Onderzoek naar bias voor allochtone leerlingen in de Cito-Eindtoets Basisonderwijs

H. Uiterwijk en T. Vallen

Samenvatting

Er is nog relatief weinig onderzoek verricht naar de bruikbaarheid van toetsen voor subgroepen uit de populatie. Het Cito en het Werkverband Taal en Minderheden van de Letterenfaculteit van de KUB verrichten onderzoek naar toets- en itembias in de Cito-Eindtoets Basisonderwijs voor allochtone leerlingen. Uit het onderzoek naar toetsbias blijkt dat het voorspellen van schoolsucces in het voortgezet onderwijs voor allochtone leerlingen minder trefzeker gebeurt dan voor autochtone leerlingen. Dit geldt zowel voor de Eindtoets Basisonderwijs als voor een onderzochte intelligentietest. Het onderzoek naar itembias laat zien dat de statistische procedures voor het opsporen van itembias in de Eindtoets Basisonderwijs verschillende uitkomsten opleveren. De meeste items blijken nooit, andere soms en enkele items blijken altijd gebiast te zijn. Een zoektocht naar de inhoudelijke oorzaken van itembias levert vooral op linguïstisch terrein gegevens op waarmee het risico op itembias in toekomstige toetsen kan worden verkleind.

1 Inleiding en probleemstelling

Het is gebruikelijk om de schoolprestaties van leerlingen te beschrijven aan de hand van scores op een of andere schoolvorderingentoets. Daarbij wordt veelvuldig gerapporteerd over de verschillen tussen de scores van bepaalde groepen in de populatie. In Nederland is er tot nu toe relatief weinig onderzoek gedaan naar de vraag of een toets die bruikbaar geacht wordt voor de hele populatie ook bruikbaar is voor daarbinnen te onderscheiden subgroepen. Als er scoreverschillen tussen bepaalde subgroepen geconstateerd worden, kan dat veroorzaakt worden door het feit dat die subgroepen de vaardigheid die de toets beoogt te meten in uiteenlopende mate beheersen. Scoreverschillen tussen subgroepen zijn op zich geen reden om aan de constructvaliditeit van een toets te twijfelen. Als bijvoorbeeld een taaltoets Nederlands voor allochtone leerlingen moeilijker is dan voor autochtone, dan wordt meestal voldaan aan de functie van de toetsitems en de toets als geheel: discrimineren tussen taalvaardige en minder taalvaardige leerlingen. Scoreverschillen kunnen echter ook geheel of gedeeltelijk een artefact zijn van de gevolgde meetprocedure.

1.1 Itembias

Scoreverschillen tussen subgroepen kunnen ook veroorzaakt worden door vaardigheidsverschillen die de toetsitems niet beogen te meten, maar die onbedoeld toch gemeten worden. Wanneer voor het correct beantwoorden van de toetsitems nog een andere vaardigheid nodig is dan de vaardigheid die de toets beoogt te meten, kan afbreuk gedaan worden aan de constructvaliditeit van de toets. Dit laatste is aan de orde wanneer die benodigde additionele vaardigheid bij de onderscheiden subgroepen niet in vergelijkbare mate aanwezig is. Voor kinderen die afkomstig zijn uit verschillende subgroepen en die even vaardig zijn in hetgeen het item beoogt te meten is de kans op een goed antwoord dan ongelijk. Wanneer dit zich voordoet is er sprake van *itembias* of *itempartijdigheid*: de toetsitems doen bij een of meer subgroepen onbedoeld een beroep op een multidimensionale vaardigheid (Mellenbergh, 1989; Hambleton & Rogers, 1989: Glas & Ouborg, 1993; Camilli & Shepard, 1994; Uiterwijk & Vallen, 1994). Itembias is bijvoorbeeld aan de orde wanneer het taalgebruik in rekenitems voor allochtone leerlingen dermate ingewikkeld is, dat deze daardoor in onvoldoende mate toekomen aan het uitvoeren van de beoogde rekenoperaties zelf. Bij dit voorbeeld gaat het om itempartijdigheid in het nadeel van allochtone leerlingen, maar het is ook mogelijk dat items partijdig zijn in hun voordeel.

Voor het correct beantwoorden van toetsitems moeten leerlingen vaak een beroep doen op additionele vaardigheden. Zo moeten leerlingen voor het beantwoorden van items inzake begrijpend lezen vaak eerst een tekst lezen. Voor het kunnen begrijpen van die tekst wordt kennis verondersteld over het onderwerp dat in die tekst aan de orde wordt gesteld. Ook bij rekenitems en items inzake wereldoriëntatie wordt bij het meten van de beoogde vaardigheid veelal gebruik gemaakt van contexten. Met het hanteren van contexten wordt gepoogd vast te stellen of de verworven kennis en vaardigheid en het verkregen inzicht bij een leerling voldoende groot en flexibel zijn om in een zo reëel mogelijke context te kunnen toepassen. De toetsconstructeur moet er bij de keuze van het contextmateriaal rekening mee houden dat de vereiste voorkennis, bijvoorbeeld inzake de beschreven situaties, het gehanteerde taalgebruik en het gebruikte beeldmateriaal (foto's, grafieken e.d.) bij alle leerlingen in vergelijkbare mate aanwezig is. Uit onderzoek naar itembias moet blijken in welke mate de toetsconstructeur daarin ten aanzien van bepaalde subgroepen is geslaagd. Door de inhoud te analyseren van partijdige items, kan gepoogd worden vast te stellen wat de oorzaak van de bias is. Deze inhoudsanalyse kan de toetsconstructeur aanwijzingen opleveren ter vermijding van itembias in nog te ontwikkelen toetsen.

1.2 Toetsbias

Toetsen worden ook gebruikt om voorspellingen te doen over buiten de toetssituatie liggend gedrag. Op grond van een behaalde toetsscore kan dan een verwachting worden uitgesproken over iemands niveau op een bepaald criterium op grond van eerder verworven kennis over de relatie tussen toetsscore en criteriumgedrag. Zo worden in Nederland aan het einde van de basisschool de taal- en reken- en vooral de totaalscores van leerlingen gebruikt om een indicatie te geven van het naar verwachting te behalen niveau in het voortgezet onderwijs, omdat uit eerder onderzoek de relatie tussen toetsscores en behaald niveau in het voortgezet onderwijs empirisch is vastgesteld. De predictieve validiteit van de toetsscore kan ook voor onderscheiden subgroepen bepaald worden. Er is sprake van *toetsbias* of *toetspartijdigheid* als systematische schattingsfouten gemaakt worden bij het voorspellen van de positie op het extern criterium als een functie van groepslidmaatschap. In feite is een toets onpartijdig wanneer de regressielijnen van toets op extern criterium van twee subgroepen (bijna) samenvallen (Jensen, 1980; Reynolds, 1982; Uiterwijk, 1994; Camilli & Shepard, 1994).

Om te kunnen beoordelen of een toets partijdig is in het voor- of nadeel van een bepaalde subgroep is het noodzakelijk om over een extern criterium te beschikken waarvan is aangetoond dat het zelf niet partijdig is voor de onderscheiden subgroepen (Jensen, 1980; Van de Vijver, Willemse & Van de Rijt, 1993; Uiterwijk, 1994). Bij de overgang van basisonderwijs naar voortgezet onderwijs kan het extern criterium 'succes in het voortgezet onderwijs' minder bruikbaar zijn in onderzoek naar toetsbias, wanneer bijvoorbeeld leden van de ene subgroep bij de toelating tot het voortgezet onderwijs eerder het voordeel van de twijfel krijgen dan gelijkpresterende leden van de andere subgroep. Ook bij de beslissingen die over de doorstroming van leerlingen in het voortgezet onderwijs worden genomen, is het niet zeker dat alleen hetgeen de toets meet (bijvoorbeeld: het algemene prestatieniveau van de leerling) als criterium wordt gehanteerd. Hoe verder een gehanteerd extern criterium in de tijd verwijderd ligt van het moment waarop de toets wordt afgenomen die object van toetsbiasonderzoek is, des te groter is de kans dat er variabelen zijn die een differentieel effect hebben op de schoolloopbanen van twee subgroepen. Hierdoor kan afbreuk gedaan worden aan de predictieve validiteit van een toets voor onderscheiden subgroepen. Deze predictieverschillen zijn echter niet te interpreteren als eigenschappen van de toets, maar als kenmerken van het extern criterium.

Hoewel er meestal geen volledig adequaat extern criterium beschikbaar is om de partijdigheid van een toets te beoordelen, is het niettemin zinvol om het verband tussen een toets en een maat voor succes in het voortgezet onderwijs voor bijvoorbeeld allochtone en autochtone leerlingen te onderzoeken. De gevonden verbanden geven immers aan of de mate van trefzekerheid waarmee het succes in het voortgezet onderwijs van allochtone en autochtone leerlingen voorspeld wordt vergelijkbaar is. Om een zo adequaat mogelijke verklaring te

kunnen geven voor eventuele verschillen, zijn echter de gegevens nodig van een groot aantal relevante onafhankelijke en afhankelijke variabelen en de effecten van die variabelen op elkaar moeten in een longitudinaal model geschat worden.

1.3 Literatuur over onderzoek naar item- en toetsbias in de Cito-Eindtoets Basisonderwijs

De Jong en Vallen hebben in 1989 in dit tijdschrift op grond van een door hen verrichte literatuurstudie bericht over mogelijke bronnen van itembias voor allochtone leerlingen. Zij hebben door een overzicht te geven van linguïstische en culturele bronnen van itembias in feite materiaal geleverd voor het formuleren van (werk)hypothesen over oorzaken van itembias. Coenen en Vallen hebben in 1991 in aansluiting op het artikel van De Jong en Vallen (1989) in dit tijdschrift verslag gedaan van een experiment waarin is nagegaan hoe vaak allochtone leerlingen bij een aantal partijdige items uit de Eindtoets Basisonderwijs 1987 van het Instituut voor toetsontwikkeling (Cito) een fout antwoord geven tengevolge van een element in het item dat waarschijnlijk de itembias veroorzaakt. Verder hebben Coenen en Vallen onderzocht hoe vaak vervolgens na manipulatie van dezelfde items tengevolge van die itemmanipulatie een goed antwoord wordt gegeven. Gemanipuleerde items zijn in dit verband items waarbij het itemelement dat vermoedelijk de biasbron vormt, is vervangen door een equivalent itemelement waarvan verwacht wordt dat het geen bias veroorzaakt. Het onderzoek van Coenen en Vallen heeft laten zien dat talige biasbronnen voor allochtone leerlingen voor een groot deel op het gebied van woordgebruik, impliciete zins- en tekstverbanden en verwijswoorden gezocht moeten worden. Verder lijken ongebruikelijke uitdrukkingen en woordvormgelijkenissen tot problemen te leiden. De geringere taalvaardigheid Nederlands van allochtone leerlingen kan er toe leiden dat ze meer moeite met complexe items hebben. Complexe items bevatten meer context en voor het oplossen van dergelijke items moet de leerling vaak een aantal tussenstappen maken. Welke dat zijn moet meestal uit de talige context afgeleid worden. Verder blijft het uiteraard mogelijk dat de itemcontext voor allochtone leerlingen minder herkenbaar is dan voor autochtone.

In Uiterwijk (1994) is gerapporteerd over het onderzoek naar zowel toets- als itembias dat in het kader van de Cito-Eindtoets Basisonderwijs 1987 en 1989 is verricht; in Uiterwijk en Vallen (1994) is het accent gelegd op het onderzoek naar itembias in beide Cito-eindtoetsen. In het onderhavige artikel, dat gezien kan worden als een vervolg op De Jong en Vallen (1989) en op Coenen en Vallen (1991) worden in de paragrafen 3 en 4 de belangrijkste resultaten gegeven uit Uiterwijk (1994). Bovendien zijn eerste resultaten opgenomen van het onderzoek naar toets- en itembias dat in het kader van de Eindtoets Basisonderwijs 1993 is gehouden.

2 Opzet van het onderzoek

Om na te gaan of de Eindtoets Basisonderwijs van het Cito voor allochtone leerlingen even bruikbaar is als voor autochtone leerlingen, voeren medewerkers van het Werkverband Taal en Minderheden van de Letterenfaculteit van de KUB en medewerkers van het project Eindtoets Basisonderwijs van het Cito een taalkundig-onderwijskundig onderzoek uit. In de eerste plaats hebben KUB en Cito onderzocht welk verband er bestaat tussen de scores op Eindtoets Basisonderwijs en een maat voor schoolsucces in het voortgezet onderwijs. Ten tweede is onderzoek gedaan naar itembias. In het onderzoek naar itembias zijn twee complementaire fasen onderscheiden: de opsporings- en de verklaringsfase. In de opsporingsfase worden met statistische technieken partijdige items gedetecteerd. In de tweede fase wordt gepoogd vast te stellen wat de oorzaak kan zijn van de bias in een item of in een reeks items. Bij het opsporen van mogelijke oorzaken van itembias zijn naast de projectmedewerkers van KUB en Cito ook een aantal experts en leerlingen uit groep acht van het basisonderwijs betrokken. Uit het onderzoek naar itembias moet blijken met welke aanpassingen de bruikbaarheid van de Eindtoets Basisonderwijs en andere toetsen voor allochtone leerlingen eventueel vergroot kan worden.

Voor de dataverzameling zijn door Cito en KUB achtergrondgegevens verzameld van de leerlingen die deelnamen aan de Eindtoets Basisonderwijs 1987 en 1989. Hierbij is aan de

leerkrachten van de leerlingen onder meer gevraagd aan te geven wat het land van herkomst van de ouders van de leerling is. Een leerling wordt tot een bepaald herkomstland gerekend wanneer beide ouders/verzorgers uit een hetzelfde land afkomstig zijn. Bij één-ouder-gezinnen geldt het herkomstland van de ouder/verzorger bij wie het kind woont. Tevens is nagegaan tot welk type voortgezet onderwijs de leerling werd toegelaten en naar welk type tweede leerjaar de leerling doorstroomde of dat er sprake was van doublure. Verder is het bestand met de resultaten van de deelnemers aan de Eindtoets Basisonderwijs 1993 gekoppeld aan het bestand dat de projectgroep Landelijke Evaluatie Onderwijsvoorrangsbeleid (LEO) van groep acht schooljaar 1992/1993 heeft opgebouwd. Door deze koppeling zijn er van 4218 deelnemers aan de Eindtoets Basisonderwijs 1993 achtergrondgegevens, zoals het herkomstland van de ouders, beschikbaar. Ook van deze leerlingen is nagegaan tot welke typen voortgezet onderwijs zij werden toegelaten en welke onderwijsposities de leerlingen na een jaar voortgezet onderwijs innamen.

Er zijn van de leerlingen uit 1987, 1989 en 1993 telkens twee databestanden opgebouwd, waarbij het tweede bestand een deelverzameling is uit het eerste. Het eerste bestand bestaat onder meer uit Eindtoets- en herkomstlandgegevens en het tweede bestand, dat door non-respons kleiner is, bevat naast Eindtoets- en herkomstlandgegevens ook gegevens over toelating tot en doorstroming in het voortgezet onderwijs. Het eerste bestand is in verband met het grotere aantal waarnemingen gebruikt voor onderzoek naar itembias, het tweede voor onderzoek naar toetsbias. Interessant is de vraag of het verwijderen van partijdige items leidt tot verhoging of verlaging van toetsbias. Door de opzet van dit onderzoek is het echter niet mogelijk om deze relatie te onderzoeken. In vervolgonderzoek is het evenwel gewenst het effect van itembias op toetsbias na te gaan.

3 Toetsbias

In het onderhavige onderzoek is nagegaan of de predictieve validiteit van de Eindtoets Basisonderwijs voor allochtone en autochtone leerlingen vergelijkbaar is. Hiermee wordt een indicatie verkregen van de mate van trefzekerheid waarmee de Eindtoets Basisonderwijs in de onderwijspraktijk het schoolsucces in het voortgezet onderwijs van beide groepen voorspelt. Een beoordeling van de predictieve validiteit als zodanig is, zoals eerder is betoogd, niet mogelijk omdat er op dit moment geen adequate maat voor schoolsucces in het voortgezet onderwijs bestaat waarvan aannemelijk gemaakt is, dat die onpartijdig is voor allochtone en autochtone leerlingen. In schoolloopbaanonderzoek bestaat een traditie ten aanzien van het construeren van een variabele die kan dienen als maat voor schoolsucces. Uitgangspunt hierbij is dat de verschillende schoolloopbanen van leerlingen in het voortgezet onderwijs hiërarchisch te ordenen zijn. Bosker (1990) en Van der Velden (1991) kwantificeren de posities die leerlingen in het voortgezet onderwijs innemen door de onderwijsniveaus (van IVBO tot VWO/Beroepsonderwijs) en leerjaren (van een tot zes leerjaren) tot een zogenaamde leerjarenladder te ordenen.

In het onderhavige onderzoek is de schoolloopbaan relatief kort: de positie die na één leerjaar (de brugklas) wordt ingenomen. Voor de kwantificering van deze posities zijn in verband met dit onderzoek de verschillen tussen de scholen met betrekking tot het algemene prestatieniveau van de leerlingen van belang. Toetsscores zijn immers gerelateerd aan het prestatieniveau van de leerlingen en vanuit dit perspectief wordt gekozen uit de schooltypen van het voortgezet onderwijs die onderling qua moeilijkheidsgraad verschillen. In dit onderzoek zijn de posities die de leerlingen in het voortgezet onderwijs innemen, geordend door aan elke onderwijspositie de waarde toe te kennen die overeenkomt met het gemiddelde prestatieniveau van alle leerlingen in dat type, zoals gemeten door de Eindtoets Basisonderwijs 1987, respectievelijk 1989 en 1993 (zie Uiterwijk, 1994). Om na te gaan hoe groot de trefzekerheid is waarmee het schoolsucces van allochtone en autochtone leerlingen in het voortgezet onderwijs door de Eindtoets Basisonderwijs wordt voorspeld is de correlatie berekend tussen toetsscore en schoolsucces. In Tabel 1 staan de correlaties vermeld van de leerlingen die in 1987, respectievelijk 1989 aan de Eindtoets Basisonderwijs deelnamen en in datzelfde jaar toegelaten werden tot het voort-

Tabel 1
Product-moment correlaties tussen Cito-score 1987 en 1989 en schoolsucces in het voortgezet onderwijs

Groep	*n*	Cito-score 1987	*n*	Cito-score 1989
Autochtonen	3274	.76	3405	.78
Alle allochtonen	2661	.76	3092	.74
Noordwest-Europa	84	.77	93	.74
China	95	.81	127	.63
Oost-Europa	23	.75	28	.69
Zuid-Europa	124	.70	147	.69
Molukken	139	.73	157	.68
Antillen	58	.71	74	.81
Suriname: Hindoestanen	184	.76	157	.68
Suriname: Creolen	202	.68	196	.71
Turkije	431	.69	534	.70
Marokko	375	.69	540	.68

gezet onderwijs. De autochtone leerlingen vormen een a-selecte steekproef uit de autochtone toetsdeelnemers. Alle in Tabel 1 vermelde correlaties zijn significant.

Uit Tabel 1 blijkt dat de voorspellende waarde van de Cito-Eindtoets 1987 bij allochtone en autochtone leerlingen even hoog is. In 1989 is de voorspellende waarde van de Eindtoets bij allochtone leerlingen lager dan bij autochtone. Verder blijkt dat de coëfficiënten per etnische groep in 1987 en 1989 soms aanzienlijk verschillen. Hierbij kan het geringe aantal waarnemingen een rol gespeeld hebben.

Van de leerlingen die zowel aan de Eindtoets Basisonderwijs 1993 als aan het LEO-onderzoek deelnamen, is nagegaan hoe de hoog de correlatie is tussen de Cito-score, respectievelijk de score op de voor het LEO-onderzoek ontwikkelde intelligentietest en schoolsucces in het voortgezet onderwijs. De score op de LEO-intelligentietest (zie Doddema-Winsemius & Van der Werf, 1989) is de som van de behaalde scores op de verbale en non-verbale subtests. In Tabel 2 geldt een leerling als allochtone leerling wanneer een van de ouders/verzorgers niet uit Nederland afkomstig is, in Tabel 1 is het herkomstland van beide ouders gehanteerd. Het effect van deze twee verschillende operationalisaties van herkomst op de uitkomsten van de analyses is niet bekend.

Tabel 2
Product-moment correlaties tussen Cito-score 1993, LEO-intelligentie-score en schoolsucces in het voortgezet onderwijs

Groep	*n*	Cito-score 1993	LEO-IQ
Autochtonen	1847	.80	.61
Allochtonen	518	.75	.56

Uit Tabel 2 blijkt dat het schoolsucces van allochtone leerlingen iets minder trefzeker wordt voorspeld dan dat van autochtone leerlingen. Dit geldt zowel voor de Cito-Eindtoets als voor de LEO-intelligentietest. Het verschil tussen allochtone en autochtone leerlingen is zowel op de Cito-Eindtoets als op de intelligentietest significant ($p<.001$, resp. $p<.05$). De voorspellende waarde van de LEO-intelligentietest ligt over de hele linie lager dan die van de Cito-Eindtoets 1993. In paragraaf 1.2 is gesteld dat niet zeker is of hoge of lage predictie toegeschreven moet worden aan de eigenschappen van het instrument of aan de partijdigheid van het extern criterium voor allochtone en autochtone leerlingen. In dit verband moet bij de verschillen in predictie tussen Cito-Eindtoets en LEO-test opgemerkt worden dat de Cito-Eindtoets afgenomen is om in schoolkeuzeprocessen een rol te spelen en dat de LEO-test louter ten behoeve van onderzoeksdoeleinden is gehanteerd.

4 Itembias

In het onderzoek naar itembias zijn, zoals eerder is opgemerkt, een opsporings- en een verklaringsfase onderscheiden. In de eerste fase

zijn met statistische technieken partijdige items opgespoord, in de tweede fase wordt gepoogd te achterhalen wat de oorzaak van de partijdigheid is.

4.1 Opsporen van partijdige items

Een item is partijdig wanneer leerlingen uit verschillende subgroepen maar met hetzelfde vaardigheidsniveau een ongelijke kans hebben om het betreffende item goed te beantwoorden. De statistische technieken voor het opsporen van een partijdig item evalueren dan ook de verschillen tussen subgroepen in de kansen op een goed antwoord gegeven een maat voor de vaardigheid die de toets als geheel beoogt te meten. Er moet dus een gevalideerde maat voor de vaardigheid beschikbaar zijn om de leerlingen van de onderscheiden subgroepen in niveaugroepen in te delen alvorens nagegaan kan worden of de kans op een goed antwoord voor de subgroepen vergelijkbaar is. Voor het opsporen van partijdige items zijn de meest algemeen toegepaste technieken gebaseerd op de Mantel-Haenszel-toets of op een model uit de Itemresponstheorie. Omdat gebleken is dat de resultaten van de verschillende detectieprocedures niet gelijk zijn, wordt algemeen geadviseerd om meer dan één techniek te hanteren. Voor uiteenzettingen over de kenmerken en de voor- en nadelen van de verschillende technieken kan verwezen worden naar Hambleton, Clauser, Mazor & Jones, 1993; Dorans & Holland, 1993; Thissen, Steinberg & Wainer, 1993; Glas & Ouborg, 1993; Camilli & Shepard, 1994.

Voor het opsporen van partijdige items in de Eindtoets Basisonderwijs 1987 en 1989 voor allochtone leerlingen is in de eerste plaats het computerprogramma One Parameter Logistic Model (OPLM) (Verhelst, 1992) gebruikt als procedure die gebaseerd is op de itemresponstheorie (IRT), vervolgens is het Mantel-Haenszel-programma (Verhelst, 1988) gehanteerd. Om de stabiliteit van de itemdetectieprocedures te bepalen, zijn de analyses telkens uitgevoerd op twee a-selecte steekproeven uit de onderscheiden subgroepen. In verband met het aantal waarnemingen per subgroep zijn alleen de Turkse, de Marokkaanse en de autochtone leerlingen in de analyses betrokken.

Als eerste stap in de procedure voor het opsporen van partijdige items is bepaald welke eendimensionale schalen elk toetsonderdeel (Taal, Rekenen en Informatieverwerking) van Eindtoets Basisonderwijs 1987 en 1989 bevat. Hierbij is eerst nagegaan welke items voor autochtone leerlingen op een schaal passen. Een schaal is als basis voor het onderzoek naar itembias geaccepteerd wanneer geen enkel item partijdig is ($p<.01$) bij de vergelijking a-selecte steekproef Autochtonen 1 versus a-selecte steekproef Autochtonen 2. Uit de uitgevoerde analyses blijkt dat de Eindtoets Basisonderwijs 1987 en 1989 voor autochtone leerlingen uit twee taal- en twee informatieverwerkingschalen bestaat. De ene taalschaal bestaat telkens voornamelijk uit spellingitems, terwijl de andere voornamelijk uit taalgebruikitems bestaat. De ene informatieverwerkingschaal wordt telkens voornamelijk gevormd door items afkomstig uit de rubriek Lezen van teksten, terwijl de andere vooral items bevat uit de rubrieken Hanteren van Informatiebronnen, Kaartlezen en Lezen van Tabellen en Grafieken. De Eindtoets Basisonderwijs 1987 bevat twee en die uit 1989 bevat drie rekenschalen. De rekenschalen zijn minder eenvoudig te karakteriseren.

Als tweede stap in de procedure is zowel met het computerprogramma OPLM als met het Mantel-Haenszel-programma voor elke steekproef autochtone versus Turkse, respectievelijk Marokkaanse leerlingen vastgesteld welke items partijdig zijn ($p<.01$). Een item is partijdig voor Turkse leerlingen wanneer het item zowel bij de vergelijking steekproef Autochtonen 1 versus steekproef Turken 1 als bij de vergelijking Autochtonen 2 versus Turken 2 partijdig is. Dezelfde procedure is ook bij de vergelijking autochtone versus Marokkaanse leerlingen gebruikt.

Uitgangspunt bij de gevolgde procedure is dat als een reeks items voor autochtone leerlingen een eendimensionale schaal blijkt te vormen, die items ook een eendimensionale schaal dienen te vormen voor Turkse en Marokkaanse leerlingen. Wanneer dezelfde items voor Turkse en Marokkaanse leerlingen niet passen dan is deze schaal voor de allochtone leerlingen multidimensionaal en de niet-passende items kunnen dan als partijdig beschouwd worden. Voor zowel de Mantel-Haenszel- als voor de OPLM-procedure moeten de twee subgroepen in niveaugroepen ingedeeld worden aan de

Tabel 3

Aantal items uit de Eindtoets Basisonderwijs 1987 en 1989 dat partijdig is voor Turkse en/of Marokkaanse leerlingen

Onderdeel	Mantel-Haenszel		IRT		IRT & MH	
	voor-deel	na-deel	voor-deel	na-deel	voor-deel	na-deel
Taal (k=120)	8	10	3	1	2	7
Rekenen (k=120)		4		1		
Informatie-verwerking (k=120)	4	6	1	1	1	3
Totaal (k=360)	12	20	4	3	3	10

hand van een meting die kan gelden als maat voor de vaardigheid die de toets als geheel beoogt te meten. Uitgangspunt bij de keuze van het criterium voor het matchen van subgroepen is geweest dat de somscore van de items, die voor autochtone leerlingen op eendimensionale schaal passen, ook geldig dient te zijn als somscore voor de Turkse en Marokkaanse leerlingen.

De resultaten van de analyses naar itembias maken duidelijk dat het moeilijk is om aan te geven hoeveel items van Eindtoets Basisonderwijs 1987 en 1989 partijdig zijn. De verschillende analyses laten een wisselend beeld zien. Uit de uitgevoerde analyses blijkt dat 13 van de 360 geanalyseerde items (= 4%) zowel met de IRT- als met de Mantel-Haenszelprocedure partijdig zijn voor Turkse en/of Marokkaanse leerlingen. Hiervan zijn 3 items partijdig in het voordeel van deze leerlingen, 10 items zijn partijdig in hun nadeel. Verder blijkt dat 7 items (= 2%) alleen partijdig zijn met de techniek volgens het IRT-model, hiervan zijn 4 items partijdig in het voordeel, 3 items zijn in het nadeel. Er zijn 32 items (= 9%) alleen partijdig met de Mantel-Haenszelprocedure, hiervan zijn er 12 in het voor- en 20 in het nadeel van Turkse en/of Marokkaanse leerlingen. Verder volgt uit Tabel 3 dat itembias het meest voorkomt bij Taal, gevolgd door respectievelijk Informatieverwerking en Rekenen.

Uit nadere analyse blijkt dat de IRT- en de Mantel-Haenszel-procedures in 87% van de gevallen overeenstemmen in het detecteren van (on)partijdige items. Dit beeld komt overeen met de resultaten van andere onderzoekers (Bügel & Glas, 1991; Hambleton & Jones, 1992; Camilli & Shepard, 1994). De stabiliteit van de IRT- en de Mantel-Haenszel-procedure is vrijwel gelijk: in twee steekproeven wijst de Mantel-Haenszel-procedure bij 86% van de items en de IRT-procedure bij 89% van de items in beide steekproeven een item als partijdig aan. De IRT-procedure spoort minder partijdige items op dan de Mantel-Haenszel-techniek. Items blijken nooit in het voordeel te zijn voor Turkse leerlingen en tegelijkertijd in het nadeel van Marokkaanse leerlingen of omgekeerd.

Van de leerlingen die zowel aan de Eindtoets Basisonderwijs 1993 als aan het LEO-onderzoek deelnamen, is nagegaan in welke mate er in de Eindtoets Basisonderwijs 1993 sprake is van itembias in het voor- of nadeel van allochtone leerlingen. Een schaal is, evenals dat gebeurde met de data uit 1987 en 1989, in 1993 als basis voor onderzoek naar itembias geaccepteerd wanneer geen enkel item partijdig is in de vergelijking steekproef Autochtonen 1 versus steekproef Autochtonen 2 ($p<.01$). Een item is partijdig ($p<.01$) voor allochtone leerlingen wanneer het item bij de vergelijking steekproef Autochtonen 1 versus alle allochtone leerlingen partijdig is. Er is niet voor gekozen om beide steekproeven autochtone leerlingen ($n = 2 \times 772$) samen te voegen, omdat de te vergelijken groepen allochtone ($n = 423$) en autochtone leerlingen bij voorkeur ongeveer even groot moeten zijn (Hambleton et al., 1993; Zieky, 1993). De steekproef Autochtonen 1 onderscheidt zich van de steekproef Autochtonen 2, omdat van de eerste steekproef de passing op de eendimensionale schalen telkens het grootst is.

Uit de resultaten van de analyses naar itembias in de Eindtoets Basisonderwijs 1993 blijkt dat 16 van de in totaal 180 geanalyseerde items (= 9%) zowel met de IRT- als met de Mantel-

Tabel 4
Aantal items uit de Eindtoets Basisonderwijs 1993 dat partijdig is voor allochtone leerlingen

Onderdeel	Mantel-Haenszel		IRT		IRT & MH	
	voor-deel	na-deel	voor-deel	na-deel	voor-deel	na-deel
Taal (k=60)	4	2	3		3	7
Rekenen (k=60)	2		1	2	1	1
Informatie-verwerking (k=60)	4	5	2	1	1	3
Totaal (k=180)	10	7	6	3	5	11

Haenszelprocedure partijdig zijn voor allochtone leerlingen. Van deze 16 items zijn er 5 partijdig in het voordeel en 11 in het nadeel van allochtone leerlingen. Verder blijkt dat 9 items (= 5%) alleen met de IRT-procedure partijdig zijn voor allochtone leerlingen. Hiervan zijn 6 items partijdig in het voordeel, 3 items zijn in het nadeel. Van de 180 geanalyseerde items zijn 17 items (= 9%) alleen partijdig met de Mantel-Haenszel-procedure. Van deze 17 items zijn er 10 in het voordeel van allochtone leerlingen, 7 items zijn in het nadeel. Ook hier spoort de IRT-procedure minder partijdige items op dan de Mantel-Haenszel-techniek. Verder volgt uit Tabel 4 dat bij Taal het aantal partijdige items het grootst is, gevolgd door respectievelijk Informatieverwerking en Rekenen.

Er is een aanzienlijke overlap tussen de Mantel-Haenszel- en de IRT-procedure, maar er zijn ook verschillen waardoor het niet altijd duidelijk is of een item partijdig is of niet. Over het algemeen wordt dan ook geadviseerd in de vorm van kruisvalidatie beide technieken te hanteren. Onderzoek naar itembias heeft als resultaat dat uitspraken over de partijdigheid van items gradueel zijn: partijdig bij alle analyses, bij een deel van de analyses, bij geen enkele analyse.

Bij alle partijdige items is in de volgende fase van het onderzoek nagegaan wat de oorzaak van de partijdigheid zou kunnen zijn. De stelligheid waarmee op dit punt conclusies getrokken worden is mede afhankelijk van het aantal keren dat er bij een item sprake is van bias.

4.2 Bronnen van itembias

Er is niet alleen in Nederland maar ook in andere landen weinig onderzoek gedaan naar inhoudelijke oorzaken van itembias voor allochtone leerlingen. Volgens Schmitt, Holland & Dorans (1993) zijn hiervoor drie redenen aan te wijzen. In de eerste plaats is onderzoek naar itembias relatief nieuw. Tot nu toe is de meeste aandacht uitgegaan naar statistische procedures voor het detecteren van partijdige items. In de tweede plaats veronderstelt het achterhalen van oorzaken van itembias voor allochtone leerlingen een theorie over de vraag waarom items voor de onderscheiden etnische groepen partijdig zijn. Omdat etnische groepen in velerlei opzichten heterogeen zijn, kunnen de verschillen tussen etnische groepen moeilijk beschreven worden. In de derde plaats is het opsporen van oorzaken van itembias uitermate complex, omdat bij een bepaald item verschillende oorzaken een rol kunnen spelen.

Omdat een theoretisch kader betreffende inhoudelijke bronnen voor itembias voor allochtone leerlingen vooralsnog niet beschikbaar is, moeten we de conclusies die op basis van het onderhavige onderzoek in dit verband getrokken worden, als voorlopig beschouwen. Om bij de inhoudsanalyse van in statistisch opzicht partijdige items toch enigszins gericht te kunnen kijken, is eerst een literatuuroverzicht gemaakt van potentiële bronnen van itembias. Dit overzicht bestaat uit een reeks van (deel)vaardigheden waarvan uit onderzoek is gebleken of waarvan te verwachten is dat allochtone en autochtone leerlingen deze aan het einde van het basisonderwijs in verschillende mate beheersen. Het gaat hier dus om verschil in moeilijkheid en niet om itembias. Verschillen in moeilijkheidsgraad zijn in het kader van dit onderzoek van belang omdat ze de basis kunnen vormen voor hypothesen omtrent oorzaken van itembias (De Jong & Vallen, 1989; Coenen & Vallen, 1991; Uiterwijk, 1994).

Bij het zoeken naar mogelijke oorzaken van itembias in de Eindtoets Basisonderwijs 1987 en 1989 zijn niet alleen de medewerkers van

het onderzoeksproject (van KUB en Cito) betrokken geweest maar ook een aantal niet aan het project verbonden experts en leerlingen uit groep acht van het basisonderwijs.

De partijdige items in het voor- of nadeel van Turkse en/of Marokkaanse leerlingen zijn op grond van de in het geding zijnde vaardigheden eerst door drie projectmedewerkers afzonderlijk inhoudelijk geanalyseerd tegen de achtergrond van de vraag welke itemelementen mogelijk een bron van itembias vormen. Hierbij ging speciale aandacht uit naar elementen die voorkomen in items die partijdig zijn in het nadeel van beide groepen allochtone leerlingen en ontbreken in items die in het voordeel zijn van deze leerlingen en omgekeerd. Bij deze inhoudsanalyse hebben de vooraf geformuleerde potentiële bronnen van itembias voor allochtone leerlingen als hypothesen gefungeerd. Vervolgens zijn de gemeenschappelijkheden in de analyses van de afzonderlijke projectmedewerkers geïnventariseerd.

De inhoudelijke analyse leverde twee fundamentele problemen op. Enerzijds blijkt het moeilijk te zijn om met zekerheid aan te geven welk element van een item de biasbron is en anderzijds blijkt dat bij sterk vergelijkbare items de ene keer wel sprake is van itembias en de andere keer niet. Tevens werd vastgesteld dat er een aanzienlijk aantal partijdige items is met onderlinge overeenkomsten. Items die samen een cluster vormen, vertonen overeenkomsten ten aanzien van hetgeen de items beogen te meten of vertonen overeenkomsten met betrekking tot additionele vaardigheden die relevant zijn voor het juist beantwoorden van de items. De inhoudelijke analyses leveren voorlopige conclusies op (vgl. Uiterwijk & Vallen, 1994). Deze conclusies worden voorlopig genoemd, omdat de conclusies in een volgende fase van het onderzoek zijn vergeleken met de oordelen van 16 niet bij het onderzoeksproject betrokken experts, elf autochtone en vijf allochtone (Van de Waal, 1992; Uiterwijk, 1994). Verder is met een kleinschalig hardopdenken-experiment nagegaan hoe vaak allochtone en autochtone leerlingen bij partijdige items een fout antwoord geven door een itemelement dat voorlopig als bron voor itembias is aangewezen. Tevens is onderzocht hoe vaak bij gemanipuleerde items tengevolge van de itemmanipulatie een goed antwoord wordt gegeven (Coenen & Vallen, 1991). Gemanipuleerde items zijn items waarbij het itemelement dat als potentiële biasbron is aangewezen (bijvoorbeeld: 'Hoeveel moet hij betalen inclusief B.T.W.'), is vervangen door een itemelement waarvan verwacht wordt dat het geen bias veroorzaakt (bijvoorbeeld: 'Hoeveel moet hij betalen met B.T.W.'). De allochtone en autochtone leerlingen moesten eerst de aan hen voorgelegde oorspronkelijke, respectievelijk gemanipuleerde items goed bestuderen, daarna konden ze het naar hun oordeel goede antwoord aankruisen. Tot slot moesten ze zo uitgebreid en nauwkeurig mogelijk mondeling toelichten hoe ze tot hun antwoord gekomen waren. In de redeneringen van de leerlingen is getracht aanwijzingen te vinden omtrent inhoudelijke oorzaken van itembias.

De partijdige items uit de Eindtoets Basisonderwijs 1993 zijn eveneens onderzocht op mogelijke oorzaken van itembias. De bevindingen uit het onderzoek met de Eindtoets Basisonderwijs 1987 en 1989 zijn in feite geverifieerd, waardoor stelligheid waarmee uitspraken gedaan worden kan toenemen. Uit de resultaten van 1993 blijkt dat het grootste deel van de resultaten uit 1987 en 1989 ondersteund wordt. In het volgende overzicht, wordt aangegeven welke items gezien het onderzoek met de Eindtoets Basisonderwijs uit 1987, 1989 en 1993 partijdig kunnen zijn in het nadeel van Turkse en Marokkaanse leerlingen.

- Items die vragen naar de betekenis van een woord, een woordcombinatie of een zin, waarbij om een min of meer woordelijke herhaling van expliciet in de tekst gegeven informatie wordt gevraagd.
- Items die vragen naar de betekenis van moeilijke woorden, woordcombinaties, functiewoorden en verwijswoorden, waarbij de betekenis niet of moeilijk uit de context kan worden afgeleid.
- Items die betrekking hebben op de kennis van de vorm van idiomatische, letterlijk of figuurlijk gebruikte woordcombinaties en/of conventies op het gebied van de zinsbouw (zinsvolgorde, inversie/vraagvormen).
- Items die veel contextmateriaal bevatten. Voor het oplossen van complexe items moet de leerling vaak een aantal tussenstappen maken. Uit de context moet dan afgeleid

worden welke tussenstappen gemaakt moeten worden en dit wordt voor allochtone leerlingen extra moeilijk wanneer er in dergelijke items (vooral op zinsniveau) veel verwijswoorden en impliciete zins- en tekstverbanden voorkomen.
- Items die betrekking hebben op contextmateriaal (bijvoorbeeld een tekst) dat specifieke Nederlandse cultuurkennis vereist. Zo kunnen allochtone leerlingen minder vertrouwd zijn met het onderwerp dat in een tekst aan de orde wordt gesteld. Dit speelt vooral wanneer dat contextmateriaal cruciaal is voor het oplossen van het item.

Verder zijn er aanwijzingen dat de volgende items partijdig kunnen zijn in het voordeel van allochtone leerlingen.
- Items die naar de hoofdgedachte van een tekst vragen.
- Items die vragen spellingfouten in werkwoorden en in woorden met een vast woordbeeld aan te geven.

5 Conclusies en discussie

Of een toets het schoolsucces in het voortgezet onderwijs van een bepaalde subgroep even goed voorspelt als van een andere subgroep, is in feite niet te beoordelen zolang er geen extern criterium is waarvan aangetoond is dat het zelf onpartijdig is. In paragraaf 1.2 is daarom gesteld dat het niet zeker is of hoge of lage predictie toegeschreven moet worden aan de eigenschappen van het instrument of aan de partijdigheid van het extern criterium voor allochtone en autochtone leerlingen. In het onderhavige onderzoek is wel gebleken dat de Cito-Eindtoets Basisonderwijs het schoolsucces in het voortgezet onderwijs van allochtone leerlingen minder trefzeker wordt voorspeld dan dat van autochtone leerlingen. Voor de verklaring van deze verschillen zou nauwkeurig vastgesteld moeten worden of bij allochtone en autochtone leerlingen de beschikbare gegevens, zoals advies basisschool en toetsscore, door de betrokkenen in het schoolkeuzeproces op dezelfde wijze worden gewogen. Het lijkt ook zinvol om na te gaan welke aanvullende informatie naast het advies basisschool en de Cito-score de trefzekerheid van de schoolkeuze van allochtone leerlingen kan vergroten. Opgemerkt moet worden dat allochtone leerlingen in het voortgezet onderwijs meer dan hun autochtone klasgenoten blijven zitten of doorstromen naar andere – meestal 'lagere' – schooltypen (Uiterwijk, 1994). De relatief lage trefzekerheid waarmee het schoolsucces van allochtone leerlingen wordt voorspeld, is voor een deel te verklaren door de grote uitval in het voortgezet onderwijs van allochtone leerlingen. Het is van groot belang om na te gaan hoe de beslissingen over toelating in het voortgezet onderwijs verbeterd kunnen worden en hoe het onderwijsaanbod in het voortgezet onderwijs beter op de allochtone leerlingen kan worden afgestemd.

Ten aanzien van het onderzoek naar itembias kan gezegd worden dat dit met veel onzekerheden is omgeven. In de eerste plaats kan niet duidelijk worden aangegeven of een item partijdig is of niet (4.1). Dit wordt veroorzaakt door het ontbreken van volledige overeenstemming tussen de resultaten van de verschillende itembiasdetectieprocedures, waardoor het in dit soort onderzoek gebruikelijk is om meer dan één analysetechniek te hanteren.

In de tweede plaats kan niet duidelijk aangegeven worden welk element van een item verantwoordelijk is voor de itembias (4.2). De bron van itembias kan bijvoorbeeld in het contextmateriaal, in de vraag, de antwoordmogelijkheden en in de te meten (deel)vaardigheid zitten. Naarmate het contextmateriaal omvangrijker wordt, is het moeilijker om met zekerheid de inhoudelijke biasbron aan te wijzen. Zo kan de voorkennis omtrent hetgeen in teksten aan de orde wordt gesteld van invloed zijn op de itemantwoorden.

Via verschillende wegen (inhoudsanalyse, expertbevraging, experimenten) zal informatie over bronnen van itembias verzameld moeten worden.

Als uiteindelijk – zoals in de onderhavige studie – met een zekere mate van waarschijnlijkheid een aantal biasbronnen opgespoord is, dan rest nog de vraag of de biasbronnen afbreuk doen aan de constructvaliditeit van de toets of niet. Wanneer een bron van itembias behoort tot de te meten vaardigheid, dan meet het item wat het beoogt te meten. Als voorbeeld geldt het volgende taalitem. Het toetsonderdeel Taal beoogt onder meer de hier aan de orde zijnde vaardigheid 'het kennen van de betekenis van woorden' te meten.

Item 14 uit de taak Taal 2 van de Eindtoets 1989

43 Ik hoop echt dat jullie komen. Jullie
44 moeten wel uiterlijk half zeven bij ons
45 thuis zijn. Jullie moeten echt zo vroeg
46 mogelijk komen hoor. Anders zijn er
47 vast geen kaartjes meer te krijgen.

Wat had de briefschrijver in plaats van *uiterlijk* (regel 44) ook kunnen schrijven?

A op zijn best
B op zijn laatst
C op zijn mooist
D op zijn vroegst

Dit item dat naar de betekenis van het specifieke woord 'uiterlijk' vraagt, is multidimensionaal voor Marokkaanse leerlingen. Toch kan niet gezegd worden dat het item iets anders zou meten dan het moet meten. Als aangenomen wordt dat leerlingen het woord 'uiterlijk' in deze betekenis aan het einde van het basisonderwijs moeten kennen, dan doet in dit geval itembias geen afbreuk aan de constructvaliditeit. Wanneer een biasbron niet tot de te meten vaardigheid behoort dan doet het item wel afbreuk aan de constructvaliditeit van de toets. Zo moeten bronnen van itembias op het gebied van woordenschat geen rol spelen bij rekenitems. Itembias als zodanig beperkt de constructvaliditeit van een toets dus niet in alle gevallen.

Gebleken is dat het aantal bronnen van itembias dat met een grote mate van zekerheid bias veroorzaakt, in feite gering is. Het detecteren van partijdige items en het verklaren van de oorzaken ervan is een lange weg met een bescheiden opbrengst. Informatie over potentiële bronnen van itembias komt ook beschikbaar uit elk empirisch onderzoek dat duidelijk maakt bij welke kennis- en vaardigheidsaspecten allochtone leerlingen significant lager scoren dan autochtone leerlingen. Wanneer de toetsconstructeur over gedetailleerde informatie op dit gebied beschikt, kan hij het risico verkleinen dat hij inzake allochtone leerlingen items construeert die afbreuk doen aan de constructvaliditeit van een toets. Hiervoor is het van belang dat in empirisch onderzoek kennis- en vaardigheidsaspecten op een zo gedetailleerd mogelijk niveau gemeten worden.

Literatuur

Bosker, R.J. (1990). *Extra kansen dankzij de school?* Nijmegen: Instituut voor Toegepaste Sociale Wetenschappen.

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.

Camilli, G., & Shepard, L.A. (1994). *Methods for identifying biased test items.* Thousand Oaks: Sage Publications.

Coenen, M., & Vallen, T. (1991). Itembias in de eindtoets basisonderwijs. *Pedagogische Studiën, 68,* 15-26.

Doddema-Winsemius, H., & Werf, G. van der (1989). *Selectie/constructie van toetsen voor sociale redzaamheid en intelligentie ten behoeve van de evaluatie van het OVB.* Groningen: Research Instituut voor onderzoek in het noorden.

Dorans, N.J., & Holland, P.W. (1993). Dif detection and description: Mantel-Haenszel and standardization. In P.W. Holland & H.W. Wainer (Eds.), *Differential item functioning* (pp. 35-66). Hillsdale: Lawrence Erlbaum Associates.

Glas, C.A.W., & Ouborg, M.J. (1993). Vraagonzuiverheid. In T.J.H.M. Eggen & P.F. Sanders (Eds.), *Psychometrie in de praktijk* (pp. 349-370). Arnhem: Cito, Instituut voor toetsontwikkeling.

Hambleton, R.K., & Rogers, H.J. (1989). Detecting potentially biased test items: comparison of IRT area and Mantel-Haenszel methods. *Applied Measurement in Education, 2,* 313-334.

Hambleton, R.K., & Jones, R.W. (1992). *Comparison of empirical and judgmental methods for detecting differential item functioning.* Princeton: Educational Testing Service.

Hambleton, R.K., Clauser, B.E., Mazor, K.M., & Jones, R.W. (1993). Advances in the detection of differentially functioning test items. *European Journal of Psychological Assessment, 9,* 1-18.

Jensen, A.R. (1980). *Bias in mental testing.* Londen: Methuen.

Jong, M. de, & Vallen, T. (1989). Linguïstische en culturele bronnen van itembias in de Eindtoets Basisonderwijs voor leerlingen uit etnische minderheidsgroepen. *Pedagogische Studiën, 66,* 390-402.

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.

Reynolds, C.R. (1982). Methods for detecting construct and predictive bias. In R.A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 199-227). Baltimore: The Johns Hopkins University Press.

Schmitt, A.P., Holland, P.W., & Dorans, N.J. (1993). Evaluating hypotheses about differential item functioning. In P.W. Holland & H.W. Wainer (Eds.), *Differential item functioning* (pp. 281-315). Hillsdale: Lawrence Erlbaum Associates.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P.W. Holland & H.W. Wainer (Eds.), *Differential item functioning* (pp. 67-113). Hillsdale: Lawrence Erlbaum Associates.

Uiterwijk, J.H. (1994). *De bruikbaarheid van de Eindtoets Basisonderwijs voor allochtone leerlingen.* Arnhem: Cito, Instituut voor toetsontwikkeling.

Uiterwijk, H., & Vallen, T. (1994). Talige bronnen van itembias voor allochtone leerlingen in de Eindtoets Basisonderwijs. *Spiegel, 12,* 9-29.

Velden, R.K.W. van der (1991). *Sociale herkomst en schoolsucces.* Monografieën onderwijsonderzoek nr. 10. Groningen: Instituut voor Onderwijsonderzoek.

Verhelst, N.D. (1988). *De Mantel-Haenszel-toetsen.* Arnhem: Cito, Instituut voor Toetsontwikkeling.

Verhelst, N.D. (1992). *Het eenparameter logistisch model (OPLM), een theoretische inleiding en een handleiding bij het computerprogramma.* Arnhem: Cito, Instituut voor toetsontwikkeling.

Vijver, F. van de, Willemse, G., & Rijt, B. van de (1993). Het testen van cognitieve vaardigheden van allochtone leerlingen. *De Psycholoog, 28,* 152-159.

Waal, M. van de (1992). *Expert-oordelen over potentiële bronnen van itembias in de Eindtoets Basisonderwijs.* Tilburg: Katholieke Universiteit Brabant.

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

Manuscript aanvaard 9-2-1996

Auteurs

H. Uiterwijk is projectleider Eindtoets Basisonderwijs bij het Instituut voor Toetsontwikkeling (Cito).

Adres: Instituut voor toetsontwikkeling (Cito), Postbus 1034, 6801 MG Arnhem

T. Vallen is universitair hoofddocent bij het Werkverband Taal en Minderheden van de Letterenfaculteit van de Katholieke Universiteit Brabant.

Adres: Faculteit der Letteren KUB, Postbus 90153, 5000 LE Tilburg

Abstract

Testbias and itembias for ethnic minority pupils in the Final Primary School Tests of the Dutch National Institute for Educational Measurement

H. Uiterwijk & T. Vallen. Pedagogische Studiën, 1997, 74, 21-32.

So far, in the Netherlands as well as in other countries hardly any substantial research has been carried out into the suitability of tests for specific subgroups within a larger population. In an interdisciplinary (educational and sociolinguistic) research project the Dutch National Institute for Educational Measurement (Cito) and the Research Group on Language and Minorities of Tilburg University have cooperated in a study into the possible existence of testbias and itembias for ethnic minority pupils in the Cito Final Primary School Test in the Netherlands. With respect to testbias it emerges that the educational success in secondary education can be predicted less accurately for ethnic minority pupils than for pupils who belong to the majority group. This conclusion holds for both the Final Primary School Test and the intelligence test used.

The research carried out into itembias shows that the several statistical procedures used for detecting itembias for ethnic minority pupils in the Final Primary School Test produce different results. Most items never appear to be biased, others sometimes, and a limited number of items always show itembias. The procedures used for discovering the reasons for itembias in terms of content revealed especially within the field of linguistics a number of potential sources of itembias (e.g. vocabulary, referential problems, amount of context) for ethnic minority pupils. The knowledge of potential sources of itembias can be useful as a tool for diminishing the risk of itembias for these pupils in future tests.