów o niskiej globalnej ocenie PF, a następnie wyboru pomiędzy oceną 2 a 3, co raczej dotyczy respondentów o niskiej globalnej ocenie PF.

<table>
<tr><td></td><td></td><td colspan="3">Gałąź progu 2 v. 3</td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td>PF_g</td><td>PF_d</td><td>PF_b</td><td></td><td></td><td></td><td></td></tr>
<tr><td>trudne</td><td>PF_a</td><td></td><td></td><td></td><td>PF_h</td><td>PF_e</td><td>PF_i</td><td>łatwe</td></tr>
<tr><td></td><td></td><td>PF_b</td><td>PF_d</td><td>PF_g</td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td colspan="3">Gałąź progu 1 v. 2</td><td></td><td></td><td></td><td></td></tr>
</table>

Rycina 2. Uporządkowanie liniowe itemów skali PF

Okazało się, że każdy próg decyzyjny, 1 v. 2 oraz 2 v. 3, nieco odmiennie porządkuje 10 itemów PF, mimo że obliczenia wykonano na podstawie dokładnie tych samych danych, przy zastosowaniu tego samego oprogramowania. Uzyskane uporządkowania tylko trzech „oczywistych" grup itemów (wysiłek, schody, chodzenie) da się przedstawić w postaci jednego modelu z rozgałęzieniem. Jak łatwo zauważyć, tego modelu nie można przedstawić w postaci uporządkowania cyklicznego, np. patrząc „z punktu widzenia" czynności PF_d. W najbliższym otoczeniu występują przeciwne uporządkowania w gałęzi progu 2 v. 3 i w gałęzi 1 v. 2. Wydaje się natomiast, że znając ogólną ocenę PF określonej osoby, można wnioskować o jej indywidualnej liniowej skali preferencji. Nic bardziej mylnego! Przecież skala PF zawiera 10 itemów (czy w rozpatrywanym przykładzie 7 itemów), a zastosowana skala ocen ma tylko 3 stopnie: 1, 2 lub 3. Siłą rzeczy kilka itemów musi uzyskać tę samą ocenę. Bardzo sprawny respondent wszystkie czynności PF_g, h, i uzna za łatwe (ranga = 3) dla siebie, a mało sprawny przeciwnie – za trudne (ranga = 1). W rezultacie przyjmuje się powszechnie, że respondent, który uznaje za oczywiste np. uporządkowanie łatwości itemów PF_i < PF_h < PF_g z tabeli 2, w trakcie wypełniania kwestionariusza SF-36 może tę samą skrajną rangę przypisać dwóm, a nawet trzem itemom. Przypisanie wszystkim itemom skali rangi pośredniej może być uznane za niesprzeczne z dowolnym uporządkowaniem tych itemów, pod warunkiem że dopuszcza się stosowanie przez respondenta punktów odniesienia spoza skali, np. trudne (ranga = 1) jest dla mnie wejście na Mont Everest, łatwe (ranga = 3) leżenie na kanapie, a te wszystkie czynności skali mają rangę równą 2. Mało tego, jak łatwo zauważyć, respondent o średniej samoocenie swojej sprawności wynik dla siedmiu itemów PF = 14 może uzyskać na 4 różne sposoby: same 2; po jednej 1 i 3, po dwie, po trzy. Kłopoty wzięły się stąd, że zwolennicy konsensusu co do porządku preferencyjnego tak długo zmniejszają liczbę stopni skali ocen, aż prawie wszystkie obserwowane odstępstwa zostają zamazane [Tennant, Penta, Tesio, Grimby, Thonnard, Slade, Lawton, Simone, Carter, Lundgren -Nilsson, Tripolski, Ring, Biering-Sorensen, Marincek, Burger, Phillips 2004]. Przecież temu, kto K itemom wystawił K identycznych ocen, można przypisać każde z K! (silnia) uporządkowań tych itemów... Może warto jeszcze raz popatrzeć na problem, z innej perspektywy?

BIBLIOGRAFIA

Bereby-Meyer Y., Meyer J., Flasher O.M. (2002). *Prospect Theory Analysis of Guessing in Multiple Choice Tests*. „Journal of Behavioral Decision Making" 15(4), s. 313–327.

Bjorner J., Ware J., Kosinski M. (2003). *The Potential Synergy Between Cognitive Models and Modern Psychometric Models*. „Quality of Life Research" 12, s. 261–274.

Chan J.C.K., McDermott K.B. (2006). *Remembering Pragmatic Information*. „Applied Cognitive Psychology" 20(5), s. 633–639.

Clemen R.T. (2001). *Naturalistic Decision Making and Decision Analysis*. „Journal of Behavioral Decision Making" 14(5), s. 359–360.

Collins D. (2003). *Pretesting Survey Instruments, An Overview of Cognitive Methods*. „Quality of Life Research" 12, s. 229–338.

Fischer G.H., Molenaar I.W. (1995). *Rasch Models – Foundations, Recent Developments, and Applications*. Berlin: Springer-Verlag.

Ghosh M. (1995). *Inconsistent MLE for the Rasch Model*. „Statistics and Probability Letters" 23, s. 165–170.

Martin M., Kosinski M., Bjorner J.B., Ware J.E., MacLean R., Li T. (2007). *Item Response Theory Methods Can Improve the Measurement of Physical Function by Combining the Modified Health Assessment Questionnaire and the SF-36 Physical Function Scale*. „Quality of Life Research" 16, s. 647–660.

Masters G.N. (1982). *A Rasch Model for Partial Credit Scoring*. „Psychometrika" 47, s. 149–173.

Neisser U. (1994). *Multiple Systems: A New Approach to Cognitive Theory*. „European Journal of Cognitive Psychology" 6(3), s. 225–241.

Schmidt F.L., Le H., Ilies R. (2003). *Beyond Alpha, An Empirical Examination of the Effects of Different Sources of Measurement Error on Reliability Estimates for Measures of Individual-Differences Constructs*. „Psychological Methods" 8(2) s. 206–224.

Slovic P. (1995). *The Construction of Preferences*. „Am. Psychologist" 50, s. 364–371.

Szałek P. (2004). *Pamięć jako akt intencjonalny (trzy teorie psychologiczne)*. „Przegląd Filozoficzny – Nowa Seria" 13(49), s. 23–38.

Tennant A., Penta M., Tesio L., Grimby G., Thonnard J.-L., Slade A., Lawton G., Simone A., Carter J., Lundgren-Nilsson A., Tripolski M., Ring H., Biering-Sorensen F., Marincek C., Burger H., Phillips S. (2004). *Disordered Thresholds: An Example from the Functional Independence Measure*. „Rasch Measurement Transactions" 17(4), s. 945–948, http://www.rasch.org/rmt/rmt174a.htm (dostęp 29.06.2011).

Ware J.E., Kosinski M., Dewey J.E. (2000*). How to Score Version 2 of the SF-36 Health Survey*. Lincoln, RI: Quality Metric Inc.

The Cognitive Approach Towards Psychometric Scales: When Difficult is Easier than Easy

Psychometric scales are used in situations in which a certain feature of the tested person cannot be measured directly, but can be estimated (as a so-called hidden – or implicit – construct or variable) on the basis of the answers to the questions (items) on a scale. The cognitive approach enables us to have a more profound insight into the psychometric measurement process which is in fact a complex process of communication between people involved in the measurement. It does not require abandoning the standard statistical methods, including Rasch procedures based on the IRT psychometric measurement theory (Item Response Theory), but enables statistics to again perform its proper role of a tool used to confirm the validity of the conclusions of the psychological examination.