FACULTEIT PSYCHOLOGIE EN PEDAGOGISCHE WETENSCHAPPEN
DEPARTEMENT PSYCHOLOGIE
ONDERZOEKSGROEP HOGERE COGNITIE EN INDIVIDUELE VERSCHILLEN
CENTRUM VOOR ORGANISATIE- EN PERSONEELSPSYCHOLOGIE
TIENSESTRAAT 102 – 3000 LEUVEN

KATHOLIEKE
UNIVERSITEIT
LEUVEN

Hebben mannen en vrouwen gelijke kansen bij selectieproeven met intelligentietests?

Samenvatting

Dr. Michel Meulders
Miek Vandenberk
Prof. Dr. Paul De Boeck
Prof. Dr. Karel De Witte
Dr. Rianne Janssen

Juni 2004

Inleiding

In het kader van het VIONA-onderzoeksproject “Psychologische testen en de effecten op de instroom van kansengroepen in het Ministerie van de Vlaamse Gemeenschap en in de Vlaamse privé-bedrijven”, werd gevraagd om voor drie kansengroepen (vrouwen, allochtonen en gehandicapten) te onderzoeken of relevante intelligentietests een bias vertonen. Deze samenvatting beschrijft het onderzoek dat gevoerd werd voor de doelgroep vrouwen. De centrale vraag van dit onderzoek is dus of mannen en vrouwen gelijke kansen hebben bij selectieproeven met intelligentietests?

Probleemstelling

In de context van personeelsselectie worden scores op psychologische tests vaak gebruikt als predictoren voor bepaalde job-criteriumvariabelen zoals, bijvoorbeeld, succes in een bepaalde job. Hierbij wordt aangenomen dat de variabele die men beoogt te meten met de test (intelligentie) empirisch predictief is voor de job-criteriumvariabele waarin men geïnteresseerd is. Het gebruiken van testscores als meting van de bedoelde meetvariabele is slechts mogelijk als voldaan is aan twee voorwaarden: (1) de testscores moeten empirisch predictief zijn voor het construct dat men wil meten en (2) de testscores mogen niet door nog andere variabelen (geslacht, etnische achtergrond, …) dan de bedoelde meetvariabele bepaald worden. Als dit toch het geval is, dan is er sprake van bias. Als de succeskans voor een item bij personen met eenzelfde intelligentie verschilt naargelang de groep waartoe men behoort, dan spreekt men van itembias of “differential item functioning” (DIF).
In dit project worden schendingen van de tweede voorwaarde onderzocht. Meer specifiek gaan we na of er een verschil is in de succeskans van mannen en vrouwen op de verschillende items van een test. Indien er verschillen worden gevonden, dan kunnen deze ook een invloed hebben op de totale testscore, zodat er discriminatie optreedt ten aanzien van een bepaalde groep.

Methodologie

Het onderzoek naar bias bij de doelgroep vrouwen bestond uit vier fasen. In de eerste fase werden testgegevens opgevraagd voor relevante intelligentietests. Hiervoor werd samenwerking gezocht met SELOR. De testbatterij van SELOR bestaat uit 14 computertests en hieruit werden volgende vier tests gekozen: LOGDED (logisch redeneren), ANAVERB (verbale analogieën), CODES (code leren) en NUMVA (cijferreeksen). Deze tests worden frequent afgenomen, ze meten verschillende soorten intelligentie, ze kunnen ontleed worden in termen van cognitieve processen en er zijn van deze tests voldoende gegevens voorhanden om een betrouwbaar biasonderzoek te doen. Daarnaast werd ook contact opgenomen met Defensie. Alle kandidaten die in de provinciale defensiehuizen informeren voor een functie bij het leger zijn verplicht om drie computertests af te leggen: DGEO (spatiale oriëntatie), TNV (algemene intelligentie) en WIMA (numerieke vaardigheden). Deze tests beantwoorden ook aan de hoger vermelde criteria.

In de tweede fase werd voor elke test onderzocht welke items een bias vertonen voor de doelgroep in kwestie. Als uit de tweede fase bleek dat bepaalde items een bias vertoonden, dan gingen we in de derde fase op zoek naar een inhoudelijke verklaring voor deze bias. Dit kan bijvoorbeeld door de samenhang tussen de bias en bepaalde itemkenmerken te onderzoeken.
In een vierde fase tenslotte, werd het effect van bias in individuele items op de testscore (aantal juiste antwoorden) en de gecorrigeerde testscore (aantal juiste antwoorden dat gecorrigeerd is voor raden bij meerkeuzevragen) onderzocht. Indien het effect van bias op de testscore substantieel was, werd aangegeven hoe de test eventueel kon worden aangepast. Onze aanpak voor biasonderzoek, zoals die aangewend wordt in de fases 2 tot 4, is gebaseerd op item-responstheorie.

Resultaten en beleidsaanbevelingen

Gemiddelde prestaties van mannen en vrouwen

De gemiddelde prestaties van mannen en vrouwen op een bepaalde test kunnen verschillend zijn. Zo kan het zijn dat één groep gemiddeld hoger scoort op de test dan de andere groep. Wanneer mannen of vrouwen gemiddeld beter presteren op een bepaalde test, dan wil dit zeggen dat zij in principe meer kans hebben om geselecteerd te worden. De groep die gemiddeld lager scoort, wordt dan gediscrimineerd. Deze discriminatie is evenwel terecht als de specifieke vaardigheden die gemeten worden door de test ook werkelijk van belang zijn voor de succesvolle uitoefening van de job. Als het daarentegen niet duidelijk is of deze specifieke vaardigheden leiden tot een betere jobperformantie dan is de discriminatie onterecht.

Het onderzoek wijst uit dat vrouwen op bepaalde tests gemiddeld slechter scoren dan mannen (LOGDED, ANAVERB, WIMA, DGEO), terwijl er voor andere tests geen verschil gevonden werd tussen mannen en vrouwen (TNV, NUMVA, CODES). Deze resultaten geven een eerste aanwijzing over het verschil in de gemiddelde vaardigheid van mannen en vrouwen voor de onderzochte tests. Verder onderzoek is echter nodig om na te gaan of het vastgestelde verschil niet gedeeltelijk kan verklaard worden door kenmerken van de onderzochte steekproeven, zoals bijvoorbeeld het opleidingsniveau van de mannelijke en vrouwelijke kandidaten. Het zou bijvoorbeeld kunnen dat toevallig meer mannen in de steekproef een exact wetenschappelijke opleiding hebben gevolgd, wat de betere prestaties van mannen voor logisch redeneren gedeeltelijk zou verklaren.

Om onterechte discriminatie van mannen of vrouwen in de selectiepraktijk te vermijden, zouden tests die specifieke vaardigheden meten en daardoor moeilijker zijn voor één bepaalde groep, alleen mogen gebruikt worden als aangetoond is dat de vaardigheden die ze meten echt noodzakelijk zijn voor succes in de job. Anders geeft men beter de voorkeur aan tests die even moeilijk zijn voor beide groepen.
Concreet betekent dit dat in het kader van algemene wervingen (waar een duidelijk functieprofiel in principe ontbreekt) de tests LOGDED, ANAVERB, WIMA en DGEO best enkel in combinatie met andere tests gebruikt worden en nog beter dat ze worden gecorrigeerd en/of er alternatieven voor voorzien worden. Daarnaast is het nodig om ook voor andere veel gebruikte tests te

onderzoeken of ze moeilijker zijn voor mannen of voor vrouwen. Om zo goed mogelijk in te schatten of tests moeilijker zijn voor een bepaalde groep, is het nodig om tijdens selecties de (anonieme) ruwe gegevens en bepaalde achtergrondvariabelen systematisch te registreren. Op dit moment is dit nog niet het geval.

DIF in de moeilijkheidsgraden van items

De resultaten van het onderzoek tonen aan dat alle onderzochte tests een aantal "discriminerende" (DIF) items bevatten die moeilijker zijn voor mannen of vrouwen met dezelfde totaalscore op de test. De meeste tests vertonen DIF in ongeveer één derde van de items. Bij CODES is er DIF in minder dan 10% van de items en bij ANAVERB is er DIF in ongeveer de helft van de items. We kunnen stellen dat bij de meeste discriminerende items de succeskansen van mannen en vrouwen niet veel verschillen maar bij een beperkt aantal items is het verschil in succeskansen wel van betekenis. Bijvoorbeeld, in ongeveer 10% van de ANAVERB en NUMVA items zijn de verschillen tussen succeskansen van mannen en vrouwen op een klein stuk van de latente schaal (2.5 % van het relevante bereik van de schaal) minstens .20. Bij deze tests is het aangeraden om een aantal sterk discriminerende items te verwijderen.
Aangezien alle onderzochte tests een aantal discriminerende items bevatten, is het aangeraden om in de toekomst ook voor nog andere tests te onderzoeken welke items discriminerend zijn voor mannen of vrouwen. Net zoals voor het onderzoek naar gemiddelde prestaties, is het hiervoor nodig dat de ruwe testgegevens en bepaalde persoonsvariabelen geregistreerd worden tijdens de selectie.

Gezamenlijk effect van DIF in individuele items op somscores en gecorrigeerde somscores

Voor alle onderzochte tests blijkt dat de DIF-effecten in individuele items elkaar opheffen en als dusdanig geen effect hebben op de (verwachte) somscores en gecorrigeerde somscores. Met andere woorden, het verband tussen de latente variabele en de prestatie op de volledige test is voor de onderzochte tests hetzelfde in beide groepen. Een uitzondering is de ANAVERB test waar mannen met een lage vaardigheid hogere somscores behalen dan vrouwen met dezelfde lage vaardigheid. Dit effect is evenwel minder uitgesproken als de somscores gecorrigeerd worden voor raden. Hoewel het verschil in testscores beperkt blijft, is het toch aanbevolen om bij ANAVERB sterk discriminerende items te verwijderen.
Men kan dus stellen dat somscores en gecorrigeerde somscores valide maten zijn om kandidaten onderling of uit verschillende groepen te vergelijken. Maar men moet zich er wel van bewust zijn dat éénzelfde testscore iets anders kan betekenen voor mannen en vrouwen, omdat ze het resultaat kan zijn van succes op verschillende verzamelingen van items die tegengestelde DIF vertonen.

Verder stellen we vast dat bij korte tests de statistische onzekerheid in de testscore tamelijk groot is. Bijvoorbeeld, in de steekproef van vrouwen bij LOGDED ligt de somscore 12 (op 22 items) met 95% zekerheid tussen 8.5 en 15.5, zodat men moeilijk een onderscheid kan maken tussen goede en slechte kandidaten. Ook bij het ordenen van kandidaten op basis van hun somscores is het van belang om rekening te houden met de onzekerheid in de scores. Kandidaten kunnen immers strikt genomen slechts geordend worden als de 95% betrouwbaarheidsintervallen rond hun somscores niet overlappen. Daarom zou het een standaardpraktijk moeten zijn om naast de

testscore de onzekerheid te rapporteren, zodat onbelangrijke verschillen tussen testscores niet kunnen doorwegen op de besluitvorming. Methoden uit de itemresponstheorie kunnen helpen om de vaardigheid van een persoon nauwkeuriger te schatten. Omdat de tests op zulke grote schaal gebruikt worden, zou men hier zeker alle inspanningen moeten doen om personen optimaal te ordenen.

Verklaren van DIF

Aangezien de onderzochte tests allemaal voor bepaalde items DIF vertonen, is het belangrijk om te onderzoeken op welke manier de DIF verklaard kan worden. Een inzicht in welke items relatief moeilijker zijn voor mannen of vrouwen geeft immers een beter inzicht in de constructvaliditeit van de test en in de betekenis van de somscores van mannen en vrouwen.

Voor elk van de onderzochte tests werd nagegaan in welke mate DIF in de moeilijkheidsgraden kan verklaard worden op basis van objectieve itemkenmerken. Daarnaast werd ook onderzocht in welke mate de moeilijkheidsgraden in elke groep kunnen verklaard worden op basis van itemkenmerken. De itemkenmerken werden telkens bepaald op basis van een cognitieve analyse van de testitems.

We stellen vast dat het voor de meeste tests erg moeilijk is om te verklaren waarom bepaalde items een verschillende moeilijkheidsgraad hebben voor mannen en vrouwen. De resultaten voor het verklaren van de moeilijkheidsgraden voor mannen en vrouwen zijn analoog. Voor de meeste tests blijft het grootste deel van de moeilijkheidsgraden onverklaard. Alleen voor LOGDED kunnen de moeilijkheidsgraden redelijk goed gevat worden in termen van een beperkt aantal cognitieve operaties die nodig zijn om de items op te lossen.

We kunnen besluiten dat het modelleren van moeilijkheidsgraden van testitems op basis van een inhoudelijke theorie over de cognitieve processen die een rol spelen bij het oplossen van de items, of op basis van andere objectieve itemkenmerken, slechts een goede kans op slagen heeft als de test geconstrueerd werd op basis van een theorie en als een beperkt aantal itemkenmerken of processen in voldoende items herhaaldelijk aan bod komen (zoals bij de LOGDED test). Als een test relatief veel verschillende itemkenmerken bevat die telkens slechts bij enkele items aan bod komen, dan is het veel moeilijker om te begrijpen wat een item moeilijk maakt en dan is het erg waarschijnlijk dat de resultaten slechts beperkt generaliseerbaar zijn naar nieuwe testitems.

Hoewel de DIF in de moeilijkheidsgraad maar beperkt verklaard kan worden, toch werden voor verschillende tests gelijkaardige verbanden gevonden tussen itemkenmerken en de geobserveerde DIF:

Bij LOGDED en TNV blijkt dat items die een beroep doen op visuele intelligentie moeilijker zijn voor vrouwen dan voor mannen, wat in overeenstemming is met de literatuur over intelligentie onderzoek. Dergelijke effecten zijn onwenselijk als de test zelf niet bedoeld is om visuele intelligentie te meten. Het is dus aanbevolen om ook voor andere tests na te gaan of er items zijn die onbedoeld visuele intelligentie meten.

2. Bij LOGDED en WIMA blijkt dat items die meer stappen vereisen om tot de oplossing te komen gemakkelijker zijn voor vrouwen. We hebben hiervoor echter geen goede verklaring.

Verdere aanbevelingen

Om een gelijkekansenbeleid te voeren, dient men te vermijden dat kansengroepen ten onrechte gediscrimineerd worden op basis van tests. Dit kan door de kwaliteit van tests op een aantal punten systematisch te bewaken. Zoals bij geneesmiddelen en voedingswaren al het geval is, zou men de kwaliteitscontrole op tests die gebruikt worden in selecties bij de overheid en in de privéondernemingen kunnen verscherpen en systematiseren. Als een nieuwe test op de markt gebracht wordt door een testproducent, zou men de test een kwaliteitslabel kunnen toekennen dat aangeeft in welke mate de test onderzocht is op aspecten als betrouwbaarheid, validiteit, afwezigheid van discriminatie ten aanzien van kansengroepen zoals allochtonen, enz. Het toekennen van een kwaliteitslabel zou bijvoorbeeld kunnen gebeuren door een onafhankelijke commissie van testexperts (zoals bijvoorbeeld de Commissie Test Aangelegenheden Nederland die is ingesteld door het Nederlands Instituut van Psychologen).

Om in de toekomst de kwaliteit van tests (bij de overheid) te bewaken en te optimaliseren, dient men de ruwe testgegevens en relevante achtergrondvariabelen na de selectie te inventariseren. Vervolgens kan men op basis van de analyse van deze gegevens de kwaliteit van de tests optimaliseren. Deze optimalisatie zou bijvoorbeeld kunnen gebeuren door testexperts in dienst van de overheid of zou bijvoorbeeld kunnen uitbesteed worden aan een wetenschappelijke instelling.

Hoe een goede en betrouwbare rangordening te maken van kandidaten op basis van de scores die behaald werden op basis van één of meerdere selectietests, blijkt een belangrijk vraagstuk te zijn dat ook ter sprake komt in het eerste deelrapport bij redelijke aanpassingen voor personen met een handicap. In de besprekingen in verband met dit onderzoek blijkt dat een afstemming op het terrein tussen enerzijds de verwachtingen van de overheid als opdrachtgever en anderzijds de selecteurs, rekening houdende met de aanbevelingen in de deelrapporten, zich opdringt.