PRZEGLĄD PSYCHOLOGICZNY, 2000, TOM 43, NR 1, 39-46

Stronniczość testów psychologicznych:
problemy metodologiczne − konsekwencje społeczne

Elżbieta Hornowska[1]
Szkoła Wyższa Psychologii Społecznej w Warszawie
Instytut Psychologii Uniwersytetu im. A. Mickiewicza

PSYCHOLOGICAL TEST BIAS:
METHODOLOGICAL PROBLEMS – SOCIAL CONSEQUENCES

Summary. The subject of this article is test and item bias. It is a brief discussion of models and statistical procedures to explore the psychometric properties of bias in standarized test. There are at least two approaches that one may take to define the term „bias". One approach is to examine the use to which test scores may be put. Selection of applicants to schools, jobs, and a variety of programs and services are the examples of such situations. This approach requires the comparison of test results for particular groups against some outside criterion. An alternative route to understanding the accusation that tests are biased is to examine the test items themselves in the absence of outside criterion.

Czy obserwowane różnice grupowe w wynikach testowych odzwierciedlają rzeczywiste różnice w poziomie mierzonej cechy? Jeżeli nie, to czy można przyjąć, że różnice te zostały spowodowane przez zastosowanie wadliwego narzędzia pomiarowego, tj. stronniczego testu? Począwszy od przeprowadzonych w 1917 r. w Stanach Zjednoczonych badań imigrantów angielskojęzycznymi testami inteligencji, zarówno psychologowie, jak i opinia publiczna coraz częściej byli bulwersowani licznymi sprawami sądowymi dotyczącymi społecznych konsekwencji stosowania testów psychologicznych. Jedną z głośniejszych spraw tego typu była sprawa *Diana v California State Board of Education* (1970), która została wniesiona do sądu po tym, jak dziewięcioro dzieci hiszpańskiego pochodzenia trafiło do szkoły specjalnej ze względu na orzeczony u nich niski iloraz inteligencji (od 30 I.I. do 72 I.I.). Tymczasem, po powtórnym przetestowaniu – tym razem w języku hiszpańskim, siedmioro z nich poprawiło swoje wyniki przeciętnie o 15 punktów (tj. o jedno odchylenie standardowe!) i znalazło się ponad poprzeczką kwalifikującą do szkoły specjalnej (por. Camilli, Shepard, 1994; McAllister, 1993). W ciągu ostatnich 10 lat uchwalono w Stanach Zjednoczonych ponad 30 aktów prawnych dotyczących procedur stosowanych przez instytucje przeprowadzające badania testowe (McAllister, 1993, s. 389; por. też Witwicki, 1928, który już wówczas zwracał uwagę na ten problem!). Gdzie zatem tkwi błąd? W źle skonstruowanym (stronniczym) narzędziu czy w niewłaściwej polityce selekcji przeprowadzanej na podstawie wyników testowych?

CO TO JEST STRONNICZOŚĆ TESTU?

Stronniczość testu (*test bias*) oznacza brak trafności lub systematyczny błąd pomiaru w stosunku do członków określonych grup. Gdybyśmy np. dokonywali pomiaru czasu wykonania określonego zadania w grupie kobiet i mężczyzn za pomocą stopera chodzącego inaczej w grupie kobiet i mężczyzn, to na pewno możemy określić trafny ranking *wewnątrz* każdej z grup, ale porównania *między* grupami będą prowadziły do fałszywych wniosków. Zauważmy, że stronniczość testu definiowana jest tu w kategoriach różnic między grupami badanych. Jednostka może mieć mniej lub bardziej zgodne z rzeczywistością wyniki w teście, ale nie powiemy, że jest to wina testu, chyba że taka sytuacja zdarzy się u wszystkich badanych (wtedy zakładamy brak trafności testu) lub w określonej grupie badanych (wtedy zakładamy stronniczość metody). Różnice grupowe *per se* nie są jeszcze dowodem istnienia stronności. Dopiero wskazanie − jako ich przyczyny − testu, który dyskryminuje jedną z badanych grup pozwala przyjąć hipotezę o istnieniu stronniczości.
Podkreślmy: stronniczość testu psychologicznego nie jest zatem wynikiem działania błędów losowych – te są bowiem równo rozłożone we wszystkich grupach. Stronniczość testu należy też odróżnić od stronniczej selekcji, tj. takich decyzji o charakterze selekcyjnym, które wykorzystują jako podstawę grupowe różnice w indywidualnych wynikach testowych. Stronniczość testu jest błędem systematycznym, wprowadzanym przez narzędzie pomiarowe.

PROBLEMY METODOLOGICZNE:

1 Adres do korespondencji: Instytut Psychologii UAM, ul. Szamarzewskiego 89, 60-568 Poznań.

ELŻBIETA HORNOWSKA

DWA SPOSOBY DEFINIOWANIA STRONNICZOŚCI

Można mówić o dwóch podstawowych sposobach definiowania, a co za tym idzie − i szacowania stronniczości testu. W pierwszym z nich odwołujemy się do kryterium niezależnego od stosowanego testu (kryterium zewnętrzne), w drugim – do właściwości pozycji tworzących test (kryterium wewnętrzne). Ponieważ celem mojego artykułu nie jest analiza proponowanych w literaturze rozwiązań problemu stronniczości (por. Osterlind, 1983), tu skoncentrujemy się na konsekwencjach, jakie niesie ze sobą przyjęcie określonego stanowiska.

Odwołanie się do kryteriów zewnętrznych

Pierwsze modelowe ujęcie zagadnienia stronniczości testu poprzez odwołanie się do kryterium zewnętrznego pochodzi od Anny Cleary (1968). Podana przez nią definicja stronniczości została sformułowana w terminach regresji wyniku kryterialnego (np. powodzenie w pracy) względem wybranych predyktorów (tu: wyników testowych). Zgodnie z tym podejściem test jest stronniczy wtedy, gdy przewidywanie zachowań badanych osób należących do różnych grup (definiowanych ze względu na np. wiek, płeć czy rasę) z tej samej populacji ogólnej jest obarczone stałym błędem. Innymi słowy, o stronniczości testu świadczą zawsze zbyt wysokie lub zawsze zbyt niskie wyniki kryterialne, otrzymane dla osób należących do różnych grup, a oszacowane na podstawie linii regresji wyznaczonej dla całej populacji (por. rys. 1).

Rys. 1. Przykład testu niestronniczego (rys. u góry) i testu stronniczego (rys. u dołu)
w świetle definicji Cleary (na podstawie: Camilli, Shepard, 1994, s. 11)

W definicji tej zakłada się, że wykorzystywane kryterium jest niestronnicze i jest ono traktowane jako zewnętrzny standard ewaluacji testu. Zarówno w tym przypadku, jak i we wszystkich pochodnych rozwiązaniach nie poszukuje się możliwości wprowadzenia zmian do testu, lecz szuka się rozwiązań pozwalających jednakowo traktować badane grupy na poziomie predykcji. A przecież już Thorndike (1971) pokazał, że równość

predykcji nie oznacza jeszcze, że dwóch kandydatów pochodzących z dwóch różnych grup, którzy otrzymali takie same wyniki kryterialne, ma jednakowe szanse na kontrakt (pozytywną decyzję) – por. rys. 2.

Rys. 2. Przytoczona przez Thorndike'a ilustracja faktu, że choć linie regresji są jednakowe, to test jest stronniczy w stosunku do grupy mniejszości, ponieważ różnice między grupami są większe na poziomie wyników testowych, niż na poziomie wyników kryterialnych (na podstawie: Thorndike, 1971, s. 66)

Analiza przytoczonego tu przypadku, a także wielu innych podobnych sytuacji zaowocowała sformułowaniem dwóch typów modeli stronniczości: biorących za podstawę równość wyników testowych – tzw. *test fair* (w rozumieniu Cleary, 1968) i biorących za podstawę wyniki kryterium – tzw. *performance fair* (w rozumieniu Thurstone'a, 1971). Bez względu na to, do której z tych koncepcji bylibyśmy sami skłonni się przychylić, warto wcześniej zadać sobie pytanie: czy opisywany efekt jest wynikiem stronniczości, czy też artefaktem wynikającym z braku idealnej korelacji między testem a kryterium?

Roztrzygnięcie przyszło ze strony *National Research Council Committee on the General Aptitude Test Battery* (Hartigan, Wigdor, 1989), który analizując sporne przypadki stosowania jednego z najważniejszych testów edukacyjnych w USA, tj. *General Aptitude Test Battery* (tzw. GATB), przypisał je artefaktowi wynikającemu z braku doskonałej korelacji pomiędzy testem a kryterium (por. też Camilli, Shepard, 1994). W tej sytuacji komitet ten zaproponował wprowadzenie specjalnej techniki wyrównującej brak doskonałej korelacji pomiędzy testem a kryterium, a mianowicie technikę dopasowywania kryterium selekcyjnego w zależności od stwierdzonej korelacji pomiędzy testem a kryterium powodzenia. Dopasowanie to – zdaniem członków Komitetu – powinno mieć następujący charakter: jeżeli pomiedzy testem i kryterium selekcji zostanie stwierdzona korelacja bliska jedności, to dopasowanie nie jest konieczne, gdy korelacja ta okaże się zerowa – należy wyrównać średnie wyniki testowe dla grupy mniejszości i grupy większości.

W tej sytuacji powstaje naturalne pytanie: jaka jest społeczna użyteczność przyjętych rozwiązań? Jak się bowiem okazuje, rozwiązanie przyszło nie ze strony metodologii, lecz z zaakceptowanej społecznie polityki. I to społeczna zgoda na przyjęcie określonej strategii wpłynęła na praktykę stosowania testów o charakterze selekcyjnym.

Odwołanie się do kryteriów wewnętrznych

Ogromne trudności w definiowaniu zewnętrznego kryterium predykcji sprawiły, że psychometrzy zaczęli poszukiwać nowych metod szacowania stronniczości, i to takich, które odwoływałyby się się do wewnętrznej struktury testu. Pionierskie badania na tym polu zostały zrealizowane przez zespół w składzie Eells, Davis, Havighurst, Herrick i Tyler (1951). Postawili oni następujące pytania:

(1) Czy obserwowane różnice grupowe w wykonaniu testu są niezmienne względem pozycji testowych, czy też zależą od ich treści (zmieniają się wraz z treścią)?

(2) Czy istnieje jakiś wzorzec formy czy treści pytań testowych, który pozwala obniżyć czy powiększyć różnice grupowe?

(3) Czy względną trudność pozycji testowych można przypisać różnicom kulturowym, charakterystycznym dla osób rozwiązujących test?

Według nich dany test można uważać za niestronniczy, jeżeli składa się on z takich pozycji testowych, które są jednakowo trafne kulturowo dla wszystkich badanych osób. A więc – mówiąc inaczej – niestronnicze są poszczególne pozycje testowe, składające się na test.

Badania te stały się punktem wyjścia dla współczesnych koncepcji stronniczości lokowanej nie na poziomie selekcji, lecz na poziomie testu. O stronniczości pozycji testowych mówimy wtedy, gdy prawdopodobieństwo udzielenia prawidłowej odpowiedzi na daną pozycję testową (bardziej technicznie nazywane „prawdopodobieństwem sukcesu") jest różne dla osób o tej samej wartości mierzonej cechy (najczęściej: taki sam poziom funkcjonowania intelektualnego), pochodzących z różnych podpopulacji (Shepard, Camilli, Averill, 1981; Hulin, Drasgow, Parsons, 1983; Holland, Wainer, 1993).

Powiemy, że pozycja testowa nie jest stronnicza, gdy poziom trudności poszczególnych pozycji testowych jest taki sam dla wszystkich badanych o tym samym poziomie funkcjonowania intelektualnego, bez względu na przynależność tych osób do różnych grup wyłonionych z tej samej populacji.

Idea tego podejścia jest prosta. Przyjmujemy, że prawdopodobieństwo udzielenia poprawnej odpowiedzi na określoną pozycję testową jest wyższe w przypadku osób uzyskujących wysokie wyniki ogólne w teście, niż w przypadku osób uzyskujących niskie wyniki. Jeżeli zatem okaże się, że dana pozycja testowa różnicuje grupy o różnych ogólnych wynikach w teście w ramach jednej podpopulacji, a nie czyni tak w ramach innej, to najczęściej przyjmujemy, iż powodem tego jest różna dostępność do materiału testowego w ramach obu podpopulacji. Jeżeli np. w teście słownikowym znajdą się wyrażenia z gwary poznańskiej, to zadania te mogą zupełnie dobrze różnicować osoby urodzone i mieszkające w Wielkopolsce, natomiast będą źle różnicować, będą za trudne dla pozostałych mieszkańców Polski. Stronniczość jest tu zatem operacjonalizowana w kategoriach

relatywnej trudności pozycji testowych, która zwiększa lub zmniejsza stałe lub typowe różnice grupowe. Badanie stronniczości poprzez odwołanie się do kryteriów wewnętrznych polega na klasyfikowaniu badanych ze względu na ogólny wynik w teście w taki sposób, aby można było sprawdzić, czy osoby badane mające taki sam wynik ogólny (będący wskaźnikiem takiego samego poziomu funkcjonowania intelektualnego), lecz pochodzące z różnych grup, uzyskują takie same wyniki w kolejnych pozycjach testowych. Posługując się ogólnym wynikiem w teście jako kryterium wewnętrznym, zyskujemy podstawę porównywania wyników otrzymanych dla poszczególnych pozycji testowych, niezależnie od przynależności grupowej badanych osób. Jeżeli w wyniku naszej analizy okaże się, że osoby o tym samym wyniku ogólnym osiągają różne wyniki w poszczególnych pozycjach testowych, to możemy przyjąć, że te właśnie pozycje testowe są potencjalnie stronnicze.

Na czym polega problem?
Miary stronniczości oparte na analizie względnej trudności pozycji testowej nie są jeszcze dowodem stronniczości. Wtedy, tylko wtedy, gdy dana pozycja testowa jest względnie trudniejsza dla jednej grupy, a źródło tej trudności wynika ze sposobu konstrukcji testu powiemy, że jest ona stronnicza. Z tego też powodu wprowadzono w literaturze termin „wartości różnicującej pozycji testowej" *(differential item functioning,* w skrócie DIF – por. Holland, Thayer, 1988), aby pokazać, że procedury statystyczne wskazują jedynie na to, czy pozycja testowa zachowuje się różnie w stosunku do różnych grup, czy też wynika to z kontekstu testowania. Statystyka DIF jest wykorzystywana do identyfikowania wszystkich tych pozycji testowych, które wypadają różnie w różnych grupach badanych osób. Następnie, na drodze analizy logicznej, szuka się odpowiedzi na pytanie, dlaczego te pozycje okazały się względnie trudniejsze dla jednej grupy. Dopiero wówczas pozycje testowe z istotnie statystyczną wartością statystyki DIF usuwa się z testu. Zwróćmy jednak uwagę, że ta ostatnia decyzja prowadzi już jednak do określonej *polityki* stosowania testu!
Omawiana tu problematyka lokowana jest w tradycyjnej psychometrii w obszarze trafności testu, a badania nad stronniczością – jako element badań walidacyjnych. Przedstawione tu procedury pozwalają odpowiedzieć jedynie na część pytań dotyczących trafności testu dla określonych grup badanych. Tak czy inaczej – żadna z technik szacowania stronniczości (zewnętrzna czy wewnętrzna) nie pozwoli odpowiedzieć na pełen zakres pytań dotyczących trafności testu, aby można było bez obaw stosować dany test w konkretnym kontekście.
Badania nad stronniczością w sposób oczywisty rozszerzają nasze rozumienie trafności testu.

KONSEKWENCJE SPOŁECZNE

Wczesne standardy dotyczące wymogu trafności – można je nazwać „wymogiem prawdziwości etykietowania" – wymagały, aby twórca testu wykazał, że test mierzy to, co z założenia ma mierzyć. Rosnące zainteresowanie testami i wadliwe ich stosowanie zwiększyło społeczny nacisk na wymóg trafności. Po to, aby można było prowadzić badania, które potwierdziłyby wnioski wyciągane na podstawie wyników testowych, badacze musieli umieć sfromułować te wnioski i następnie je badać. Stało się wyraźne, że wnioski te zależą od konkretnego zastosowania testu. Już w 1971 roku Cronbach (1971) twierdził, że jeżeli wyniki testowe mają stać się podstawą decyzji (zwłaszcza selekcyjnych), to konsekwencje tych decyzji muszą być elementem badań trafnościowych. Stąd w *Standardach dla testów stosowanych w psychologii i pedagogice*, wydanych w 1985 roku (American Psychological Association, 1985) sformułowano już – idąc za Cronbachem (1971) – następującą dyrektywę: jeżeli test jest wykorzystywany np. do podejmowania decyzji o kierowaniu do różnych placówek (np. szkół specjalnych) należy wykazać, że idący za tym inny sposób oddziaływania stanie się skuteczny. Kluczowym pojęciem staje się skuteczność; dzieci skierowane do szkół specjalnych muszą się lepiej rozwijać tam, niż gdyby zostały w dotychczasowym środowisku. Podczas gdy tradycyjne badania walidacyjne można określić jako udowadnianie prawdziwości w etykietowaniu, to współczesne można porównać do testowania nowego leku – z jednakowym naciskiem zarówno na efekty uboczne, jak i zamierzone korzyści.
Techniczne badania stronniczości nie mogą zatem być wykorzystywane jedynie jako źródło informacji na temat trafności zastosowania testu dla określonych grup społecznych. Trafność to również skuteczność zastosowanego postępowania posttestowego. Bez określonej polityki społecznej, wyraźnie fromułowanej, opartej na wynikach badań, techniczne analizy stronniczości testu mogą stać się przykrywką dla niekompetencji urzędników. Taka polityka społeczna zaś nie może powstawać poza psychologami, specjalistami w zakresie pomiaru psychologicznego. Najwyższy zatem czas, aby powołać do życia polski komitet ds. stosowania testów psychologicznych i ukrócić samowolę wielu instytucji (zwłaszcza tych zajmujących się doradztwem personalnym), które stosują testy w sposób całkowicie niezgodny z kanonami sztuki (często są to metody bez określonej rzetelności, trafności i polskich norm) i – co gorsza – na ich podstawie podejmują decyzje o przyszłości badanych przez siebie osób.

BIBLIOGRAFIA

American Psychological Association, American Educational Research Association, National Council on Measurement in Education (1985). *Standards for educational and psychological testing.* Washington, DC.: American Psychological Association.
Camilli, G., Shepard, L. A. (1994). *Methods for identyfying biased test items*. Beverly Hills−London: Sage Publications.
Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and white students in integrated colleges. *Journal of Educational Measurement,* 5, 115-124.
Cronbach, L. J. (1971[2]). Test validation. [W:] R. L. Thorndike (red.), *Educational measurement* (s. 443-507). Washington D.C.: American Council of Education.
Diana v. California State Board od Education (1970). U.S. District Court for the Northern District of California.
Eells, K., Davis, A., Havighurst, R. J., Herrick, V. E., Tyler, R. W. (1951). *Intelligence and cultural differences*. Chicago: University of Chicago Press.
Hartigan, J. A., Wigdor, A. K. (red) (1989). *Fairness in employment testing*: *validity generalization, minority issues, and the General Aptitude Test Battery.* Washington, DC.: National Academy Press.
Holland, P. W., Thayer, D. T. (1988). Differentiae item performance and the Mantel-Haenszel procedure. [W:] H. Wainer, H. Braun (red.), Test Validity (s. 129-145). Hillsdale, N.J.: Erlbaum.
Holland, P. W., Weiner, P. (1993). *Differential item functioning*. Hillsdale, N.J.: Lawrence Erlbaum.
Hulin C. L., Drasgow F., Parsons C. K. (1983). *Item response theory. Application to psychological measurement*. Homewood, Il.: Dow Jones-Irwin.
McAllister, P. H. (1993). Testing, DIF, and public policy. [W:] P. W. Holland, H. Wainer (red.), *Differential item functioning* (s. 389-396). Hillsdale, N.J.: Erlbaum.
Osterlind, S. J. (1983). *Test item bias*. Beverly Hills, CA.: Sage.
Shepard, L., Camilli, G., Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. *Journal of Educational Statistics*, 6, 317-376.
Thorndike, R. L. (1971). Concepts of culture fairness. *Journal of Educational Measurement*, 8, 63-70.
Witwicki, W. (1928). O narodowych testach amerykańskich do badania inteligencji. *Psychotechnika*, 7, 23-32.