Henny Uiterwijk & Ton Vallen

Talige bronnen van itembias voor allochtone leerlingen in de Eindtoets Basisonderwijs

1. Inleiding

In deze bijdrage wordt gerapporteerd over een gedeelte van een door het Cito samen met het werkverband Taal en Minderheden van de Letterenfaculteit van de KUB uitgevoerd taalkundig-onderwijskundig onderzoek naar de bruikbaarheid van de Eindtoets Basisonderwijs voor allochtone leerlingen. In dat project stonden twee thema's centraal. Allereerst is nagegaan hoe hoog de voorspellende waarde van de Eindtoets Basisonderwijs is voor allochtone en autochtone leerlingen in vergelijking met de voorspellende waarde van het schoolkeuze-advies van de basisschool. Dit onderzoeksgedeelte heeft met andere woorden betrekking op een vergelijking van de trefzekerheid waarmee door toets en advies het schoolsucces van beide subgroepen van leerlingen in het voortgezet onderwijs wordt ingeschat. Wanneer er wat de trefzekerheid van de Eindtoets als voorspeller van schoolsucces betreft verschillen tussen de onderscheiden subgroepen bestaan, is er sprake van 'toetspartijdigheid' of 'toetsonzuiverheid', kortweg aangeduid met de term *toetsbias*. Dit deel van het uitgevoerde project blijft in onderhavig artikel verder buiten beschouwing. Wel kan er samenvattend over worden opgemerkt dat het advies van de basisschool voor zowel autochtone als allochtone leerlingen een iets hogere voorspellende waarde heeft dan de Eindtoets en dat zowel het advies als de Eindtoets het schoolsucces van allochtone leerlingen in het voortgezet wat minder trefzeker voorspellen dan dat van autochtone leerlingen. Uiterwijk (1994) biedt uitvoerige informatie over dit gedeelte van het onderzoek.

De tweede centrale onderzoekscomponent binnen het project was niet gericht op de toets als geheel maar op de afzonderlijke opgaven (items) waaruit de Eindtoets Basisonderwijs bestaat. Daarbij werd in eerste instantie statistisch onderzocht of de toets items bevat waarbij allochtone leerlingen in vergelijking met autochtone leerlingen die over dezelfde vaardigheden beschikken, toch nog een ongelijke kans hebben om die items goed te beantwoorden. Bij die items waarbij dat het geval is, kan in navolging van Kok (1988) gesproken worden van 'vraagpartijdigheid'. Ook wordt hiervoor wel de aanduiding 'vraagonzuiverheid' gehanteerd; de internationaal gangbare term voor dit fenomeen is *itembias* en deze zal, naast het Nederlandse equivalent *partijdigheid*, in de rest van dit artikel worden gebruikt. In een tweede fase van dit onderzoeksgedeelte is een poging ondernomen om na te gaan wat bij de statistisch gebiaste items de inhoudelijke, met name talige oorzaken van de partijdigheid zouden kunnen zijn. In het onderstaande zal hierop het accent liggen. Behalve via een zoekexercitie in de literatuur werden de voor dit gedeelte van het onderzoek relevante gegevens vooral ingewonnen door middel van een tweetal

procedures waarmee getracht werd talige bronnen van itembias op het spoor te komen.

Na de start van het project bleek al snel dat met name het onderzoek naar itembias in meerdere opzichten een ontdekkingsreis vol voetangels en klemmen zou worden. Zo werd duidelijk dat in de Verenigde Staten weliswaar veel aandacht is geschonken aan statistische procedures voor het opsporen van itembias (teneinde partijdige items uit toetsen te kunnen verwijderen), maar tevens bleek dat vergelijkbare procedures vaak niet tot identieke resultaten leiden. Op de vraag bij hoeveel items van een bepaalde toets sprake is van bias, zijn dan ook verschillende antwoorden mogelijk. Verder bleek dat met het zoeken van inhoudelijke oorzaken van itembias zowel in als buiten Nederland bijzonder weinig ervaring is opgedaan. Goed gefundeerde taalkundige verklaringen inzake itembias voor allochtone leerlingen ontbreken geheel en al. Door de afwezigheid van een theoretisch kader voor inhoudelijke bronnen van itembias voor allochtone leerlingen dragen de conclusies die op basis van het onderhavige onderzoek in dit verband worden getrokken, dan ook een voorlopig karakter.

In de volgende paragrafen worden van de beide procedures ter detectie van inhoudelijke bronnen van itembias in het kort de opzet en de belangrijkste resultaten besproken. Daaraan voorafgaand wordt in paragraaf 2 allereerst nader ingegaan op het onderscheid tussen de moeilijkheidsgraad van een item en itembias (2.1) en op de opzet en de resultaten van het uitgevoerde statistisch onderzoek naar itembias (2.2). In paragraaf 3 geven we eerst een korte samenvatting van de belangrijkste resultaten van een literatuurstudie naar potentiële talige en culturele bronnen van itembias (3.1); daarna beschrijven we doelstelling en opzet (3.2) en de resultaten (3.3) van een bij leerlingen uitgevoerde hardop-denken-procedure. In paragraaf 4 komen doelstelling en opzet (4.1) en de resultaten (4.2) aan de orde van een procedure waarmee een aantal experts is gevraagd naar hun oordelen over problemen met het maken van Eindtoetsitems voor allochtone basisschoolverlaters. In paragraaf 5 ten slotte, zal een samenvattend overzicht worden gegeven van de belangrijkste potentiële talige bronnen van itembias en zullen enkele suggesties worden gedaan ter voorkoming van dit soort ongewilde benadelingen voor allochtone kinderen in toetsen en andere evaluatie-instrumenten.

2. Probleemstelling, opzet en resultaten van het statistisch itembias-onderzoek

2.1. Probleemstelling

Hoewel de schoolresultaten in het basisonderwijs en de doorstroming van de basisschool naar het voortgezet onderwijs bij de meeste groepen allochtone leerlingen de laatste jaren wat zijn verbeterd, blijven hun scores nog steeds beduidend achter bij die van hun autochtone klasgenoten (zie bijvoorbeeld CALO, 1992 en Tesser,

1993). Ook de scores op de Eindtoets Basisonderwijs, waarvan het jaarlijks aantal deelnemers inmiddels tot boven de 100.000 gestegen is, laten dit beeld zien. Zoals bekend, bestaat deze toets uit 180 items, evenredig verdeeld over de onderdelen Taal, Rekenen en Informatieverwerking. Gemiddeld hebben de Marokkaanse en Turkse leerlingen de meeste moeite met deze opgaven, maar ook de kinderen van Surinaamse en Antilliaanse afkomst scoren door de bank genomen toch ook gemiddeld nog één standaarddeviatie onder het gemiddelde van de autochtone leerlingen (zie Uiterwijk & Vallen, 1991 en Uiterwijk, 1994).

Bij het zoeken naar verklaringen voor de soms aanzienlijke verschillen in (Eind)toetsscores tussen autochtone en allochtone leerlingen moet rekening worden gehouden met twee, principieel van elkaar verschillende, mogelijkheden. De eerste is dat de scoreverschillen veroorzaakt worden door de uiteenlopende mate waarin juist die vaardigheden worden beheerst die de toets beoogt te meten. Dit is op zich niets bijzonders en geen reden tot twijfel over de kwaliteit/constructvaliditeit van de toets: het komt regelmatig voor dat de resultaten van verschillende bevolkingsgroepen, i.c. subgroepen van leerlingen, op toetsen/toetsitems verschillend zijn, omdat de ene subgroep gemiddeld vaardiger in het te meten construct is dan de andere. Als bijvoorbeeld de items van een taaltoets voor bepaalde (subgroepen) leerlingen moeilijker zijn dan voor andere, wordt in de meeste gevallen voldaan aan de functie van die items of de taaltoets als geheel: het discrimineren tussen meer en minder taalvaardige (groepen) leerlingen met betrekking tot de taal die wordt getoetst.

De tweede mogelijkheid is dat scoreverschillen tussen subgroepen veroorzaakt worden door verschillen die de toets/een item niet beoogt te meten, maar die ongewild toch meespelen of gemeten worden. Wanneer voor het juist beantwoorden van de items nog andere vaardigheden nodig zijn dan de vaardigheden die de items beogen te meten, kan afbreuk gedaan worden aan de constructvaliditeit van het meetinstrument. De constructvaliditeit is in het geding wanneer die benodigde additionele vaardigheden niet bij alle onderscheiden subgroepen (bijvoorbeeld autochtonen en allochtonen) in gelijke mate aanwezig zijn. In dat geval is er sprake van itembias. Dat kan bijvoorbeeld het geval zijn wanneer het niet tot het te meten construct behorende taalgebruik in een rekenitem voor bijvoorbeeld allochtone leerlingen dermate ingewikkeld is, dat ze daardoor niet of in onvoldoende mate aan het uitvoeren van de beoogde rekenoperatie toekomen of daaraan onvoldoende aandacht kunnen besteden. Allochtone leerlingen die over dezelfde (reken)vaardigheid beschikken als autochtone hebben dan ten onrechte en onbedoeld een geringere kans op een goed antwoord. Bij dit voorbeeld is er sprake van itembias ten nadele van allochtone leerlingen, maar het kan natuurlijk ook voorkomen dat een toets items bevat die partijdig zijn ten voordele van deze subgroep van leerlingen. Dat is het geval wanneer voor het correct kunnen oplossen van een item additionele vaardigheden vereist zijn waarover allochtone leerlingen in grotere mate beschikken dan gelijkpresterende autochtone.

Uit het bovenstaande zal duidelijk zijn dat itembias niet hetzelfde is als de moeilijkheidsgraad van een item. Daarnaast is gebleken dat het, alvorens onderzoek

naar itembias van start kan gaan, allereerst duidelijk moet zijn welke specifieke vaardigheid toetsitems beogen te meten.

Zoals uit 1 reeds naar voren is gekomen, was de bedoeling van het uitgevoerde onderzoek niet alleen om in de Eindtoets opgaven op te sporen waarbij van itembias sprake is, maar ook om potentiële inhoudelijke (met name talige) bronnen daarvan aan te geven.

2.2. Opzet en resultaten van het statistisch itembias-onderzoek

Ten behoeve van het totale project zijn over alle leerlingen die in 1987 (n = 80.685) en 1989 (n = 92.448) aan de Eindtoets Basisonderwijs deelnamen via een schriftelijke vragenlijst, die samen met de toetsopgaven aan hun leerkrachten was toegestuurd, een aantal achtergrondgegevens ingewonnen. De respons op deze vragenlijst bedroeg respectievelijk 73.2 (1987) en 67.8% (1989).

In het kader van het deelonderzoek naar itembias voor allochtone leerlingen was het uiteraard van belang om via de betreffende vragenlijst na te gaan tot welke etnische groepen de toetsdeelnemers behoren. De vaststelling daarvan is een ingewikkelde opgave (zie o.a. Extra & Verhoeven, 1993a), zeker wanneer daaromtrent slechts één vraag gesteld kan worden die voor leerkrachten snel en eenduidig beantwoordbaar moet zijn. Uiteindelijk werd besloten deze vraag te operationaliseren door te vragen naar het herkomstland van beide ouders, waarbij echter aangetekend moet worden dat bij éénouder gezinnen het herkomstland van de ouder bij wie het kind woont gold en bij tweede-generatiekinderen (bijvoorbeeld bij Chinezen en Molukkers) de herkomst van de grootouders in de beschouwing werd betrokken. Bij de indeling in etnische groepen werd in eerste instantie aangesloten bij de zesdeling van Extra & Vallen (1985) en Extra & Verhoeven (1993b): Nederland, mediterrane landen, ex-koloniale gebieden, Chinezen, politieke vluchtelingen en overige landen. Deze zes globale en intern zeer heterogene hoofdgroepen werden op basis van een schatting van het aantal te verwachten leerlingen in groep acht van het basisonderwijs en rekening houdend met het feit dat voor de geplande statistische analyses een ondergrens van zo'n 500 leerlingen per subgroep nodig is (vgl. bijvoorbeeld Intraprasert, 1986 en Zieky, 1993), opgesplitst in twaalf subgroepen: Nederlanders, Turken, Marokkanen, Chinezen, Molukkers, Antillianen, Surinamers (Creolen), Surinamers (Hindoestanen) en vier, meerdere herkomstlanden omvattende, restgroepen. Uiteindelijk bleken slechts de subgroepen van leerlingen uit twee etnische minderheidsgroepen omvangrijk genoeg ($n > 500$) om in de statistische itembiasanalyses te kunnen worden betrokken: Turkse leerlingen ($n = 797$ in 1987 en $n = 919$ in 1989) en Marokkaanse leerlingen ($n = 720$ in 1987 en $n = 907$ in 1989). Als referentiegroep werd een steekproef van autochtone leerlingen genomen ($n = 4969$ in 1987 en $n = 5000$ in 1989).

Voor statistisch onderzoek naar itembias zijn verschillende procedures beschikbaar, die in twee groepen verdeeld kunnen worden: procedures die werken volgens

klassieke testtheorie en procedures die werken volgens itemresponsetheorie (IRT). Zonder hier nader op de ins en outs en de pro's en contra's van en de verschillen tussen beide typen procedures in te gaan (zie daarvoor Uiterwijk 1994) zij opgemerkt dat uit eerder uitgevoerd statistisch itembiasonderzoek naar voren is gekomen dat IRT- en klassieke-testtheorieprocedures niet tot identieke resultaten leiden. Daarom werden in het onderhavige onderzoek technieken gebruikt die op beide typen procedures gebaseerd zijn. Als klassieke testtheorieprocedure werd de in de laatste jaren veel gebruikte *Mantel-Haenszel* techniek (MH) toegepast (zie o.a. Verhelst, 1988). Voor het itembiasonderzoek onder het IRT-model werd het computerprogramma *One Parameter Logistic Model* (OPLM) gebruikt (Verhelst, 1992).

Zoals verwacht, laten de resultaten van de statistische analyses naar itembias zien dat het uitermate lastig is om exact aan te geven hoeveel opgaven van de Eindtoets 1987 en 1989 (in totaal 360 items) partijdig zijn. De beide typen analyses laten namelijk een verschillend beeld zien. Globaal samengevat (zie voor verdere uitwerkingen Uiterwijk, 1994) komt dit op het volgende neer. Uit de analyses met de IRT-procedure blijkt dat 20 van de in totaal 360 geanalyseerde items (6%) partijdig zijn voor Turkse en/of Marokkaanse leerlingen. De analyses met de MH-procedure leveren een aantal van 45 partijdige items op (13%) voor Turken en/of Marokkanen. In totaal zijn er 13 items (4%) partijdig bij zowel de IRT- als de MH-procedure. Bij alle analyses werd een significantiegrens van 1% gehanteerd. Het zal duidelijk zijn dat de IRT-procedure minder gebiaste items opspoort dan de MH-techniek, een gegeven dat ook uit ander onderzoek naar voren komt.

Zoals reeds aangegeven in 2.1. kunnen gebiaste items partijdig zijn in het voordeel of in het nadeel van Turkse en Marokkaanse leerlingen. Uit de analyses komt naar voren dat van de 13 items die volgens de beide statistische procedures gebiast zijn er drie partijdig zijn in het voordeel van één of beide etnische groepen en tien in het nadeel. Verder blijken gebiaste items nooit partijdig te zijn in het voordeel voor Turkse leerlingen en tegelijkertijd in het nadeel voor Marokkaanse en omgekeerd.

Omdat, ondanks de kleine overlap, de resultaten op de beide gehanteerde statistische biasdetectieprocedures nogal verschillend zijn en omdat niet eenduidig kan worden vastgesteld welke van beide procedures duidelijk de voorkeur verdient (zie voor een vergelijking en een afweging van de voor- en nadelen Uiterwijk, 1994) zijn de opgaven die volgens beide procedures partijdig zijn betrokken in de activiteiten die in het kader van het project werden uitgevoerd om de inhoudelijke oorzaken van itembias op het spoor te komen. De twee experimenten die in dat opzicht werden ontwikkeld en uitgevoerd staan in de volgende paragrafen centraal.

3. Doelstelling, opzet en resultaten van een bij leerlingen afgenomen hardop-denken-procedure

3.1. Potentiële bronnen van itembias

Voorafgaand aan en tijdens de statistische biasanalyses werd op basis van een uitgebreide literatuurstudie een inventarisatie gemaakt van de problemen die allochtone basisschoolverlaters met het Nederlands als T2 ondervinden en van de problemen waarmee ze mogelijk geconfronteerd worden op grond van het feit dat hun culturele achtergrond verschillend is van de autochtone *mainstream* in Nederland. Uiteraard werd daarbij speciaal aandacht geschonken aan problemen die relevant geacht kunnen worden wanneer betreffende kinderen de Eindtoets moeten maken. De linguïstische en culturele factoren die uit die inventarisatie naar voren kwamen kunnen uiteraard niet zonder meer als inhoudelijke biasbronnen worden beschouwd, maar ze geven wel een indruk van de richting waarin gedacht moet worden voor het verkrijgen van een beeld omtrent potentiële linguïstische en culturele bronnen van itembias. In De Jong & Vallen (1989), Coenen & Vallen (1991), Uiterwijk & Vallen (1991) en Uiterwijk (1994) is uitvoerig verslag gedaan van deze inventarisatie. Daarom kan hier worden volstaan met een opsomming van de belangrijkste linguïstische en culturele probleemvelden. Wat het eerste betreft doen zich vooral problemen voor op het terrein van de Nederlandse woordenschat (o.a. woordkennis en kennis van woordcombinaties, moeilijkheidsgraad van woorden mede in samenhang met woordfrequentie, woordambiguïteit, woordsamenstelling), zinscomplexiteit (o.a. zinslengte, onderschikking, inbedding, vraagzinnen) en tekstcomplexiteit (o.a. tekstuele referenties/verwijswoorden, tekstsignalen). Wat de potentiële culturele biasbronnen betreft wordt in de literatuur vooral gewezen op de culturele lading van teksten en op cultureel bepaalde toetservaring. De elementen van deze groslijst kunnen elkaar op verschillende punten overlappen of met elkaar samenhangen. Het is dan ook niet toevallig dat bijvoorbeeld in veel onderzoek culturele voorkennis wordt gemeten met behulp van een woordenschattoets.

Vanwege de grote hoeveelheid en diversiteit van factoren was het noodzakelijk in het verdere onderzoek een inperking aan te brengen in de inhoudelijke biasanalyses. Besloten werd een accent op de linguïstische factoren te leggen en in een eventueel vervolgonderzoek de culturele factoren nader onder de loep te nemen.

3.2. Doelstelling en opzet

De bedoeling van het uitgevoerde hardop-denken-experiment was om na te gaan of en hoe vaak door qua niveau vergelijkbare allochtone en autochtone leerlingen bij een aantal oorspronkelijke (statistisch sterk ten nadele van allochtonen partijdige) opgaven van de Eindtoets 1987 een fout antwoord wordt gegeven ten gevolge van itemelementen die op grond van de uitgevoerde literatuurstudie als potentiële biasbronnen werden beschouwd. Tevens werd middels het experiment beoogd te

onderzoeken of en hoe vaak bij gemanipuleerde items ten gevolge van de itemmanipulatie een goed antwoord wordt gegeven. Dezelfde items werden dus in hun oorspronkelijke en in een gemanipuleerde vorm aan de leerlingen voorgelegd. Ter verduidelijking zij opgemerkt dat het bij gemanipuleerde items gaat om items waarbij het itemelement dat als potentiële biasbron wordt beschouwd (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 deelnemende autochtone en allochtone leerlingen moesten eerst de hen voorgelegde oorspronkelijke, respectievelijk gemanipuleerde items goed bestuderen, daarna het naar hun oordeel goede antwoord aankruisen en tot slot zo uitgebreid en nauwkeurig mogelijk mondeling toelichten hoe ze de taakstelling van elk item opgelost hadden. Als een leerling een item fout oploste of een onduidelijke toelichting gaf werd door de proefleidster doorgevraagd enerzijds om achter de foutenbron te komen en anderzijds om na te gaan of de leerling de door het item gemeten vaardigheid beheerste. De gesprekken werden op audio-cassette vastgelegd. Bij het afluisteren van de opnamen werd er vooral op gelet of de itemelementen die een potentiële biasbron konden zijn, voor de leerlingen inderdaad een probleem vormden bij het oplossen van het item.

Tegen de achtergrond van de problematiek bias versus moeilijkheidsgraad (zie 2.1.) is het volgende nog van belang met het oog op de uitgevoerde itemmanipulaties. Er mag in verband met de constructvaliditeit verondersteld worden dat in bijvoorbeeld rekenopgaven de talige context in hoge mate communaal is of op z'n minst niet zo'n hoge taalvaardigheid Nederlands veronderstelt dat hij op grond daarvan discrimineert tussen allochtone en autochtone leerlingen. Indien de talige context wel tussen beide subgroepen van leerlingen zou discrimineren, dan moeten de in talig opzicht lastige itemelementen vervangen worden door talige elementen die wel communaal zijn. Bij taalopgaven kan ook onderscheid gemaakt worden tussen de (taal)vaardigheid die het item beoogt te meten en de daarvoor benodigde talige context. Een taalitem moet de verschillen tussen leerlingen blootleggen in zoverre het gaat om hetgeen het item wil meten, maar net als bij rekenen moet ook hier de talige context communaal zijn. In het onderdeel Taal van de Eindtoets staan teksten waarin opzettelijk talige tekortkomingen zijn aangebracht. Deze taalopgaven vragen aan de leerlingen om te beoordelen of een woord of een zinsconstructie in een bepaalde tekst of de opbouw van die tekst al dan niet in orde is. In een aantal gevallen moeten de leerlingen daarbij ook verbeteringen in de tekst aanbrengen. In de praktijk is het bij taalopgaven vaak moeilijk om aan te geven waar de scheiding ligt tussen de taalvaardigheid die een item meet en de talige context ervan. Zo kan een item vragen na te gaan of in een tekst een bepaald verwijswoord juist gebruikt is. De inhoud van de tekst als geheel alsmede die van de zin(nen) waarin het verwijswoord en zijn referent staan (de talige context) spelen voor het correct beantwoorden van de opgave een belangrijke rol, naast inzicht in de linguïstische conventies inzake de relatie verwijswoord en referent (de te meten vaardigheid). De verwevenheid van context

en te meten (taal)vaardigheid maakt het bij een groot deel van de taalopgaven uitermate moeilijk om deze te manipuleren zonder de constructvaliditeit aan te tasten.

Als partijdige opgaven op de juiste wijze gemanipuleerd worden, is te verwachten dat allochtone leerlingen bij de gemanipuleerde opgaven minder fouten maken dan bij de oorspronkelijke items en dat allochtone en autochtone leerlingen met hetzelfde (reken)vaardigheidsniveau ongeveer evenveel (gemanipuleerde) opgaven goed maken. Wellicht dat de items dan bij beide subgroepen leerlingen beter de vaardigheid meten die ze beogen te meten. Van belang hierbij blijft uiteraard wel dat het oorspronkelijke en het gemanipuleerde item dezelfde vaardigheid blijven meten.

Ten behoeve van het hardop-denken-experiment werden 17 opgaven uit de toetsonderdelen Taal (5 items), Rekenen (8 items) en Informatieverwerking (4 items) van de Eindtoets 1987 geselecteerd die bij de MH-analyses sterk partijdig in het nadeel van Turkse en/of Marokkaanse leerlingen waren. Bij de keuze van de items speelde tevens een rol dat ze zodanig gemanipuleerd moesten kunnen worden dat de vaardigheid die ze in hun oorspronkelijke vorm pretenderen te meten door de manipulatie niet werd aangetast. Dit had tot gevolg dat de opgaven slechts minimaal gemanipuleerd werden. De manipulaties hadden vooral betrekking op onnodig moeilijke woorden, complexe grammaticale en/of ambigue zinsconstructies en op impliciete zins- en tekststructuren. In enkele gevallen vonden ook manipulaties van de grafische contexten plaats om veronderstelde onduidelijkheden in tekeningen, kaarten en tabellen te verwijderen. De gekozen items kunnen niet beschouwd worden als een representatieve steekproef uit alle Eindtoetsopgaven. De 17 items zijn in toetsversie A in hun oorspronkelijke vorm getoetst en in toetsversie B in de gemanipuleerde vorm. Ze zijn voorgelegd aan 44 leerlingen uit groep acht van vijf basisscholen. In concreto ging het daarbij om 22 paren van telkens qua niveau vergelijkbare autochtone en allochtone leerlingen. Aan de ene helft van de paren werden de oorspronkelijke items (toetsversie A) ter oplossing en bespreking voorgelegd, aan de andere helft de gemanipuleerde (toetsversie B).

Om het effect van de manipulaties goed te kunnen nagaan hadden in principe beide toetsversies aan dezelfde leerlingen moeten worden voorgelegd, uiteraard met een interval van enkele maanden. Het nadeel van een dergelijke opzet is echter dat er dan een grote kans bestaat dat de leerlingen zich bij de tweede afname bepaalde opgaven herinneren. Met andere woorden, de geheugenfactor speelt dan een niet te controleren rol. In overleg met de leerkrachten werd daarom telkens bij elke allochtone leerling een autochtone leerling geselecteerd die vergelijkbaar was op factoren die voor schoolsucces van belang zijn, zoals sociaal-economische achtergrond, taalvaardigheid Nederlands, rekenvaardigheid, motivatie, doubleergeschiedenis, Eindtoetsscore en schoolkeuze-advies van de basisschool. Alle allochtone leerlingen moesten bovendien voldoen aan twee criteria. Ze moesten thuis een etnische minderheidstaal spreken en ze moesten minimaal vanaf groep drie het Nederlandse basisonderwijs volgen. Bij de verdeling van de leerlingen over beide toetsversies is erop gelet dat het prestatieniveau van beide groepen leerlingen zoveel mogelijk vergelijkbaar was.

Onder de leerlingen die toetsversie A dan wel B maken, zitten evenveel leerlingen met een LBO-, MAVO-of HAVO-advies. Bij de autochtone en allochtone leerlingen was het aantal jongens en meisjes gelijk. De groep allochtone leerlingen bestond uit elf Turkse en acht Marokkaanse leerlingen en één Chinese, één Antilliaanse en één Braziliaanse leerling. De Turkse en Marokkaanse leerlingen waren nagenoeg gelijk verdeeld over beide toetsversies.

3.3. Resultaten

De gemiddelde scores van de autochtone en allochtone leerlingen die toetsversie A, respectievelijk toetsversie B maakten, geven een eerste indicatie voor het effect dat de itemmanipulatie heeft gehad. Het betreft slechts een indicatie, omdat het aantal leerlingen per cel (n=11) en het aantal items (k=17) gering is en omdat niet met volledige zekerheid kan worden gezegd of de onderscheiden subgroepen exact even vaardig zijn in wat de items beogen te meten. In tabel 1 staat per onderscheiden subgroep het gemiddelde percentage goed gemaakte opgaven.

Tabel 1: Gemiddeld percentage goed gemaakte antwoorden per subgroep

Toetsversie	Allochtonen (n = 11)	Autochtonen (n = 11)
Taal		
Versie A: k=5	58.2	78.2
Versie B: k=5	67.3	87.3
Verschil B-A	9.1	9.1
Rekenen		
Versie A: k=8	36.4	55.7
Versie B: k=8	53.4	58.0
Verschil B-A	17.0	2.3
Informatieverwerking		
Versie A: k=4	52.3	84.1
Versie B: k=4	72.7	84.1
Verschil B-A	20.4	0
Totaal		
Versie A: k=17	46.5	69.0
Versie B: k=17	62.0	72.7
Verschil B-A	15.5	3.7

Uit Tabel 1 blijkt dat de allochtone leerlingen die de gemanipuleerde items gemaakt hebben, in totaal gemiddeld 15.5% meer items goed maken dan de allochtone leerlingen die de oorspronkelijke items hebben gemaakt. Bij de autochtone leerlingen bedraagt het verschil slechts 3.7%. Het verschil tussen het gemiddelde percentage goed van de allochtone en autochtone leerlingen die de oorspronkelijke toetsversie maakten, is 22.5%. Bij de gemanipuleerde versie bedraagt het gemiddeld

verschilpercentage tussen beide groepen nog 10.7%. De gegevens uit Tabel 2 laten verder zien dat de allochtone leerlingen grosso modo meer geprofiteerd hebben van de itemmanipulaties dan de autochtone. Verder lijken de itemmanipulaties differentiële effecten te hebben voor de items uit de drie toetsonderdelen van de Eindtoets. Bij het onderdeel Informatieverwerking maken de allochtone leerlingen alle gemanipuleerde items beter dan de oorspronkelijke (wat een toename in het gemiddeld aantal goed gemaakte opgaven van ruim 20% oplevert), terwijl de autochtonen geen profijt van de manipulaties hebben gehad (hun score blijft hetzelfde). Ook drie van de acht rekenitems zijn door de allochtone leerlingen in de gemanipuleerde versie beter gemaakt dan in de oorspronkelijke versie. Bij de overige rekenitems zijn er nauwelijks verschillen. De toename in de correctscore bij de gemanipuleerde rekenitems bedraagt voor de allochtonen 17%, terwijl de correctscore voor de autochtonen nauwelijks stijgt (ruim 2%). Zowel de oorspronkelijke als de gemanipuleerde items worden door de autochtonen beter gemaakt dan door de allochtonen, maar het verschil tussen beide subgroepen is bij de gemanipuleerde items veel kleiner geworden. Bij Rekenen is het verschil gedaald van 19.3 naar 4.6% en bij Informatieverwerking van 31.8 naar 11.4%.

De itemmanipulaties bij het onderdeel Taal zijn minder succesvol geweest; de autochtonen lijken er in gelijke mate van te hebben geprofiteerd als de allochtonen (ruim 9% toename van de correctscore). Dit zou erop kunnen wijzen dat met de talige manipulatie van de taalitems niet de mogelijke biasbron geëlimineerd is, maar dat de moeilijkheidsgraad van de items veranderd is (de items zijn voor iedereen gemakkelijker geworden). Ook het gegeven dat er naast de drie gemanipuleerde taalitems die door de allochtone leerlingen beter werden gemaakt twee zijn die in hun oorspronkelijke vorm beter werden gemaakt, zou een indicatie in die richting kunnen geven.

Over de vraag of de veronderstelde inhoudelijke biasbronnen in de 17 bij het hardop-denken-experiment betrokken opgaven ook werkelijk voor allochtonen een rol speelden, moesten de geregistreerde leerlinguitspraken en de op basis daarvan gemaakte protocolanalyses uitsluitsel geven. Dat was geen eenvoudige opgave, omdat uit de protocollen voor de afzonderlijke items en de afzonderlijke kinderen verschillende beelden naar voren kwamen. Bij sommige items die door allochtone leerlingen vaker fout werden beantwoord dan door autochtone bestond wel de indruk dat de te meten vaardigheid werd beheerst, maar bij andere niet. Bij sommige items, die in afzonderlijke concrete stappen opgelost moeten worden, konden deze (deel)vaardigheden apart bevraagd worden, maar dat was minder eenvoudig of onmogelijk bij items waarbij dat niet het geval is. Om een zo duidelijk mogelijke indruk te geven behandelen we daarom in het onderstaande bij wijze van voorbeeld twee rekenopgaven en een opgave voor informatieverwerking. De doelstelling van die drie statistisch sterk in het nadeel van allochtone leerlingen gebiaste items heeft primair met de meting van andere vaardigheden van doen dan met de meting van taalvaardigheid Nederlands. De kwestie van interferentie van taalbias en beoogde moeilijkheid behoort daarom bij deze items niet aan de orde te zijn. De items worden

telkens eerst in hun oorspronkelijke en daarna in hun gemanipuleerde vorm gepresenteerd. Het gemanipuleerde deel van het item is steeds gecursiveerd.

Oorspronkelijke versie (1987; Rekenen 57)

Vader koopt een naaimachine. Deze kost *f* 800,- zonder B.T.W. De B.T.W. is 20%. Hoeveel moet vader betalen inclusief B.T.W.?

A *f* 160,-
B *f* 640,-
C *f* 820,-
D *f* 960,-

Gemanipuleerde versie

Vader koopt een naaimachine. Deze kost *f* 800,- zonder B.T.W. De B.T.W. is 20%. *Wat moet vader voor de naaimachine betalen met B.T.W.?*

A *f* 160,-
B *f* 640,-
C *f* 820,-
D *f* 960,-

De talige manipulatie van dit rekenitem bestaat met name hierin dat de vraag anders is geformuleerd (o.a. een w-vraag in plaats van een h-vraag) en dat het woord *inclusief* is vervangen door het woord *met*.

De oorspronkelijke versie wordt door zes van de elf allochtone leerlingen fout beantwoord en door twee van de elf autochtone. Bij de gemanipuleerde versie geven nog twee allochtone leerlingen en één autochtone leerling een fout antwoord. Uit de protocol-analyses blijkt dat bij drie allochtone leerlingen en één autochtone het woord *inclusief* duidelijk de oorzaak van de fout is, waarvan er één (allochtone) echter ook problemen heeft met het berekenen van procenten. Als in plaats van *inclusief* de aanduiding *met* in het oorspronkelijke item zou zijn gebruikt, had dat (naar het eigen oordeel van de leerlingen) voor twee allochtone leerlingen geen verschil uitgemaakt, maar vier allochtone en twee autochtone leerlingen geven te kennen dat de opgave daardoor voor hen begrijpelijker zou zijn geweest.

Oorspronkelijke versie (1987; Rekenen 52)

Per jaar gaf een gezin gemiddeld *f* 1500,- uit aan aardappels en groenten. Ze wilden bezuinigen. Ze huurden een jaar een tuin van 200 vierkante meter voor *f* 2,- per vierkante meter. De overige onkosten waren *f* 80,-. Ze moesten nog voor een bedrag van *f* 400,- aan aardappelen en groenten in de winkel kopen. De rest kwam uit de tuin.
Hoeveel had dat gezin bespaard met tuinieren?

A *f* 620,-
B *f* 820,-
C *f* 920,-
D *f* 1020,-

Gemanipuleerde versie

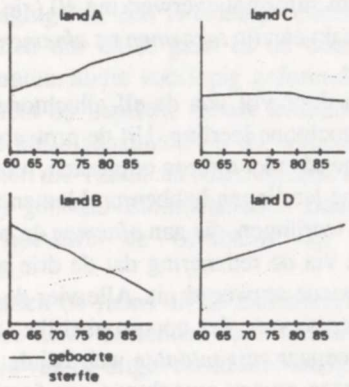
Per jaar koopt mevrouw Knol voor f 1500,- groenten. Dat vindt ze teveel geld. Daarom huurt ze voor een jaar een groentetuin van 200 vierkante meter. Die tuin kost f 2,- per vierkante meter. De spullen die ze voor de tuin nodig heeft kosten nog f 80,-. Omdat de tuin veel groenten oplevert hoeft mevrouw Knol nog maar voor f 400,- aan groenten in de winkel te kopen. Hoeveel geld heeft mevrouw Knol met haar groentetuin bespaard?

A *f* 620,-
B *f* 820,-
C *f* 920,-
D *f* 1020,-

De leerlingen moeten bij dit item uit de tekst opmaken welke bewerkingen ze achtereenvolgens moeten uitvoeren. Door het relatief grote aantal getallen dat in een omvangrijke talige context is ingebed en gebruikt moet worden, doet deze rekenopgave een groot beroep op taalvaardigheid Nederlands. Het is dan ook niet zo verwonderlijk dat de meerderheid van de allochtone leerlingen grote problemen met de oorspronkelijke opgave heeft. Daarom is in de gemanipuleerde versie (die met name door de allochtonen beduidend beter werd gemaakt) met meer expliciete en duidelijker op elkaar aansluitende formuleringen gewerkt. Daarnaast heeft personificatie plaatsgevonden (*mevrouw Knol* in plaats van de verzamelnaam *gezin*). Verder bevat de oorspronkelijke opgave niet alleen veel maar ook een aantal complexe verwijzingen (bijvoorbeeld het verwijswoord *ze* in de tweede zin, dat getalsmatig afwijkt van zijn antecedent in de eerste zin, *een gezin*). Uit de literatuur is het bekend dat de verwijzingssystematiek van het Nederlands voor allochtonen een specifiek probleem vormt dat ze pas in een vrij laat T2-verwervingsstadium onder de knie krijgen.

Oorspronkelijke versie (1987; Informatieverwerking 40)

Sterfte- en geboortecijfers in vier landen



Van welk land kan men zeggen dat het aantal geboorten toeneemt en het aantal sterfgevallen afneemt?

A van land A
B van land B
C van land C
D van land D

Gemanipuleerde versie

Zie de grafieken 'Sterfte- en geboortecijfers in vier landen' van de bovenstaande oorspronkelijke versie van het item.

In welk land is tussen 1960 en 1985 het aantal geboorten gestegen en het aantal sterfgevallen gedaald?

A in land A
B in land B
C in land C
D in land D

Bij de overige rekenitems doen zich soortgelijke en andere talige problemen bij allochtonen voor. Bij een opgave waarin de vraag gesteld wordt *Welk bootje heeft in verhouding tot zijn lengte de langste mast?* wordt *zijn lengte* vaak opgevat als de lengte van de mast. Ook blijkt een aanduiding *de overige onkosten* (zie hiervoor bij het ongemanipuleerde rekenitem 52) zorg voor heel wat problemen en de aanduiding

een half procent wordt door sommige allochtone leerlingen begripsmatig verward met *de helft*. Zo nu en dan leveren ook de bij de rekenitems behorende tekeningen problemen voor allochtonen op.

In de gemanipuleerde versie van item Informatieverwerking 40 (zie vorig pagina) is de vraagstelling meer expliciet gemaakt en zijn *toenemen* en *afnemen* vervangen door respectievelijk *stijgen* en *dalen*.

De oorspronkelijke versie wordt door vijf van de elf allochtone leerlingen fout beantwoord en door geen enkele autochtone leerling. Uit de protocol-analyses blijkt dat twee allochtone leerlingen de legenda niet meteen opmerken, omdat die erg dicht op de grafieken staat. Vier allochtone leerlingen hebben problemen met de woorden *toenemen* en *afnemen*. Een van deze leerlingen, die aan *afnemen* de betekenis *iemand zijn spullen afpakken* toekent, komt via de redenering dat de drie andere grafieken 'raar' zijn toch uiteindelijk op het goede antwoord uit. Alle vier de allochtonen die problemen hebben met de betekenis van de oorspronkelijke combinatie van werkwoorden, begrepen het begrippenpaar *stijgen/dalen* wel. Bij de gemanipuleerde versie gaven twee allochtone leerlingen en een autochtone een fout antwoord.

Bij de overige opgaven voor Informatieverwerking spelen vooral visuele onduidelijkheden een rol, maar het kon niet vastgesteld worden of dat voor allochtonen en autochtonen in verschillende mate geldt.

Met inachtneming van de in het voorafgaande genoemde beperkingen (o.a. klein aantal deelnemers en items, geen complete zekerheid of de oorspronkelijke en de gemanipuleerde items hetzelfde meten en of de leerlingen die beide versies maakten even vaardig zijn in wat de items beogen te meten) kan gezegd worden dat de hardop-denken-procedure aanwijzingen heeft gegeven dat talige biasbronnen voor allochtone leerlingen voor een groot deel op het gebied van woordgebruik en impliciete zins- en tekstverbanden gezocht moeten worden. Daarnaast leiden ongebruikelijke uitdrukkingen en woordvormgelijkenissen tot problemen. De complexiteit van de opgaven speelt eveneens een rol. Complexe items vereisen doorgaans meer context en voor het oplossen van dergelijke items moet de leerling meestal een aantal tussenstappen maken. Welke dat zijn moet meestal uit de talige context afgeleid worden. Door hun geringere taalvaardigheid Nederlands kunnen allochtone leerlingen derhalve meer moeite met complexe items hebben. Dat geldt des te meer wanneer er in dergelijke items veel verwijswoorden voorkomen. Verder blijft het uiteraard mogelijk dat de itemcontext voor allochtone leerlingen minder herkenbaar is dan voor autochtone. Dat geldt met name voor contextmateriaal dat cruciaal is voor het oplossen van een item. Meer uitgebreide informatie over het hardop-denken-experiment wordt gegeven in Coenen & Vallen (1991) en Uiterwijk (1994).

4. Doelstelling, opzet en resultaten van de expert-bevraging

4.1. Doelstelling en opzet

De expert-bevraging had een tweeledige doelstelling. Binnen het project werd de noodzaak ingezien om na te gaan of de door de onderzoekers op basis van de uitgevoerde literatuurstudie voorlopig geformuleerde (zie 3.1.) - en via het hardop-denken-experiment op beperkte schaal tentatief onderzochte (zie 3.3.) - potentiële bronnen van itembias aansluiten bij de oordelen die ter zake deskundigen hebben over de problemen die Turkse en Marokkaanse leerlingen ondervinden bij het maken van (statistisch) gebiaste Eindtoetsitems. Daarnaast werd met dit deelonderzoek nagegaan in hoeverre de oordelen van de bevraagde experts onderling overeenstemmen.

In totaal werden 84 items uit de Eindtoetsen van 1987 en 1989 aan 16 experts voorgelegd. Deze items bestonden voor verreweg het grootste deel uit opgaven die bij minimaal drie van de uitgevoerde MH-analyses (zie 2.2) significant partijdig zijn in het voordeel of het nadeel van Turkse en/of Marokkaanse leerlingen. Om het aandeel van de items ten voordele van beide groepen leerlingen wat te verhogen bestond een relatief klein deel van de items uit opgaven die bij alle MH-analyses niet-significant in het voordeel van beide groepen leerlingen zijn. De totale aan de experts voorgelegde itemlijst bestond uit 37 taalopgaven, 31 rekenopgaven en 16 opgaven voor informatieverwerking.

De experts is gevraagd om kenbaar te maken welke items en itemelementen naar hun oordeel moeilijker dan wel gemakkelijker zijn voor allochtone leerlingen en om de oorzaken daarvoor zo uitgebreid mogelijk te expliciteren. Bovendien werd hen gevraagd om bij hun beoordeling en explicatie zoveel mogelijk te differentiëren naar Turkse en Marokkaanse kinderen. De experts wisten uiteraard vooraf niet of een item partijdig is ten voor- of nadele van deze leerlingen.

Zoals uit het bovenstaande blijkt, is de experts niet gevraagd om bij hun beoordeling een onderscheid te maken tussen items die gebiast zijn ten voor- of ten nadele van beide allochtone leerlinggroepen, omdat de experts dan ook duidelijk voor ogen zouden moeten hebben ten aanzien van welke (deel)vaardigheden allochtone en autochtone leerlingen een gelijk prestatieniveau hebben (zie de omschrijving van itembias in 2.1), wat natuurlijk niet het geval kon zijn. De dimensie 'moeilijk(er)-gemakkelijk(er)' geeft weliswaar op een andere wijze aan of een item in het voor- of het nadeel van allochtone leerlingen is, maar kan niettemin als een variatie op de dimensie 'partijdig in het voor- of nadeel' worden beschouwd.

De beoordelingen werden door de experts per afzonderlijk item gegeven in een mondeling interview door één van de projectmedewerkers. Daardoor bestond de mogelijkheid om verdere toelichting en explicatie te vragen. Alle interviews werden op audio-cassette opgenomen en achteraf bestudeerd. De complete lijst met te beoordelen opgaven werd enkele weken voorafgaand aan het interview aan betrokkenen toegezonden, zodat deze zich terdege konden voorbereiden.

Van de 16 bevraagde deskundigen waren er 11 van autochtone en 5 van allochtone afkomst. Drie deskundigen waren werkzaam in de dagelijkse praktijk van het basisonderwijs (2 allochtone onderwijsgevenden en 1 autochtone) en 2 autochtone experts waren bij het Cito werkzaam als toetsconstructeurs. De overige 11 waren als taalkundigen (2 allochtone en 7 autochtonen) of als onderwijskundigen (1 allochtoon en 1 autochtoon) werkzaam aan de universiteit, maar hadden allen ervaring in en/of gebleken belangstelling voor toets- en/of curriculumontwikkeling. Omdat ze op linguïstisch en/of cultureel terrein mogelijk verschillende inzichten, suggesties of ideeën zouden kunnen verschaffen, werden zowel autochtone als allochtone deskundigen bij het onderzoek betrokken.

Verdere informatie over opzet, afnameprocedure, wijze van analyseren alsmede een samenvatting van de gevoerde gesprekken en een uitgebreidere bespreking van de onderstaande resultaten zijn opgenomen in Van de Waal-Heijkants (1992) en Uiterwijk (1994).

4.2. Resultaten

Uit de analyses van de met de experts gevoerde gesprekken komt in de eerste plaats naar voren dat het geven van een oordeel over de moeilijkheidsgraad van items of itemonderdelen voor allochtone basisschoolverlaters niet eenvoudig is. Zo maken de experts qua moeilijkheidsgraad voor allochtonen bijvoorbeeld geen onderscheid tussen de rekenitems die op de volgende pagina staan, terwijl de statistische biasanalyses daar overduidelijk wel aanleiding toe geven. De eerste opgave (1989; Rekenen 27) is (significant) partijdig *in het nadeel* van zowel Turkse als Marokkaanse leerlingen, terwijl het tweede (1989; Rekenen 57), weliswaar niet significant, een bias *ten voordele* van beide groepen leerlingen laat zien.

Nagenoeg alle experts beschouwden deze twee items als gelijkwaardig op de dimensie moeilijk-gemakkelijk. Een van hen merkt over item 27 op dat de referentie *dat* en de formulering *1 van elke 2* moeilijke elementen voor allochtone leerlingen zijn. Twee experts zijn van oordeel dat bij item 57 voor allochtone leerlingen de moeilijkheid in de complexiteit van de opdracht zit. Alle andere deskundigen voorzien voor allochtone leerlingen geen problemen bij beide opgaven.

Er is uiteraard onderzocht in welke mate de experts erin slagen aan te geven of een item voor allochtone leerlingen moeilijker is dan voor autochtone. Per item is daarom nagegaan of meer dan de helft van de respondenten zegt dat het item moeilijker is voor allochtone leerlingen en of dat item inderdaad ook partijdig is in het nadeel van deze groep leerlingen. Op dit punt bleek er bij 31 van de 84 items (37%) geen overeenstemming tussen het oordeel van de meerderheid van de experts en de richting (voor- of nadeel) van de partijdigheid.

1989; Rekenen 27

Op de Arkschool is 1 van elke 2 kinderen lid van een club.
Hoeveel procent is dat?

A ½%
B 33¼%
C 50%
D 100%

1989; Rekenen 57

De olieprijs daalde van 20 tot 15 dollar per vat.
Hoeveel procent daalde de prijs?

A 3%
B 4%
C 5%
D 25%

Tevens is vastgesteld hoe hoog de samenhang is tussen het aantal experts dat zegt dat een item moeilijker voor allochtone leerlingen is en de mate van itembias (aantal keren in het voordeel, respectievelijk in het nadeel van allochtone leerlingen). Deze samenhang is niet hoog te noemen: $r = .30$ ($p<.01$). Feit is natuurlijk dat de dimensies 'partijdig in het voor- of nadeel' en 'gemakkelijk-moeilijk' niet hetzelfde zijn, maar het is niet aannemelijk te veronderstellen dat dit onderscheid de geringe trefzekerheid van de experts verklaart.

Verder kwam uit de analyses naar voren dat de experts niet of nauwelijks onderscheid maken tussen moeilijke/gemakkelijke items voor Turkse en Marokkaanse leerlingen.

De item-oordelen van de experts op de dimensie 'moeilijk-gemakkelijk' komen onderling sterk overeen en sluiten bovendien in hoge mate aan bij de bevindingen van de projectonderzoekers. Ook de experts benadrukken dat items die vragen naar woordkennis en kennis van idiomatische uitdrukkingen een grote kans maken moeilijk te zijn voor allochtonen. Daarnaast wijzen de experts inzake de moeilijkheden voor allochtonen op de problematiek van de cultureel bepaalde voorkennis die voor het kunnen maken van sommige items vereist is. Deze voorkennis kan volgens hen bijvoorbeeld een aanzienlijke rol spelen bij die items waarbij een uitvoerige tekst als contextmateriaal fungeert. Bij een groot aantal van de voorgelegde items waren de experts van oordeel dat deze geschreven zijn vanuit een Nederlandse cultuurkennis en daardoor voor veel allochtone leerlingen minder herkenbaar zijn.

Zeer opmerkelijk is dat de experts niet erg trefzeker zijn in het onderscheiden van items in het voor- dan wel in het nadeel van Turkse en Marokkaanse leerlingen, maar toch in hoge mate aansluiten bij de oordelen van de onderzoekers over bronnen van itembias. De voor de experts onbekende reden hiervan zou kunnen zijn dat bij partijdige items in het voordeel van Turkse en/of Marokkaanse leerlingen vaak dezelfde inhoudelijke bron van itembias aan de orde is als bij items die partijdig zijn in het nadeel. Het aantal van eerstgenoemde is echter aanzienlijk kleiner (zie 2.2.). Door dit gegeven wordt mogelijk de betekenis afgezwakt van enkele eerder als inhoudelijke bron van itembias genoemde categorieën. Itemclusters die uitsluitend partijdige items in het nadeel van Turkse en/of Marokkaanse leerlingen kennen zijn: woordkennis en kennis van woordcombinaties en rekenitems met relatief veel context. Bij het clusters Spelling is de situatie overigens omgekeerd aan de hierboven geschetste: de meeste partijdige spellingitems zijn in het voordeel van Turkse en/of Marokkaanse leerlingen en slechts enkele in hun nadeel.

5. Samenvattende conclusies en (voorzichtige) praktijksuggesties

Op basis van de uitgevoerde literatuurstudie en de voorafgaande paragrafen kan allereerst worden geconcludeerd dat onderzoek naar itembias nog met veel onzekerheden is omgeven. Ook het door ons uitgevoerde onderzoek naar itembias in de Eindtoetsen Basisonderwijs van 1987 en 1989 laat duidelijk zien dat het op dit moment, gezien de stand van de theorie-ontwikkeling en de onderzoeksmatige mogelijkheden, uitermate lastig is om empirisch stevig gefundeerde uitspraken te doen over de vraag welke items om welke inhoudelijke reden(en) partijdig zijn in het voor- dan wel in het nadeel van allochtone (met name Turkse en Marokkaanse) leerlingen. Een en ander neemt niet weg dat ons onderzoek ook inzichten heeft opgeleverd over de vraag in welke richting verder gewerkt zou moeten worden en bovendien plausibele bronnen van talig-inhoudelijke itembias heeft gegenereerd. Die plausibiliteit is vooral daarin gelegen dat de resultaten van de diverse procedures om inhoudelijke biasbronnen op het spoor te komen bij items die zowel volgens de MH- als de IRT-procedure partijdig zijn op een aantal punten overeenstemmen. Wanneer immers verschillende procedures in de richting van dezelfde biasbronnen wijzen, dan is de mate van zekerheid hierover groter. Wat ons betreft geldt dat laatste vooral ten aanzien van de volgende bronnen van itembias.

Tekst- en zinsbegrip (in termen van begrijpend lezen)
Items die op macroniveau naar globaal tekstbegrip vragen, hebben kans op itembias in het voordeel van Turkse en Marokkaanse leerlingen. Hierbij gaat het om vragen als *Wat wil de schrijver in de eerste 10 regels vooral duidelijk maken?* of *Welke conclusie kun je uit de laatste alinea trekken?*

Items die op meso- en microniveau een beroep doen op tekstbegrip, zoals dat het geval is bij vragen naar de betekenis van zinnen of verbanden tussen zinnen of vragen naar de verbanden tussen of de betekenis van woordgroepen of combinaties

van woordgroepen, hebben een gerede kans op itembias in het nadeel van allochtone kinderen. Dit geldt in het bijzonder wanneer om een woordelijke of geparafraseerde herhaling van expliciet in de tekst gegeven informatie wordt gevraagd. Het verschillende beeld dat naar voren komt ten aanzien van tekstbegrip op macro- versus meso-/microniveau zien we ook in Hacquebord (1989).

In aansluiting bij het bovenstaande kan nog worden opgemerkt dat ook moeilijke referenties (bijvoorbeeld over langere tekstpassages heen of verwijzingen waarbij het bijwoord *er* een functie vervult) of potentieel ambigue referenties (zoals het verwijzende *het* en verwijswoorden als *die*, *deze* en *ze*, die een enkelvoudig of een meervoudig antecedent kunnen hebben) tot itembias in het nadeel van allochtonen kunnen leiden.

Woordkennis en kennis van woordcombinaties

Items die vragen naar de betekenis van woorden en combinaties van woorden en waarbij de betekenis van die woorden niet of moeilijk uit de context van het item kan worden afgeleid, hebben eveneens een sterke kans op itembias in het nadeel van allochtone leerlingen.

Correct taalgebruik

Items die betrekking hebben op de kennis van de vorm van vaste (letterlijk of figuurlijk gebruikte) woordcombinaties (bijvoorbeeld *ergens zonder kleerscheuren afkomen*) en/of conventies op het gebied van de zinsbouw (zinsvolgorde, inversie/vraagvormen) hebben eveneens een kans om partijdig te zijn in het nadeel van Turkse en Marokkaanse leerlingen.

Spelling

Items die vragen om spellingfouten in werkwoorden en in woorden met een vast woordbeeld aan te geven, hebben een kans op itembias in het voordeel van allochtone leerlingen.

De mate van zekerheid inzake het optreden van bias en bronnen van itembias is bij andere, hierboven niet gememoreerde clusters geringer, omdat die items bijvoorbeeld alleen partijdig zijn volgens de MH-procedure. Dat geldt ook voor de hierboven vermelde opgaven waarin referenties een cruciale rol spelen. Toch kan daarbij sprake zijn van een wat grotere mate van zekerheid, omdat referenties bij twee typen vraagclusters als belangrijke biasbron in alle uitgevoerde exercities naar voren komen: bij tekst- en zinsbegrip wanneer expliciet naar de betekenisrelaties wordt gevraagd en bij rekenitems met relatief veel context.

Omdat er geen sprake kan zijn van volledige overeenstemming tussen de uitkomsten van de verschillende statistische itembiasdetectieprocedures kan op dit moment niet met volstrekte zekerheid ten aanzien van het merendeel van de 360 onderzochte items worden aangegeven of deze partijdig zijn of niet. Bij een klein gedeelte van de items is die overeenstemming er wel en is de duidelijkheid dus groter. Door het

hanteren van meerdere procedures kunnen verder echter ook verschillen in de gradaties in de partijdigheid van items worden opgespoord, die richting kunnen geven aan verder onderzoek.

Ook de vraag welk element in een item verantwoordelijk is voor itembias kan in de meeste gevallen vooralsnog nog niet eenduidig worden beantwoord. De bron van itembias kan bijvoorbeeld in het contextmateriaal, in de vraagstelling, de antwoordmogelijkheden en in de te meten (deel)vaardigheid zitten. Naarmate het contextmateriaal omvangrijker wordt, is het moeilijker de biasbron op het spoor te komen en aan te wijzen. Verder kan ook de voorkennis over hetgeen in teksten aan de orde wordt gesteld een potentiële biasbron zijn. Scheuneman (1985), Uiterwijk & Vallen (1991) en Schmitt e.a. (1992) geven aan dat via verschillende wegen (inhoudsanalyse, expertbevraging, experimenten) informatie over biasbronnen verzameld kan worden. Als uiteindelijk - zoals in ons onderzoek - met een zekere mate van waarschijnlijkheid een aantal biasbronnen zijn opgespoord, dan blijft nog de vraag of die biasbronnen ook daadwerkelijk afbreuk doen aan de constructvaliditeit van een toets of niet.

Algemeen kan daarover worden opgemerkt dat wanneer een bron van itembias tot de te meten vaardigheid behoort, het item meet wat het behoort te meten. Wanneer een biasbron echter niet tot de te meten vaardigheid behoort, dan doet het item afbreuk aan de constructvaliditeit van de toets. Zo moeten bronnen van itembias op het gebied van de woordenschat geen rol spelen bij rekenitems, maar mogen ze wel voorkomen in taalitems inzoverre met die items beoogd wordt een bijdrage te leveren aan het meten van woordkennis. Itembias als zodanig beperkt de constructvaliditeit van een toets dus niet in alle gevallen. Het aantal tot nu toe bekende bronnen van itembias dat met een redelijk grote mate van zekerheid bias veroorzaakt, is in feite gering. Met inachtneming van het hierboven gestelde lijkt het zinvol dat toetsconstructeurs, leerkrachten en anderen die bij de ontwikkeling van allerlei tests, toetsen en andere evaluatie-instrumenten betrokken zijn, in ieder geval rekening houden met de vier behandelde clusters van biasbronnen voor allochtone leerlingen. Verdere uitwerking van dit punt en een bespreking van een aantal relevante aandachtspunten voor toekomstig wetenschappelijk en toepassingsgericht itembiasonderzoek komen aan de orde in Uiterwijk (1994).

Bibliografie

CALO (Commissie Allochtone Leerlingen in het Onderwijs), *Ceders in de tuin. Naar een nieuwe opzet van het onderwijsbeleid voor allochtone leerlingen.* Zoetermeer: Ministerie van Onderwijs en Wetenschappen, 1992.

Coenen, M. & T. Vallen, Itembias in de Eindtoets Basisonderwijs. In: *Pedagogische Studiën* 68(1), 15-26, 1991.

Extra, G. & T. Vallen, Languages and ethnic minorities in the Netherlands: Current issues and research areas. In: G. Extra & T. Vallen (eds.), *Ethnic minorities and Dutch as a second language* (1985). Dordrecht: Foris Publications, 1-13.

Extra, G. & L. Verhoeven, Community languages in cross-cultural perspective. In: G. Extra & L. Verhoeven (eds.), *Community languages in the Netherlands* (1993a). Amsterdam etc.: Swets & Zeitlinger, 1-28.

Extra, G. & L. Verhoeven, A bilingual perspective on Turkish and Moroccan children and adults in the Netherlands. In: G. Extra & L. Verhoeven (eds.), *Immigrant languages in Europe* (1993b). Clevedon etc.: Multilingual Matters, 68-100.

Hacquebord, H., *Tekstbegrip van Turkse en Nederlandse leerlingen in het voortgezet onderwijs*. Dordrecht: Foris Publications, 1989.

Intraprasert, D., *An investigation of the reliability of five methods for detecting test item bias: an empirical study*. Denton: North Texas State University, 1986.

Jong, M. de & T. Vallen, Linguïstische en culturele bronnen van itembias in de Eindtoets Basisonderwijs voor leerlingen uit etnische minderheidsgroepen. In: *Pedagogische Studiën* 66(2)(1989), 390-402.

Kok, F., *Vraagpartijdigheid*. Amsterdam: Universiteit van Amsterdam, 1988.

Scheunemann, J., *Exploration of causes of bias in test items*. Princeton: Educational Testing Service, 1985.

Schmitt, e.a., *Evaluating hypotheses about differential item functioning*. Princeton: Educational Testing Service, 1992.

Tesser, P., *Rapportage minderheden 1993*. Rijswijk: Sociaal en Cultureel Planbureau, 1993.

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

Uiterwijk, H. & T. Vallen, De bruikbaarheid van de Cito-Eindtoets Basisonderwijs voor leerlingen uit etnische minderheidsgroepen; een eerste analyse. In: R. van Hout & E. Huls (red.), *Artikelen van de Eerste Sociolinguïstische Conferentie* (1991), Delft: Eburon, 395-410.

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

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

Zieky, M., Practical questions in the use of DIF statistics in test development. In: P. Holland & H. Wainer (eds.), *Differential item functioning*. Hillsdale: Lawrence Erlbaum Associates (1993), 82-104.

Waal-Heijkants, M. van de, *Expert-oordelen over potentiële bronnen van itembias in de Eindtoets Basisonderwijs*. Tilburg: KUB (doctoraalscriptie Faculteit Letteren), 1992.

(manuscript binnengekomen 2 september 1994)
(manuscript aanvaard 8 september 1994)