onderzoeksvoorstel voor BEPERKTE VIONA-ONDERZOEKSOPROEP 2002 rond de onderzoeksvraag "Psychologische testen en de effecten op de instroom van kansengroepen in het Ministerie van de Vlaamse Gemeenschap en in de Vlaamse privébedrijven"

1. promotoren

promotor-woordvoerder

Naam: Michel Meulders
Functie: postdoctoraal onderzoeker onderzoeksfonds KU Leuven
Instelling: KU Leuven
Faculteit: Psychologische en Pedagogische Wetenschappen
Onderzoekseenheid: Hogere Cognitie en Individuele Verschillen
Contactadres: Tiensestraat 102, 3000 Leuven
Telefoon.: 32-16-325985 Fax: 32-16-325916
E-mail: michel.meulders@psy.kuleuven.ac.be

co-promotoren

Naam: Paul De Boeck
Functie: gewoon hoogleraar
Instelling: KU Leuven
Faculteit: Psychologische en Pedagogische Wetenschappen
Onderzoekseenheid: Hogere Cognitie en Individuele Verschillen
Contactadres: Tiensestraat 102, 3000 Leuven
Telefoon.: 32-16-325980 of 32-16-326004 Fax: 32-16-325916
E-mail: paul.deboeck@psy.kuleuven.ac.be

Naam: Karel De Witte
Functie: hoofddocent
Instelling: KU Leuven
Faculteit: Psychologische en Pedagogische Wetenschappen
Onderzoekseenheid: Centrum voor Organisatie en Personeelspsychologie
Contactadres: Tiensestraat 102, 3000 Leuven
Telefoon: 32-16-326062
E-mail: karel.dewitte@psy.kuleuven.ac.be

Naam: Rianne Janssen
Functie: postdoctoraal onderzoeker FWO
Instelling: KU Leuven
Faculteit: Psychologische en Pedagogische Wetenschappen
Onderzoekseenheid: Leuvens Instituut voor Onderwijsonderzoek
Contactadres: Dekenstraat 2, 3000 Leuven
Telefoon: 32-16-326184 Fax: 32-16-326274
E-mail: rianne.janssen@ped.kuleuven.ac.be

2. Titel van het onderzoeksproject

De bruikbaarheid van intelligentietesten voor de selectie van kansengroepen

3. Bondige omschrijving van het onderzoeksproject

Dit project bevat een onderzoeksplan in antwoord op de VIONA-onderzoeksoproep met als onderwerp “Psychologische testen en de effecten op de instroom van kansengroepen in het ministerie van de Vlaamse Gemeenschap en in de Vlaamse privébedrijven”. De bedoeling van het project is te onderzoeken of relevante intelligentietests een bias vertonen voor elk van drie kansengroepen, namelijk, vrouwen, allochtonen, en gehandicapten.

Conceptueel kader

In de context van personeelsselectie worden psychologische testscores 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, en dus ook dat de testscore predictief is. De bedoelde meetvariabele (de intelligentie) wordt bepaald door determinerende factoren zoals vooropleiding, motivatie, genetische aanleg, de maatschappelijke groep waartoe men hoort, enz… 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, d.w.z. de constructvaliditeit van de test moet gewaarborgd zijn, en (2) de testscores mogen niet nog door andere variabelen dan de bedoelde meetvariabele bepaald worden, m.a.w. de test mag geen bias vertonen. In dit project worden schendingen van de tweede voorwaarde onderzocht.

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). Testbias of "differential test functioning" (DTF) kan op verschillende manieren opgevat worden. Naast de gangbare opvatting in itemrespons theorie (IRT) van testbias als het effect van bias in individuele items op testscores (Shealy & Stout, 1993) hanteren we nog een tweede opvatting, namelijk testbias als gelijke bias in al de items van een bepaalde test. In de literatuur wordt de term bias soms gereserveerd voor het geval dat de constructvaliditeit van de test gewaarborgd is, maar wij zullen in het vervolg de termen itembias (testbias) en DIF (DTF) als synoniemen gebruiken.

Het voorgestelde onderzoeksplan bevat vier fasen: (1) bepalen welke intelligentietests dienen onderzocht te worden, (2) onderzoeken of deze tests een bias vertonen voor de relevante subgroepen, (3) in geval er een bias wordt vastgesteld, een verklaring zoeken hiervoor, (4) aangeven hoe bias verwijderd kan worden of hoe de test eventueel kan worden aangepast.

Eerste fase

Om te bepalen welke tests onderzocht zullen worden, gebruiken we drie criteria:
(1) Het soort intelligentie moet belangrijk zijn in de selectieprocedures van de overheid en van andere organisaties. Door rekening te houden met het soort van tests kunnen we het gemeenschappelijke belichten in de tests van verschillende organisaties, wat het generaliseren van de resultaten naar andere contexten ten goede komt. (2) Tests die frequent gebruikt worden voor selectie van de relevante doelgroepen verdienen de voorkeur. (3) Tests waarvoor al gegevens voorhanden zijn of waarvoor gegevens verzameld kunnen worden bij de relevante doelgroepen verdienen de voorkeur aangezien hierbij de ecologische validiteit gewaarborgd is.

Voor de concrete uitvoering van de eerste onderzoeksfase kunnen de volgende stappen onderscheiden worden: (1) Nagaan welke intelligentietests gebruikt worden voor personeelsselectie bij de overheid en in privéondernemingen. (2) Classificatie van deze tests naar het soort intelligentie dat gemeten wordt. Hiervoor kan een beroep gedaan worden op erkende classificatiesystemen zoals het systeem van de French kit (Ekstrom, French, Harman, & Dermen, 1976) en het syteem van Caroll (1993).
(3) Samenwerking zoeken met selectie centra om gebruik te maken van bestaande gegevens en om nieuwe gegevens te verzamelen. Voor biasonderzoek naar geslacht zijn wellicht bestaande gegevens voorhanden. Voor biasonderzoek naar etnische afkomst is het nodig om nieuwe gegevens te verzamelen en zal de vrijwillige medewerking gevraagd worden van sollicitanten die deelnemen aan bestaande

selectieprocedures. De bevraagde informatie over lidmaatschap van een bepaalde groep zal alleen gebruikt worden voor dit project en zal niet doorgegeven worden aan de organisatie.

Biasonderzoek met betrekking tot een handicap is waarschijnlijk niet haalbaar omdat voor een specifieke handicap (vb. dyslexie) te weinig gegevens voorhanden zijn (te kleine aantallen) om betrouwbare conclusies te kunnen trekken. We stellen voor om dit probleem in eerste instantie theoretisch te onderzoeken door middel van een literatuurstudie.

Wat betreft de concrete gegevens die zullen geanalyseerd worden voor bias streven we ernaar om via twee uitgebreide selectieprocedures (één voor de overheid en één voor de privé) met grote aantallen kandidaten gegevens van drie intelligentietests te bekomen die afgenomen zijn van dezelfde kandidaten.

Tweede en derde fase

In deze fasen worden DIF en DTF beschrijvend (tweede fase) en verklarend (derde fase) onderzocht. Voor het beschrijven van DIF en DTF zullen we vooral gebruik maken van procedures uit de IRT omdat deze benadering de meeste mogelijkheden biedt en meer verfijnde analyses mogelijk maakt. Bij IRT-modellen worden de succeskansen van personen voor de itemantwoorden gemodelleerd als een functie van persoons- en itemparameters. De persoonsparameters beschrijven de posities van personen op één of meerdere continue latente variabelen (bedoelde meetvariabelen). De itemparameters beschrijven de moeilijkheidsgraad van items en de mate waarin individuele items toelaten een onderscheid te maken tussen personen met een hoge en lage vaardigheid (verder "discriminatiegraad" genoemd). Bij onderzoek naar DIF zullen we IRT-modellen op twee manieren gebruiken: een itemsgewijze manier en een verklarende manier.

De itemsgewijze methode bestaat erin om per item achtereenvolgens een interactie toe te laten tussen de groepsvariabele en het item betreffende de moeilijkheidsgraad en de discriminatiegraad. Er is sprake van uniforme DIF als de moeilijkheidsgraad van een item anders is naargelang de groep en er is sprake van niet-uniforme DIF als daarnaast ook de discriminatiegraad verschilt naargelang de groep. Door de interactie-effecten te modelleren als random effecten kan onderzocht worden in welke mate DIF verschilt binnen groepen (Van den Noortgate & De Boeck, 2002).

De verklarende methode bestaat erin om niet itemsgewijs te werken, maar om naar itemkenmerken te zoeken die ofwel gemeenschappelijk zijn voor items met DIF, of om itemkenmerken te gebruiken die men op voorhand bepaalt door een inhoudsanalyse van de items. Voor de itemkarakteristieken die aldus werden bepaald, kunnen interacties met de groepsvariabele opgenomen worden in het IRT-model, in de eerste plaats voor de moeilijkheidsgraad, maar ook voor de discriminatiegraad. In geval van een significante interactie tussen de groepsvariabele en de itemkarakteristiek spreken we van "Differential Feature Functioning" (DFF). DFF modellen met en zonder individuele verschillen werden beschreven door Meulders en Xie (2002).

Het onderzoek naar DTF met IRT-modellen kan op verschillende manieren geconceptualiseerd worden.

Een elegante manier die mogelijk is als we per persoon beschikken over gegevens van meerdere testen (zie eerste fase), is een multidimensioneel DFF model te gebruiken met itemkarakteristieken die aangeven tot welke test een item behoort. Het model is multidimensioneel omdat per test een verschillende latente variabele nodig is. Daarnaast zullen we ook het effect van bias in individuele items op testscores nagaan (zie vierde fase).

Vierde fase

Indien bias vastgesteld wordt, zal eerst nagegaan worden in welke mate dit effect heeft op de testscores. Als het effect aanzienlijk is, kan de bias verwijderd worden door individuele items te verwijderen (itemsgewijze methode) of door items met itemkarakteristieken die DFF vertonen te verwijderen (verklarende methode). Indien nodig zal de test aangepast worden door nieuwe items te formuleren, of door richtlijnen te formuleren voor een mogelijke aanpassing.

4. Gedetailleerd tijdschema

De voorgestelde duur voor het project is 21 maanden. Het project zal opgestart worden vanaf 15 december en er zal zo snel mogelijk een medewerker aangeworven worden. Omdat er waarschijnlijk pas een medewerker kan aangeworven worden vanaf 1-1-2003, wordt in onderstaand tijdschema en bij de berekening van de personeelskost de periode 1-1-2003 tot 30-9-2004 gebruikt. De verschillende fasen van het onderzoek worden besproken in de projecttekst. Hieronder volgt enkel een schematisch overzicht:

<u>Fase 1: selectie van relevante intelligentietests</u>

januari-februari 2003

- inventariseren van intelligentietests die vooral gebruikt worden voor selectie
- literatuurstudie met betrekking tot DIF in intelligentietests bij doelgroepen (speciale aandacht voor gehandicapten)
- bepalen welke tests dienen gescreend te worden op bias bij een bepaalde doelgroep op basis van voorgestelde criteria

maart-augustus 2003

- indien beschikbaar, bestaande gegevens met geslacht als groepsvariabele klaarmaken voor de psychometrische analyse (als dit onderdeel is uitgevoerd kan al begonnen worden met de psychometrische analyses)
- nieuwe gegevens verzamelen met etnische afkomst als groepsvariabele

<u>Fase 2 en 3 : beschrijven (fase 2) en verklaren (fase 3) van DIF en DTF</u>

september 2003

- psychometrische analyses detectie DIF

oktober-december 2003

- itemkarakteristieken formuleren voor verklaring van DIF op basis van de resultaten van fase 3 of door middel van analyse itemfeatures
- psychometrische analyses DFF

januari 2004

- psychometrische analyses DTF

Fase 4 : DIF verwijderen of aangepaste tests formuleren

februari-mei 2004

- effect nagaan van DIF op testscores
- nagaan of tests kunnen aangepast worden door items met DIF te verwijderen of door items met bepaalde itemkarakteristieken te verwijderen, eventueel aangepaste items formuleren

juni-september 2004

- toetsen onderzoeksresultaten bij betrokkenen en opmaak eindrapport

5a. Valorisatie van de onderzoeksresultaten

De resultaten van het onderzoek zullen op verschillende manieren gevaloriseerd worden:

a. toetsen van beleidsaanbevelingen bij de betrokken organisaties waardoor deze reeds geïnformeerd worden
b. toetsen van beleidsaanbevelingen bij Algemene Vergadering/ Raad van Beheer Federgon
c. meedelen van resultaten in een eindrapport
d. organisatie van een studienamiddag waar de resultaten gepresenteerd worden
e. publicatie van de belangrijkste resultaten in vaktijdschriften zoals HR magazine en algemene pers zoals vacature en job@
f. publicatie van de belangrijkste resultaten op de website van Federgon
g. voorstelling en bespreking van de onderzoeksresultaten in FWS-opleiding (Federatie voor werving en selectie) voor consulenten
h. voorstellen onderzoeksresultaten in programma's en cursussen HRM aan universiteiten
i. formuleren van richtlijnen voor de adviescommissie private arbeidsbemiddeling

5b. Beleidsrelevantie van het onderzoeksproject

Er bestaan op dit ogenblik geen tests die gebruikt worden in de sector van de personeelsselectie die onderzocht zijn op bias. De beleidsrelevantie van het voorgestelde onderzoek kan op verschillende niveaus gesitueerd worden. Door dit project

- wordt de aandacht gevraagd en wordt de sector gesensibiliseerd voor de problematiek
- zullen er tests beschikbaar worden die wel op bias onderzocht zijn en die aangepast worden zodat ze vrij zijn van bias
- worden dankzij het verklaringsgedeelte de principes zichtbaar, zodat men ook in de praktijk enige feeling krijgt voor waar bias zou kunnen optreden
- wordt duidelijk hoe men tests effectief kan controleren, zodat deze controle in principe ook kan uitgevoerd worden door wie dat wil

Er wordt steeds meer aandacht besteed aan het probleem van discriminatie. Via het bias-onderzoek kan er daadwerkelijk iets gebeuren dat concreet en haalbaar is en de discriminatie tegen gaat. Het probleem is vermoedelijk veel groter, maar in zijn geheel ook moeilijk op te lossen binnen de context van een project.

6. de onderzoeksverantwoordelijken

Michel Meulders

vijf belangrijkste publicaties

- Meulders, M., Gelman, A., Van Mechelen, I., & De Boeck P. (1998). Generalizing the probability matrix decomposition model: An example of Bayesian model checking and model expansion. In J. Hox, & E. De Leeuw (Eds.), *Assumptions, robustness, and estimation methods in multivariate modeling* (pp. 1-19). TT Publicaties: Amsterdam.

- Meulders, M., De Boeck, P., Van Mechelen, I., Gelman, A., & Maris, E. (2001). Bayesian inference with probability matrix decomposition models. *Journal of Educational and Behavioral Statistics, 26,* 153-179.

- Meulders, M., De Boeck, P., & Van Mechelen, I. (2001). Probability matrix decomposition models and main-effects generalized linear models for the analysis of replicated binary associations. *Computational Statistics and Data Analysis, 38,* 217-233.

- Meulders, M., De Boeck, P., & Van Mechelen, I. (2002). A taxonomy of latent structure assumptions for probability matrix decomposition models. Aanvaard bij *Psychometrika.* (37 pagina's).

- Meulders, M., De Boeck, P., Kuppens, P., & Van Mechelen, I. (in druk). Constrained latent class analysis of three-way three-mode data. *Journal of Classification*. (25 pagina's)

Paul De Boeck

-

vijf relevante publicaties

- Rijmen, F., & De Boeck, P. (2002). The random weights linear logistic test model. *Applied Psychological Measurement*, *26*, 271-285.

- Hoskens, M., & De Boeck, P. (2001). Multidimensional componential IRT models. *Applied Psychological Measurement, 25,* 19-37.

- Tuerlinckx, F., & De Boeck, P. (2001). The effect of ignoring item interactions on the estimated discrimination parameters. *Psychological Methods, 6,* 181-195.

- Verguts, T., & De Boeck, P. (2001). Some Mantel-Haenszel tests of Rasch model assumptions. *British Journal of Mathematical and Statistical Psychology, 54,* 21-37.

- De Boeck, P., & Wilson, M. (Eds.) *Psychometrics using logistic mixed models*.

Dit boek is een gezamenlijk initiatief van de BEAR groep in UC Berkeley en de groep in Leuven. Elk hoofdstuk is geschreven door iemand van Leuven en iemand van Berkeley. We hebben een contract van Springer voor dit boek.

Karel De Witte

Personalia

Naam: Karel De Witte
Werkadres: Psychologisch Instituut
Tiensestraat 102
3000 Leuven
tel. 016/326062; fax 016/326055; gsm 0495/274041
e-mail: karel.dewitte@psy.kuleuven.ac.be

Vijf relevante publicaties

- De Witte, K., & Andriessen, M. (2002). Werving en selectie: een zootje (on)geregeld? In: P. Vlerick, F. Lievens & R. Claes, *Mens en organisatie: Liber Amicorum Pol Coetsier* (pp. 37-50). Gent: Academia Press.

- De Witte K. (1989). Recruiting and advertising. In P. Herriot (Ed.), *Assessment and selection in Organizations* (pp. 206-217). London: Wiley & Sons.

- De Witte K. (1992). *Occupational Assessment techniques: towards a new approach*. Proceedings of the European Workshop on Psycho-social aspects of Employment, Sofia, Bulgaria, september (dit boek werd ook in het Bulgaars gepubliceerd).

- Derous, E., & De Witte, K. (2001). Looking at selection from a social process perspective: towards a social process model on personnel selection. *European Journal of Work and Organizational Psychology, 10,* 319-342.

- Derous, E., & De Witte, K. (2001). Sociale procesfactoren, testmotivatie en testprestatie: Een procesperspectief of selectie geëxploreerd via een experimentele benadering. *Gedrag en Organisatie, 14, nr.3,* 153-170.

Rianne Janssen

vijf relevante publicaties

- Janssen, R., De Boeck, P., & Vander Steene, G. (1996). Verbal fluency and verbal comprehension abilities in synonym tasks. *Intelligence, 22*, 291-310.

- Janssen, R., & De Boeck, P. (1999). Confirmatory analyses of componential test structure using multidimensional Item Response Theory. *Multivariate Behavioral Research, 34*, 245-268.

- Janssen, R., De Boeck, P., Viaene, M., & Vallaeys, M. (1999). Simple mental addition in children with and without mild mental retardation. *Journal of Experimental Child Psychology, 74*, 261-281.

- Janssen, R., De Corte, E., Verschaffel, L. Knoors. E., & Colémont, A. (2002). National assessment of new standards for mathematics in elementary education in Flanders. *Educational Research and Evaluation,* 8, 197-225.

- Janssen, R., Tuerlinckx, F., Meulders, M., & De Boeck, P. (2000). A hierarchical IRT model for criterion-referenced measurement. *Journal of Educational and Behavioral Statistics, 25*, 285-306.

Lopende onderzoeksprojecten

- Janssen, R., De Corte, E., De Boeck, P., Verschaffel, L. , Daems, Fr. *Peilingsonderzoek wiskunde en begrijpend lezen in het basisonderwijs.* Onderzoeksproject in opdracht van het Departement Onderwijs van het Ministerie van de Vlaamse Gemeenschap, lopende van 1 december 2001 tot en met 30 juni 2003 en gefinancierd voor een totaalbedrag van 9.889.972 Bfr.

- Janssen, R., Crauwels, M., Laevers, F., & De Meuter, F. *De constructie van een peilingsinstrument wereldoriëntatie (domein natuur) voor het basisonderwijs.* OBPWO-project 01.08, lopende van 1 september 2002 tot en met 31 augustus 2003 en gefinancierd voor een totaalbedrag van € 247 982.

- Janssen, R. , Crauwels, M., Van Damme, J., & De Meuter, F. *De constructie van een peilingsinstrument biologie voor de A-stroom van de eerste graad secundair onderwijs.* OBPWO-project 01.07, lopende van 1 augustus 2002 tot en met 31 juli 2003 en gefinancierd voor een totaalbedrag van € 248 796.

- Janssen, R., Daems, Fr., Janssens, D., Verschaffel, L., & Van Damme, J. *De ontwikkeling van een begin- en eindtoets wiskunde en Nederlands voor de eerste graad secundair onderwijs.* OBPWO-project 2002, lopende van 1 oktober 2002 tot en met 31 maart 2004 en gefinancierd voor een totaalbedrag van € 328 333.50.

BIJLAGE: UITGEBREIDE OMSCHRIJVING VAN HET PROJECT

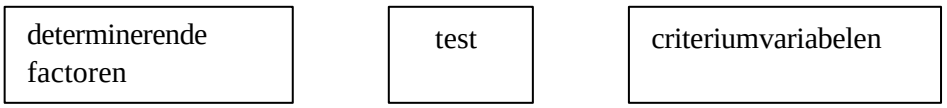
Dit project bevat een onderzoeksplan in antwoord op de VIONA-onderzoeksoproep met als onderwerp "Psychologische testen en de effecten op de instroom van kansengroepen in het Ministerie van de Vlaamse Gemeenschap en in de Vlaamse privébedrijven". De bedoeling van het project is te onderzoeken of relevante intelligentietests een bias vertonen voor elk van drie kansengroepen, namelijk vrouwen, allochtonen en gehandicapten.

Een test of item uit een test vertoont een bias als naast de variabele die men wil meten ook de groep waartoe men behoort het resultaat van een persoon voor een item of voor de gehele test bepaalt. Bij gelijke intelligentie moeten de succeskansen dezelfde zijn, ongeacht de groep waartoe men behoort. De bias kan bestudeerd worden voor individuele items of voor de test in zijn geheel. Veronderstel dat twee personen even intelligent zijn, maar dat de persoon die tot de allochtone groep hoort voor een bepaald item een kleinere kans heeft op een juist antwoord dan de persoon die tot de autochtone groep hoort. Als dit zich voordoet bij een bepaald item, dan werkt het betreffende item discriminerend. Het item vertoont dan "itembias". Het functioneert anders in de onderscheiden subgroepen. In het Engels wordt dit "differential item functioning" of kortweg DIF genoemd. Het item zou dan eigenlijk verwijderd moeten worden als men de twee groepen gelijke kansen wil geven. Als een test een bias vertoont voor verschillende items dan is het interessant om "testbias" (de bias voor de gehele test) of "differential test functioning" (DTF) te onderzoeken. Testbias kan op verschillende manieren geconceptualiseerd worden: (1) In de klassieke testtheorie (KTT) spreekt men van testbias als testscores van verschillende subgroepen een verschillende predictieve validiteit hebben voor het criterium dat men beoogt te meten met de test. (2) In het kader van de itemrespons theorie (IRT) definieert men testbias als het effect van bias in individuele items op de testscores van subgroepen (Shealy en Stout, 1993). (3) In dit project wordt nog een derde manier toegevoegd om testbias te onderzoeken in het kader van IRT-modellen, namelijk als constante itembias in elk van de items van een test. Beide opvattingen over testbias in IRT worden in dit project onderzocht. Tenslotte merken we op dat in de literatuur de term "bias" soms wordt gereserveerd voor gevallen waarin de constructvaliditeit van de test gewaarborgd is (zie Shealy en Stout, 1993). We zullen in het vervolg van de tekst dit onderscheid niet expliciet maken en de termen DIF (DTF) en itembias (testbias) als synoniemen beschouwen voor hetzelfde statistische fenomeen.

In de volgende paragrafen schetsen we eerst het conceptuele kader van bias bij intelligentietests die gebruikt worden voor personeelsselectie. Vervolgens beschrijven we drie verschillende manieren om van testscores gebruik te maken, namelijk een pragmatische, een theoretische en een validiteitsbewakende aanpak. Daarna beschrijven we de conceptuele en de praktische beperkingen van het biasonderzoek in dit project. Tenslotte beschrijven we de verschillende fasen van dit onderzoeksplan.

Conceptueel kader

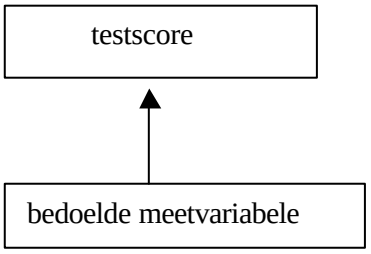
Het conceptueel kader van de studie van tests kan als volgt beschreven worden:
a. Er wordt een onderscheid gemaakt tussen determinerende factoren, de test en criteriumvariabelen.



De *determinerende factoren* zijn variabelen zoals vooropleiding, vertrouwdheid met de taal, motivatie, genetische aanleg, de maatschappelijke groep waartoe men behoort (man of vrouw, allochtoon of autochtoon, met of zonder handicap, enz.). De *test* bestaat uit een reeks opgaven of vragen, items genoemd. Doorgaans wordt de som bepaald van de itemscores (bijvoorbeeld de som van het aantal juiste antwoorden) en wordt die "ruwe uitslag" genoemd. Soms wordt die ruwe score omgezet in een afgeleide uitslag op basis van een normering. De *criteriumvariabelen* zijn de variabelen waarin men is geïnteresseerd. Het zijn de variabelen die men wil voorspellen of verklaren. Intelligentietests worden vaak aangewend als predictoren. Ondermeer voor schoolsucces, succes in een job of in de loopbaan. Als men bijvoorbeeld iedereen die een score heeft lager dan een kritische grens verder niet in aanmerking neemt, dan neemt men aan dat wie lager scoort dan die kritische grens slecht zou presteren. Het is ook mogelijk dat de testscores zelf of de variabele die ze meten een causale invloed hebben op andere variabelen. Het gaat dan om predictoren met een causale rol.

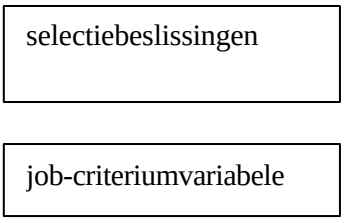
b. Een test is altijd bedoeld om een bepaalde persoonskarakteristiek te meten: de variabele die de test bedoelt te meten, verder ook de *bedoelde meetvariabele* genoemd. Die karakteristiek hoeft geen onveranderlijke karakteristiek te zijn, het kan ook gaan om een niveau dat men tijdelijk heeft bereikt of een toestand waarin men tijdelijk vertoeft. Idealiter wordt de testscore geheel bepaald door de bedoelde meetvariabele, maar in de praktijk is het meestal slechts voor een substantieel deel dat de testscore bepaald

wordt door de bedoelde meetvariabele. Hoe groter dat deel, des te groter de constructvaliditeit van de test, dat wil zeggen, des te sterker sluit de test aan bij het construct dat men wil meten.

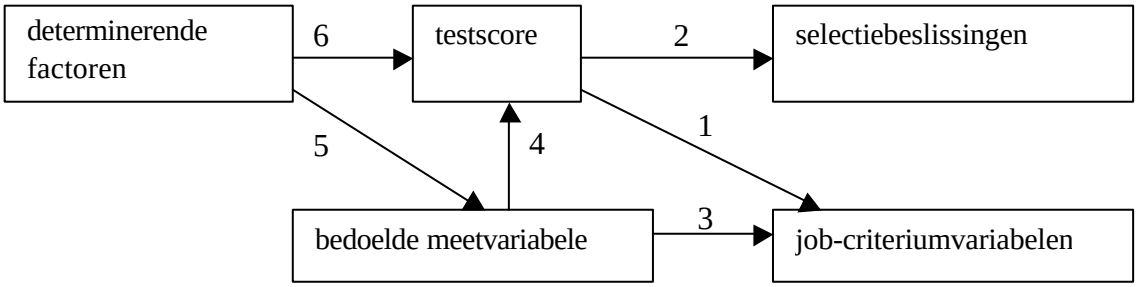


Men kan de band met de bedoelde meetvariabele per item bekijken. In principe speelt de bedoelde meetvariabele een rol in elk item, maar naargelang van het item kan die variabele sterker of minder sterk doorwegen. Het gewicht van de bedoelde meetvariabele in een item noemt men de “discriminatiegraad” van het item. Hoe groter de *discriminatiegraad* des te sterker differentieert het item tussen hoge en lage waarden van de bedoelde meetvariabele en, des te beter is het item als indicator van de bedoelde variabele. Items hebben naast hun discriminatiewaarde ook nog een moeilijkheidsgraad. Voor juist/fout items is de *moeilijkheidsgraad* het niveau van de bedoelde variabele (de intelligentie) dat nodig is om één kans op twee te hebben om het item juist te beantwoorden.

c. In de context van personeelsselectie zijn de twee belangrijke types van criteriumvariabelen: (1) *selectiebeslissingen*, zoals de preselectie en de eigenlijke selectie, en (2) gedrag in de job, zoals bijvoorbeeld het prestatieniveau, promoties, het verlaten van de job, enz. , die samen de *job-criteriumvariabelen* worden genoemd. Alleen van wie geselecteerd wordt, kan men de waarde op de job-criteriumvariabelen bepalen.



d. Om een volledig beeld te krijgen van de rol die tests kunnen spelen, moet ten eerste het onderscheid tussen de testscore en de bedoelde variabele worden ingebouwd in het schema tussen de determinerende factoren en de criteriumvariabelen. Ten tweede moeten de twee types van criteriumvariabelen onderscheiden worden. Op basis daarvan kan men een globaal schema opstellen met de mogelijke invloeden tussen de verschillende bouwstenen.



Een pragmatische, theoretische en validiteitsbewakende visie op testscores

Een *zuiver pragmatische* aanpak bestaat er in om een test te gebruiken omdat de testscore empirisch predictief is voor de job-criteriumvariabelen waarin men geïnteresseerd is. Men doet dan een beroep op de band die in het schema is aangegeven door pijl 1 die de predictierelatie aangeeft. Pijl 1 geeft de empirische validiteit weer van de test. Er is geen verantwoording van de pijl nodig op grond van een hypothese of theorie. Alleen de feitelijke predictieve waarde van de testscore is van belang. Op grond van de empirische predictierelatie wordt de testscore medebepalend voor de selectiebeslissing, zoals weergegeven door pijl 2. Deze zuiver pragmatische aanpak kan men blind volgen, zonder enige hypothese of theorie. Alleen de pijlen 1 en 2 spelen een rol.

Een *theoretisch geïnspireerde aanpak* bestaat er in om een beroep te doen op hypothesen of een theorie over welke de variabelen zijn die een rol spelen in de job-criteriumvariabelen. De hypothese of theorie betreft de persoonskarakteristieken die bevorderlijk of hinderlijk zijn in de job of de loopbaan. Bijvoorbeeld, bij jobs voor hoger opgeleiden wordt dikwijls aangenomen dat er een minimum aan intelligentie nodig is, naast persoonlijkheidseigenschappen en motivatie. Afhankelijk van de job neemt men aan dat een hogere intelligentie beter is. In een meer gedifferentieerde aanpak bepaalt men ook welke soorten van intelligentie van belang zijn voor de betreffende job of loopbaan. Op basis van dergelijke hypothesen, weergegeven in pijl 3, kiest men bedoelde meetvariabelen en voor deze bedoelde meetvariabelen kiest men tests die constructvaliditeit hebben voor die variabelen, weergegeven in pijl 4. In de theoretisch geïnspireerde aanpak spelen dus ook hypothesen (en theorie) over de job en/of de loopbaan een rol, alsook de constructvaliditeit van tests. Op grond van de pijlen 3 en 4 verwacht men dat de testscore predictief is (pijl 1) en zal de testscore medebepalend zijn voor de selectiebeslissing (pijl 2). Idealiter wordt de predictieve waarde van de testscore ook in de feiten nagegaan, maar dat gebeurt niet

altijd. Soms stelt men zich tevreden met de theoretische ondersteuning. Samengevat, spelen in de theoretisch geïnspireerde aanpak de pijlen 1, 2, 3 en 4 een rol.

De basis van dit project is een derde aanpak: de *validiteitsbewakende* aanpak. De aandacht gaat daarbij naar de pijlen 4, 5 en 6. Idealiter verlopen alle invloeden van de determinerende factoren op de testscore via de bedoelde meetvariabele. Dat wil zeggen, als iemand een lagere score haalt op een intelligentietest, dan mag dat alleen maar een reflectie zijn van een lagere intelligentie en niet van iets anders, zoals bijvoorbeeld van de maatschappelijke groep waartoe men behoort, of van de motivatie. Elke invloed op de testscore buiten de bedoelde meetvariabele om is een bedreiging van de constructvaliditeit (pijl 4), want dan spelen naast die bedoelde meetvariabele ook nog andere factoren een rol. Een bedreiging van de validiteit die speciale aandacht vraagt is dat de groep waartoe men behoort een rol speelt in de testscore, los van de bedoelde meetvariabele. Er is dan immers sprake van discriminatie. Pijl 5 geeft de invloed weer van de determinerende factoren op de bedoelde meetvariabele. Pijl 6 geeft de invloed weer van de determinerende factoren op de testscore. De aanwezigheid van pijl 6 is een bedreiging van de validiteit en houdt een discriminatie in als de determinerende variabele betrekking heeft op de groep waartoe men behoort. Het probleem van DIF en DTF heeft betrekking op pijl 6. De invloed op (de items van) een test kan twee vormen aannemen: een differentiële moeilijkheidsgraad of een differentiële discriminatiegraad. Een differentiële moeilijkheidsgraad betekent dat bepaalde items moeilijker zijn voor de ene groep dan voor de andere. Een differentiële discriminatiegraad betekent dat voor bepaalde items de bedoelde meetvariabele een verschillend gewicht heeft naargelang van de groep. Het is bijvoorbeeld mogelijk dat een item in de ene groep wel een indicator is van intelligentie en in een andere groep niet, of een minder goede indicator. De oorzaak van deze twee vormen van bias (moeilijkheid en discriminatie) kan velerlei zijn: een ander soort voorkennis, een minder goede taalbeheersing, een andere belangstelling. In de validiteitsbewakende aanpak onderzoekt men of naast pijl 4 en 5 niet ook pijl 6 een rol speelt.

Conceptuele en praktische afbakening van biasonderzoek

Het geschetste kader heeft twee belangrijke implicaties die betrekking hebben of de aflijning van bias-onderzoek:

a. Bias-onderzoek handelt niet over de invloed die met pijl 5 is weergegeven. Het is mogelijk dat twee bevolkingsgroepen verschillen inzake de meetvariabele zonder dat er van bias sprake is, d.w.z. zonder dat pijl 6 een rol speelt. Als de bedoelde meetvariabele intelligentie is, dan zou dit betekenen dat de ene bevolkingsgroep intelligenter is dan de andere. Vermoedelijk is er een verklaring voor een dergelijk

verschil, zoals geringere kansen in het onderwijs, een minder intellectuele opvoeding, genetische aanleg, en dergelijke, maar hoe belangrijk deze invloeden (pijl 5) ook zijn, we rekenen ze niet tot het bias-onderzoek (wel tot de differentiële psychologie van de intelligentie). Als men het onderscheid tussen de twee pijlen niet zou maken, dan leidt dat tot onduidelijkheid met als risico dat men niet op de juiste bal speelt als men aan de effecten iets wil doen. Ongewenste effecten die op pijl 6 betrekking hebben kan men oplossen door de tests aan te passen. Ongewenste effecten die op pijl 5 betrekking hebben vergen veel meer, bijvoorbeeld een verandering van het onderwijs.

b. Bias-onderzoek kan gebeuren zonder kennis te hebben van selectiebeslissingen of resultaten die geselecteerden behalen in de job of de loopbaan. Het gaat immers alleen maar om de pijlen 4, 5 en 6: het linkse gedeelte uit het schema. Men kan één of meer tests op bias onderzoeken ongeacht wat er verder aan beslissingen op de test volgt en wat de predictieve waarde is van de testscore. Dat een test vrij is van DIF en DTF is een belangrijke verworvenheid die noodzakelijk goede gevolgen heeft voor de selectiepraktijk.

We hebben in het geschetste kader geen aandacht gegeven aan de mogelijkheid dat de pijlen in het rechtergedeelte van het schema (1,2 en 3) zelf verschillen naargelang van de groep. Toch kunnen ook dergelijke verschillen voor discriminatie zorgen. Bijvoorbeeld, als een vrouw hogere testscores zou moeten halen dan een man om aangeworven te worden, dan is er een verschil in pijl 2 met discriminatie als gevolg. Dergelijke praktijken komen voor en zijn afkeurenswaardig, maar we rekenen onderzoek daarover niet tot het bias-onderzoek. Het is ook mogelijk dat de bedoelde meetvariabele een ander verband vertoont met prestaties in de job of met het verloop van de loopbaan (een verschil in pijl 3 en dus ook in pijl 1), bijvoorbeeld omdat er verschillende manieren zijn om een job uit te voeren: alternatieve manieren om succes te halen (bijvoorbeeld een mannelijke en een vrouwelijke). Ook dit is een interessant en belangrijk probleem, maar ook dat probleem rekenen we niet tot het bias-onderzoek. Geen van beide voorbeelden heeft betrekking op de validiteit van de test.

We hebben niet alleen conceptuele redenen om het onderwerp af te bakenen maar ook twee soorten praktische redenen. De eerste praktische reden heeft betrekking op de remediëring. Als men de geschetste validiteitsbewakende aanpak volgt, kan men zeer doelgericht bepaalde vormen van discriminatie uitschakelen met een grote kans op succes, namelijk die vormen van discriminatie die rechtstreeks betrekking hebben op de tests. De andere vormen van discriminatie vergen een maatschappelijke hervorming die de beperktheid van het project overstijgt. De tweede praktische reden is dat de beschikbaar gestelde middelen een gerichte en afgelijnde benadering vergen om tot concrete tastbare resultaten te komen. Deze keuze betekent geen onderschatting van de andere problemen. Ze is slechts door realisme

ingegeven.

Beschrijving van het onderzoeksplan in vier fasen

Het voorgestelde onderzoeksplan bevat vier fasen: (1) bepalen welke intelligentietests dienen onderzocht te worden, (2) onderzoeken of deze tests een bias vertonen voor de relevante doelgroepen, (3) in geval er een bias wordt vastgesteld, een verklaring zoeken hiervoor, (4) aangeven hoe de bias kan verwijderd worden of hoe de test eventueel kan worden aangepast.

Eerste fase

In de eerste fase van het onderzoek wordt bepaald welke intelligentietests het meest nuttig zijn voor bias-onderzoek bij een bepaalde doelgroep. Met andere woorden, het doel van deze fase is het kiezen van enkele "relevante" tests uit de ganse populatie van intelligentietests. De criteria voor de keuze van tests zijn drievoudig:

a. Opdat een test betrokken zou worden in het onderzoek moet het soort van intelligentie dat wordt gemeten belangrijk zijn in de selectieprocedures van de overheid en van andere organisaties. Wij willen ons niet fixeren op de frequentie van de tests zelf om de volgende redenen. Ten eerste worden er in vele gevallen "eigen" tests gebruikt, gemaakt of aangepast door het selectiebureau of het bedrijf in kwestie. Ten tweede is het belangrijk om te kunnen generaliseren. Het leek ons een interessantere aanpak om tests te kiezen als representanten van een type van tests, zodat de resultaten maximaal gegeneraliseerd kunnen worden naar andere tests. Deze aanpak impliceert ook een bepaalde keuze voor het bias-onderzoek zelf, door het niet in de eerste plaats te concentreren op de detectie van problematische individuele items, maar door gebruik te maken van itemkarakteristieken en verklaringsgericht onderzoek (zie tweede fase).

b. Een tweede belangrijke factor is de frequentie waarmee een type van test in het algemeen afgenomen wordt bij personen van de doelgroepen. Meer bepaald, in de mate dat zou blijken dat allochtonen relatief meer deelnemen aan selecties waar een lagere scholingsgraad is vereist, moet het onderzoek ook toegespitst worden op de intelligentietests die vooral gebruikt worden bij deze selectieprocedures.

c. Het moet mogelijk zijn om voor de belangrijke doelgroepen gebruik te maken van gegevens die al beschikbaar zijn of relatief gemakkelijk verzameld kunnen worden zonder apart onderzoek op te zetten. De ecologische validiteit vereist dat de gegevens representatief zijn voor de selectiepraktijk. Een

afzonderlijk opgezet onderzoek betekent dat men met het onderzoek buiten de selectiepraktijk gaat staan, zodat het betwistbaar is of de resultaten ook gelden voor die selectiepraktijk. Een bijkomende reden om geen afzonderlijk onderzoek op te zetten is dat er relatief grote aantallen personen nodig zijn om bias-onderzoek te doen. Meerbepaald, bij Educational Testing Service (ETS) in de Verenigde Staten geldt als regel dat bij biasonderzoek de kleinste van de te vergelijken groepen minstens 200 personen moet bevatten (Zieky, 1993).

Om aan de eerste twee criteria tegemoet te komen zal in een *eerste stap* nagegaan worden welke tests gebruikt worden bij de selectie van overheidspersoneel en bij de selectie van werknemers in Vlaamse privéondernemingen. Als onze onderzoeksgroep wordt aangewezen om het project uit te voeren, zal contact genomen worden met SELOR, jobpunt en de VDAB. Daarnaast zullen we ook contact nemen met organisaties die bedrijvig zijn in de personeelsselectie. Via Karel De Witte en Paul De Boeck hebben we goede contacten met verschillende van die organisaties, bijvoorbeeld dankzij de constructie van intelligentietests of dankzij onderzoek over intelligentietests in samenwerking met die organisaties. In een *tweede stap* zal nagegaan worden welk type van intelligentie wordt gemeten in die tests. We zullen daarvoor een beroep doen op twee algemeen erkende classificatiesystemen die gebaseerd zijn op factoranalytisch onderzoek: (a) het systeem van de French kit (Ekstrom, French, Harman, & Dermen, 1976) dat ontwikkeld is in ETS en (b) het systeem van Caroll (1993), die het meest volledige overzicht ooit heeft gemaakt van factoranalytisch onderzoek over intelligentie. We beschikken over een gecombineerd systeem om op grond van het onderzoek dat is uitgevoerd door Ekstrom e.a. (1976) en Carroll (1993) uit een analyse van de opdrachten van gelijk welke intelligentietest af te leiden welke intelligentie in een test wordt gemeten, zonder de tests zelf ook daadwerkelijk te moeten afnemen (De Boeck & Bijttebier, 2001). Dit is mogelijk gebleken omdat de variëteit van tests die door Ekstrom e.a. en door Carroll is onderzocht allesomvattend is. Speciaal het onderzoek van Carroll (1993) is ongelooflijk breed. Het boek dat hij heeft geschreven is monnikenwerk en het resultaat van een volledige carrière.

Om aan het derde criterium tegemoet te komen zullen we samenwerking zoeken om gebruik te maken van bestaande gegevens en om nieuwe gegevens te verzamelen. In de meeste gevallen beschikt men voor reeds verzamelde testgegevens wel over informatie over het geslacht, maar niet over informatie over de culturele achtergrond. Voor informatie over handicaps is het onderzoeksprobleem nog veel groter omdat er gewoon weinig gegevens (kunnen) zijn uit lopende selectieprocedures. Voor wat de culturele achtergrond betreft (allochtoon/autochtoon), stellen we voor om met garanties voor de privacy en op vrijwillige basis de nodige informatie te verzamelen. Dat wil zeggen, er zal aan de kandidaten gevraagd

worden om hun culturele achtergrond op vrijwillige basis kenbaar te maken. Er zal aan de kandidaten worden uitgelegd
(a) dat de informatie wordt gevraagd om onderzoek mogelijk te maken dat discriminatie tegengaat,
(b) dat de informatie niet wordt doorgegeven aan de organisatie, maar slechts aan de onderzoekers.

Voor wat personen met een handicap betreft, vrezen we dat het aantal personen per type van handicap te klein zal zijn voor een degelijk onderzoek. Daarom stellen we voor om het probleem in de eerste plaats theoretisch te onderzoeken via een literatuurstudie en om het empirisch gedeelte te beperken tot het nagaan van wat men "person fit" noemt (Meijer & Sijtsma, 2001), omdat die methode bedoeld is voor individuele personen en dus toepasselijk bij kleine aantallen. Een lage person fit betekent dat voor de betreffende persoon andere factoren een rol spelen bij de antwoorden dan voor de meerderheid. Om informatie over handicaps te krijgen zullen we dezelfde werkwijze volgen als voor de informatie over de culturele achtergrond.

Tweede en derde fase

Wanneer we over de nodige gegevens beschikken, bestaan deze fasen er uit dat de DIF en DTF worden beschreven (tweede fase) en verklaard (derde fase). We bespreken de tweede en de derde fase samen omdat ze intrinsiek verbonden zullen uitgevoerd worden. We streven ernaar om via twee uitgebreide selectieprocedures (één uitgevoerd bij de overheid en één uitgevoerd in de privé) met grote aantallen kandidaten gegevens van drie intelligentietests (drie of meer subtests) te bekomen die afgenomen zijn van dezelfde kandidaten.

DIF-onderzoek

In het algemeen kunnen we stellen dat een item een bias vertoont (of partijdig is) ten nadele van de doelgroep als blijkt dat, in de populatie van personen met gelijke score op de trek die de test beoogt te meten (de bedoelde meetvariabele), personen die niet tot de doelgroep behoren (ook referentiegroep genoemd) een hogere kans hebben om het item juist op te lossen. Voor DIF-onderzoek bestaan er verschillende methodes, zowel vanuit de KTT (Holland & Thayer, 1988) als vanuit de IRT (Kelderman, 1989; Mellenbergh, 1982, 1985, 1989; Bügel & Glas, 1991). Millsap en Everson (1993) geven een overzicht van procedures voor biasdetectie. Welkenhuysen-Gybels en Billiet (2002) vergelijken verschillende methoden voor biasdetectie in de context van cross-cultureel onderzoek.

In de KTT neemt men aan dat het aantal correct opgeloste opgaven een goede schatting is van de trek die men beoogt te meten. Bij de Mantel-Haenszel techniek (Holland & Wainer, 1993), de populairste methode uit de KTT, worden leerlingen van de onderscheiden subgroepen (bijvoorbeeld allochtonen en autochtonen) op basis van de totaalscore op de test ingedeeld in niveaugroepen. Vervolgens wordt de hypothese getoetst of binnen elk van de niveaugroepen de subgroepen dezelfde *p*-waarde (proportie juist) hebben. Omdat de totaalscore kan gebaseerd zijn op partijdige items kan de indeling in niveaugroepen onnauwkeurig zijn. Daarom wordt in herhaalde analyses een uitgezuiverde totaalscore berekend op basis van alleen maar onpartijdige items. Bij het einde van de procedure is het belangrijk om na te gaan of de onpartijdige items nog steeds een goede inhoudelijke beschrijving geven van het construct dat de oorspronkelijke test beoogde te meten. Anders gezegd, men dient de inhoudsvaliditeit van de onpartijdige items na te gaan. Een belangrijk nadeel van de Mantel-Haenszel procedure is dat er geen statistische test wordt uitgevoerd om na te gaan of de totaalscore wel een goede schatting is van de latente trek. Deze assumptie kan wel getoetst worden met IRT procedures. De Mantel-Haenszel techniek kan ook verantwoord worden vanuit de IRT, namelijk als het Rasch model blijkt op te gaan voor de onpartijdige items. Verguts en De Boeck (2001) ontwikkelden een uitbreiding van de techniek.

We zullen in dit project hoofdzakelijk gebruik maken van de procedures uit de IRT, omdat deze benadering de meeste mogelijkheden biedt en meer verfijnde analyses mogelijk maakt. In modellen van het IRT-type, verder IRT-modellen genoemd, wordt het geheel van itemantwoorden verklaard op basis van continue latente variabelen. Dit zijn de bedoelde meetvariabelen van de test. In de context van intelligentietests kunnen deze latente variabelen geïnterpreteerd worden als de soorten intelligentie die een test meet. Bij unidimensionele IRT modellen volstaat één latente variabele om de itemantwoorden te beschrijven, terwijl bij multidimensionele IRT modellen meerdere latente variabelen nodig zijn. IRT modellen geven een beschrijving van de kans dat een persoon een bepaald item van de test juist oplost. Meer bepaald wordt bij Rasch modellen verondersteld dat de kans op een juist antwoord een logistische functie is van het verschil tussen de vaardigheid van de persoon en de moeilijkheidgraad van het item (Fischer & Molenaar, 1995). Bij multidimensionele modellen wordt de vaardigheid van de persoon bijvoorbeeld opgevat als een gewogen som van de latente variabelen. Wij zullen de IRT-modellen op twee manieren gebruiken: een itemsgewijze manier en een verklarende manier.

De *itemsgewijze methode* bestaat er in om in het model beurtelings per item een interactie toe te laten tussen de groepsvariabele (man/vrouw en allochtoon/autochtoon) waartoe iemand behoort en het item, betreffende (a) de moeilijkheidsgraad, en (b) de discriminatiegraad indien mogelijk. (Interacties voor de discriminatiegraad kunnen in principe in het model opgenomen worden, maar in de praktijk leidt de

schatting soms tot problemen). Als de interactie statistisch significant is, dan is er sprake van DIF: "uniforme" DIF als de interactie slechts de moeilijkheidsgraad betreft (Mellenbergh, 1982), en niet-uniforme DIF als de interactie (ook) de discriminatiegraad betreft (Mellenbergh, 1982). De term "uniform" verwijst ernaar dat de verstoring uniform is (d.w.z. een constant additief effect heeft) over het gehele bereik van de latente variabele, hetgeen niet het geval is als de discriminatiegraad verschilt. Bijkomend zullen we nagaan of binnen de groepen individuele verschillen zijn in DIF, door het interactie-effect te modelleren als een zogenaamd random effect, d.w.z. een effect met individuele verschillen die een bepaalde verdeling volgen. Deze modellen zijn beschreven en toegepast door Van den Noortgate en De Boeck (2002). Het is mogelijk dat dergelijke individuele verschillen correleren met de latente trek, zodat de DIF in dat geval opnieuw niet-uniform zou zijn.

De *verklarende methode* bestaat er in om bij het modelleren niet itemsgewijs te werken, maar om naar itemkenmerken te zoeken die ofwel gemeenschappelijk zijn voor de items met DIF, of om itemkenmerken te gebruiken die men op voorhand bepaalt door analyse van de items. Het gaat om een verklarende methode omdat de DIF op die manier aan kenmerken van de items kan worden toegeschreven. De eerste werkwijze vertrekt van de resultaten van de itemsgewijze aanpak. Deze werkwijze wordt beschreven door Meulders en Xie (2002). Voor de tweede werkwijze moeten de items eerst geanalyseerd worden volgens de kenmerken die ze vertonen, bijvoorbeeld het belang van de complexiteit van de taal die wordt gebruikt in de uitleg per item, het visuele karakter van het item, het aantal inferentieregels dat moet afgeleid en/of bijgehouden worden, enz. Dit vergt een kennis van de cognitieve psychologie van items uit intelligentietests. Het is bijvoorbeeld mogelijk dat oppervlakkig bekeken items niet visueel lijken te zijn, terwijl men ze wel oplost door zich dingen visueel voor te stellen. Dit is bijvoorbeeld het geval bij items voor deductief redeneren waarbij lineaire syllogismen worden gebruikt (Paul is groter dan Michel, Karel is groter dan Paul, is Karel groter dan Michel?), zoals aangetoond door Sternberg (1980). In onze eigen onderzoekgroep hebben we cognitieve analyses gemaakt van deductieve items (Rijmen & De Boeck, 2001), van inductieve items (Verguts, Maris & De Boeck, 2002) en van verbale intelligentie-items (Janssen, De Boeck & Vander Steene, 1996). Voor de itemkarakteristieken die worden afgeleid uit de itemsgewijze resultaten of uit een cognitieve analyse van de items, worden er in het model interacties opgenomen met de groepsvariabele (man/vrouw, autochtoon/allochtoon), in de eerste plaats voor de moeilijkheidsgraad en indien mogelijk ook voor de discriminatiegraad (zie een eerdere bemerking). Een dergelijke interactie noemen we "differential feature functioning" (DFF) omdat een itemkenmerk (item feature) een effect heeft dat afhangt van de groep. Er zal nagegaan worden of de DFF statistisch significant is en hoeveel beter het model met DFF is. Ook voor de DFF kan men nagaan of er binnen de

groepen individuele verschillen bestaan. Voor een beschrijving van DFF-modellen, met en zonder individuele verschillen in DFF, zie Meulders en Xie (2002).

DTF-onderzoek

In feite is het DTF-onderzoek formeel equivalent met DFF onderzoek, omdat het behoren tot een (sub)test een kenmerk is van de items. Men kan inderdaad de items van verschillende (sub)tests samen modelleren en de (sub)test waartoe het item behoort gebruiken als itemkenmerk. In deze opvatting wordt DTF beschouwd als gelijke DIF in alle items van een test.
De modellen die men voor dit onderzoek nodig heeft zijn multidimensioneel, omdat men mag aannemen dat naargelang van de (sub)test een andere dimensie (bedoelde meetvariabele) gemeten wordt. In compensatorische modellen (het meet courante type) kan men in de moeilijkheidsgraad geen onderscheid maken tussen de dimensies. De moeilijkheidsgraad is gewoon het hoofdeffect van het item. Daarom is het mogelijk om op een relatief eenvoudige wijze de interactie te bepalen tussen de groepsvariabele (man/vrouw, autochtoon/allochtoon) en het itemkenmerk dat erin bestaat dat het item tot deze of gene (sub)test behoort. Het verschil met modellen voor DFF is alleen maar dat modellen voor DTF multidimensioneel zijn. Het gaat evenwel niet om moeilijke multidimensionele modellen omdat de dimensie waartoe een item behoort niet meer geschat moet worden, want het is de test die bepalend is voor de dimensie. Een andere opvatting over DTF, die aan bod komt in de vierde fase van het onderzoek, beschouwt DTF als het effect van individuele items op de testscores (Shealy & Stout, 1993).

Als er zich inzake DIF, DFF of DTF individuele verschillen binnen de groepen voordoen en als er andere gegevens (bijvoorbeeld, andere testuitslagen) beschikbaar zijn, dan is het mogelijk om de DIF, DFF en DTF te koppelen aan deze gegevens, zodat de verklaring niet alleen berust op itemkenmerken, maar ook op kenmerken van de personen, bijvoorbeeld zoals die uit andere testgegevens blijken.

<u>Vierde fase</u>

De vierde fase bestaat er in dat wordt aangegeven hoe de bias verwijderd kan worden of hoe de test kan aangepast worden. Vanuit praktisch oogpunt is het interessant om eerst te berekenen hoe groot het effect van de bias is, d.w.z. hoe groot het effect is in termen van het aantal punten of in termen van het percentage variantie van de testscores dat wordt verklaard door de bias. Als het effect groot is, dan gaat het om een prangend probleem.

Voor *bias van individuele items* ligt de oplossing voor de hand. De afwijkende items kunnen uit de test verwijderd worden. Een minder drastische oplossing bestaat er in om de betreffende items niet te verrekenen in de testscore.

Als de *bias van het DFF-type is*, dan gaat het om een algemeen probleem dat vermoedelijk niet beperkt is tot de onderzochte tests. De bias is dan immers gekoppeld aan kenmerken van items en die kenmerken komen meestal ook voor bij items van andere tests van hetzelfde intelligentietype en eventueel nog breder. Omwille van deze generalisatiemogelijkheid, is DFF zo belangrijk. De remedie bestaat er dan in om de test aan te passen zodat die alleen nog uit items bestaat met kenmerken die geen DFF vertonen. Het zal dan meestal nodig zijn om items bij te maken ter vervanging van de items met de problematische kenmerken. Er zal afhankelijk van het concrete geval dat zich voordoet een advies geformuleerd worden voor het aanpassen van de onderzochte tests en meer algemeen voor tests van hetzelfde type.

Voor *bias van het DTF-type* zal eerst onderzocht worden of de bias afhangt van bepaalde itemkenmerken. Zonder aanwijzing van itemkenmerken die een rol spelen is er immers niet veel meer mogelijk dan het vermijden van de betreffende test en gelijkaardige tests. Als bepaalde kenmerken wel een rol spelen, dan kan men proberen om een nieuwe test te maken waarin de problematische kenmerken gereduceerd worden zonder het essentieel karakter van de test aan te tasten. Dat zal evenwel niet in alle gevallen mogelijk zijn. Daarom zal ook hier, afhankelijk van het concrete geval, een advies geformuleerd worden voor de te volgen weg.

Er is één soort van bias waar we machteloos tegenover staan en dat zelfs niet opgemerkt kan worden. Elke vorm van bias die we kunnen ontdekken berust op verschillen tussen groepen die niet gelden voor alle items of voor alle tests. Het is in principe mogelijk dat voor alle intelligentietests en in het bijzonder voor alle tests die we in het onderzoek betrekken de bias even groot is. Omdat een dergelijke bias niet onderscheiden kan worden van het effect dat met pijl 5 is weergegeven is het niet mogelijk om die bias te identificeren als bias die op de invloeden via pijl 6 berusten. Als de ene groep systematisch, ongeacht het item en ongeacht de test, eenzelfde verschil vertoont inzake kans op succes, dan zijn er twee mogelijke hypothesen: de groepen verschillen echt van elkaar op de bedoelde meetvariabele(n), of de bias is algemeen en voor alle items en tests dezelfde. Dat de bias algemeen zou zijn is misschien niet onwaarschijnlijk, maar het is wel onwaarschijnlijk dat een algemene bias ongeacht het item en de test precies even groot zou zijn. Daarom voelen we ons relatief goed beveiligd tegen de geschetste blinde vlek in het bias-onderzoek. Onder andere om die beveiliging te verbeteren is het van belang om gegevens te

gebruiken van verschillende tests die van dezelfde personen zijn afgenomen. Met slechts één test is het gevaar nog relatief groot, vooral als de test relatief homogeen is.

Referenties

Bügel, K., & Glas, C. (1991). Item-specifieke verschillen in prestaties van jongens en meisjes bij tekstbegripexamens moderne vreemde talen. *Tijdschrift voor Onderwijsresearch, 16*, 337-351.

Carroll, J.B. (1993). *Human cognitive abilities. A survey of factor-analytic studies*. New York: Cambridge University Press.

De Boeck, P., & Bijttebier, P. (2001). *Principes en methoden van de psychodiagnostiek*. Licentiecursus K.U.Leuven.

Ekstrom, R. B., French, J. W., Harman, H. H., & Dermen, D. (1976). *Manual for kit of factor-referenced cognitive tests.* Princeton, NJ: Educational Testing Service.

Fischer, G. H., & Molenaar, I. W. (1995) (Eds.). *Rasch models: Foundations, recent developments, and applications.* New York: Springer-Verlag.

Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer and H. I. Braun (eds.), *Test validity* (pp. 129-145). Hillsdale: Lawrence Erlbaum.

Holland, P. W., & Wainer, H. (1993) (Eds.). *Differential item functioning.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Janssen, R., De Boeck, P., & Vander Steene, G. (1996). Verbal fluency and verbal comprehension abilities in synonym tasks. *Intelligence*, 22, 291-310.

Kelderman, H. (1989). Item bias detection using loglinear IRT. *Psychometrika, 54*, 681-697.

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person-fit. *Applied Psychological Measurement, 25*, 107-135.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7*, 105-118.

Mellenbergh, G. J. (1985). Vraag-onzuiverheid: definitie, detectie en onderzoek. *Nederlands Tijdschrift voor Psychologie, 40*, 425-435.

Mellenbergh, G. J. (1989). Itembias and itemresponse theory. In R. K. Hambleton (ed.), Applications of itemresponse theory (special issue). *International Journal of Educational Research, 13*, 127-143.

Meulders, M., & Xie, Y. (2002). Person-by-item predictors. To appear in P. De Boeck and M. Wilson (Eds.), *Psychometrics using logistic mixed models*. New York: Springer-Verlag.

Millsap, R. E., & Everson, H. T. (1993). Methodology Review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17*, 297-334.

Rijmen, F., & De Boeck, P. (2001). Propositional reasoning: the different contribution of 'rules' to the difficulty of complex reasoning problems. *Memory and Cognition, 29*, 165-175.

Shealy, R. T., & Stout, W. F. (1993). An item response theory model for test bias and differential test functioning. In P. W. Holland and H. Wainer (eds.), *Differential item functioning.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Sternberg, R. J. (1980). Representation and process in linear syllogistic reasoning. *Journal of experimental Psychology: General, 109*, 119-159.

Van den Noortgate, W., & De Boeck, P. (2002). Assessing and explaining differential item functioning (DIF) using logistic mixed models. Submitted to *Journal of Educational and Behavioral Statistics*.

Verguts, T., & De Boeck, P. (2001). Some Mantel-Haenszel tests of Rasch model assumptions. *British Journal of Mathematical and Statistical Psychology, 54*, 21-37.

Verguts, T., Maris, E., & De Boeck, P. (2002). A dynamic model for rule induction tasks. *Journal of Mathematical Psychology, 46*, 455-485.

Welkenhuysen-Gybels, J., & Billiet, J. (2002). A comparison of techniques for detecting cross-cultural inequivalence at the itemlevel. *Quality and Quantity, 36(3)*, 197-218.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland and H. Wainer (eds.), *Differential item functioning* (pp. 337-347). Hillsdale NJ: Lawrence Erlbaum Associates.