Itembias in de eindtoets basisonderwijs

M. COENEN en T. VALLEN*
Werkverband Taal & Minderheden, Faculteit der Letteren, Katholieke Universiteit Brabant

Samenvatting

In aansluiting bij de door De Jong en Vallen (1989) gerapporteerde inzichten over mogelijke linguïstische en culturele bronnen van itembias voor allochtone leerlingen in de Eindtoets Basisonderwijs, wordt in deze bijdrage voornamelijk verslag gedaan van een vervolgexperiment. Daarin is op bescheiden schaal nagegaan of – bij gelijkblijvende moeilijkheidsgraad – talige manipulatie van een aantal in de Eindtoets 1987 voor allochtonen in statistisch opzicht sterk gebiaste items andere resultaten oplevert dan de resultaten op die items in hun oorspronkelijke vorm. Voor deelname aan het experiment werden, in overleg met de leerkrachten van de 44 deelnemers, paren 'vergelijkbare' autochtone en allochtone (voornamelijk Turkse en Marokkaanse) leerlingen uit groep 8 van de basisschool geselecteerd. Beide leerlingen van de op deze wijze geselecteerde paren moesten ofwel de oorspronkelijke ofwel de gemanipuleerde versie van de items maken en zo uitgebreid mogelijk mondeling toelichten hoe ze de opgaven opgelost hadden. De gemanipuleerde items werden door de allochtonen beter gemaakt dan de niet-gemanipuleerde, terwijl er voor de autochtonen nauwelijks verschil tussen beide versies bestond. Ondanks een aantal interpretatieproblemen kwamen uit het experiment vooral woordgebruik, impliciete zins- en tekstverbanden en itemcomplexiteit als belangrijkste biasbronnen voor allochtonen naar voren.

* Met dank aan H. Uiterwijk en A. Kerkhoff voor hun kritisch commentaar op eerdere versies van dit artikel.

1 *Inleiding*

Door De Jong en Vallen (1989) is op basis van literatuurstudie uiteengezet welke linguïstische en culturele factoren mogelijk van invloed zijn bij het behalen van lagere scores op de Eindtoets Basisonderwijs door allochtone subgroepen die qua kennis- en vaardigheidsniveau vergelijkbaar zijn met autochtone subgroepen. Daarbij kwam onder andere naar voren dat niet alleen een relatief omvangrijke woordkennis van groot belang is voor het succesvol kunnen maken van de Eindtoets, maar dat ook een goede kennis van vaak verborgen culturele aspecten vereist is om cultuurgeladen tekstelementen te kunnen begrijpen en taken goed te kunnen interpreteren. Ook is in bedoeld artikel aangegeven hoe in het onderzoekproject *Item- en testbias voor etnische groepen in de Eindtoets Basisonderwijs* (uitgevoerd door het CITO en de KUB) geprobeerd wordt deze onbedoelde benadelingen (biases) voor leerlingen uit etnische minderheidsgroepen op het spoor te komen. De uit De Jong en Vallen (1989) voortkomende inzichten zijn richtinggevend geweest voor het formuleren van verwachtingen over de mogelijke oorzaken van linguïstische en culturele biases voor allochtone leerlingen in de Eindtoets. Vervolgens zijn deze verwachtingen vergeleken met statistische biasanalyses van de Eindtoets 1987. Uitgebreide informatie over de uitkomsten van deze statistische analyses wordt gegeven in Uiterwijk (1990). Op basis van deze analyses worden de inhoudelijke biasverwachtingen bijgesteld en daarna getoetst aan de resultaten van de Eindtoets 1989. Laatstgenoemde activiteiten zijn momenteel nog in volle gang. Indien er uiteindelijk duidelijke linguïstische en/of culturele biases voor allochtone subgroepen in de Eindtoets worden gesignaleerd en inhoudelijk adequaat kunnen worden onderbouwd, ligt het in de bedoeling op termijn voorstellen voor bijstellingen van toetsonderdelen te formuleren teneinde de kwaliteit van de toets voor autochtone en allochtone groepen identiek te maken.

De hoofdmoot van dit artikel (paragraaf 3 t/m 6) wordt gevormd door het verslag van een in het kader van voornoemd project op

beperkte schaal uitgevoerd hardop-denken-experiment. Daarin is nagegaan of manipulatie van een aantal in de Eindtoets 1987 gebiaste items andere resultaten oplevert in vergelijking met de resultaten op die items in hun oorspronkelijke vorm. De concrete itemmanipulaties hadden uiteraard betrekking op dié items waarvan op grond van de inhoudelijke analyses werd verwacht dat deze niet bedoelde problemen voor leerlingen uit etnische (sub)groepen zouden kunnen opleveren.

2 *Itembias*

Bij het zoeken naar verklaringen voor de soms aanzienlijke verschillen in Eindtoetsscores tussen autochtone en allochtone basisschoolverlaters moet grosso modo met twee mogelijkheden rekening worden gehouden. Enerzijds kan het zijn dat de scoreverschillen veroorzaakt worden door verschillen in juist dié vaardigheden die de Eindtoets wil meten. In dat geval is er in termen van bias niets aan de hand: de allochtone kinderen beheersen de beoogde vaardigheden minder goed en scoren derhalve lager. Anderzijds kan het zijn dat de scoreverschillen tussen autochtonen en allochtonen veroorzaakt worden door verschillen die de toets niet beoogt te meten, maar die ongewild toch gemeten worden. Zoals reeds aangegeven in De Jong en Vallen (1989, p. 391) is er in dergelijke gevallen sprake van itembias, omdat immers de kans op een goed antwoord voor leerlingen met gelijke kennis en vaardigheden maar uit onderscheiden subgroepen verschillend is (zie ook Scheunemann 1988).

Het lijkt vooralsnog het meest raadzaam om bij onderzoek naar itembias in twee stappen te werk te gaan. Wanneer het mogelijk is leerlingen uit onderscheiden subgroepen te matchen naar gelijke kennis en vaardigheden, zou langs statistische weg itembias gedetecteerd kunnen worden. Om uiteindelijk wegen te vinden voor concrete veranderingen of verbeteringen in de Eindtoets is het echter tevens noodzakelijk om alle items afzonderlijk inhoudelijk op potentiële talige en culturele biasbronnen te bestuderen. Dat kan onder meer gebeuren door per item na te gaan of de context waarbinnen de operationalisatie van de itemdoelstelling tot stand is gekomen (in het vervolg *itemcontext*) reëel is in vergelijking met de beschrijving van de doelstelling in het Doelenboek (CITO 1986). Bij een niet-reële context kan bijvoorbeeld gedacht worden aan onbedoelde meting van Nederlandse taalvaardigheid bij rekenopgaven, aan culturele voorkennis of cultureel bepaalde toetservaring bij informatieverwerkingopdrachten etcetera. Beide manieren om itembias op het spoor te komen zijn in het project gehanteerd en worden – deels mede aan de hand van Uiterwijk (1990) – in het onderstaande kort aan de orde gesteld.

2.1 *Statistisch vastgestelde itembias in de Eindtoets 1987*

Het matchen van leerlingen naar gelijke kennis en vaardigheden, wat bij onderzoek naar itembias noodzakelijk is, vormt een fundamenteel probleem. Een toets als de Eindtoets is zowel als geheel als ook in de afzonderlijke onderdelen (Taal, Rekenen en Informatieverwerking) zeer waarschijnlijk multidimensioneel van aard. Van alle items van bijvoorbeeld het onderdeel Rekenen wordt weliswaar verondersteld dat deze de te meten vaardigheden representeren, maar het is zeker niet zo dat er voor de beantwoording van al deze 60 items slechts één welbepaald type rekenvaardigheid relevant is. Mutatis mutandis geldt dat natuurlijk ook voor de onderdelen Taal en Informatieverwerking. Meerdimensionaliteit van een toets is een probleem bij het matchen van leerlingen naar gelijke kennis en vaardigheden, omdat het uitermate lastig is vast te stellen welke vaardigheden de toets precies meet en op grond van welke criteria de leerlingen derhalve gematcht moeten worden. Omdat de itemresponsen niet verklaard kunnen worden op basis van één onderliggende vaardigheid (de latente trek), ligt het niet voor de hand te veronderstellen dat de Eindtoetsdata voldoen aan een item-response-model (Tobi 1989).

Een tweede probleem ligt in het vinden van een extern criterium voor de beoogde matching, dat niet of in beduidend mindere mate van bias wordt verdacht dan het object van onderzoek. Met andere woorden: welke taal- of rekentoets pretendeert hetzelfde te meten als de onderdelen Taal en Rekenen van de Eindtoets en wordt bovendien niet of beduidend minder van bias verdacht? Een dergelijk criterium is niet beschikbaar. Wanneer nu bijvoorbeeld de totaalscore op een of meerdere Eindtoetsonderdelen als maat wordt gebruikt om de leerlingen te matchen, is het

natuurlijk nog altijd mogelijk dat alle items of een gedeelte ervan bias bevatten. Een oplossing kan zijn om het criterium totaalscore te zuiveren door met behulp van een iteratieve procedure de gebiaste items voor de berekening van de totaalscore te verwijderen (Holland & Thayer, 1986; Verhelst, 1988; Kok, 1988). Bij het verwijderen van dergelijke items moet uiteraard wel gecontroleerd worden of de resterende opgaven de te meten vaardigheid nog voldoende representeren (het dekkingsprobleem; zie Uiterwijk, 1990).

In eerste instantie is in het onderhavige onderzoek gestart met de totaalscoreprocedure van Mantel-Haenszel. Via deze veelbelovende techniek (Kok, 1988; Holland & Thayer, 1986; Uiterwijk, 1990) zou het mogelijk zijn, met inachtneming van de bovengeschetste eisen, een van bias gezuiverde totaalscore te genereren. Deze score kan dan als alternatieve meting worden gebruikt om groepen te matchen naar gelijke kennis en vaardigheden.

Uit de door Uiterwijk (1990) op de Eindtoetsscores 1987 uitgevoerde Mantel-Haenszel-analyses komen nogal wat items naar voren die een negatieve bias (significantieniveau 1%; $z > +2.58$) bevatten voor de onderzochte Turkse (n = 797), Marokkaanse (n = 720), Surinaams-Creoolse (n = 391) en Surinaams-Hindoestaanse (n = 334) deelnemers. Het aantal Eindtoetsdeelnemers uit de overige etnische minderheidsgroepen was te gering ($n < 300$) om de beoogde analyses te kunnen uitvoeren.

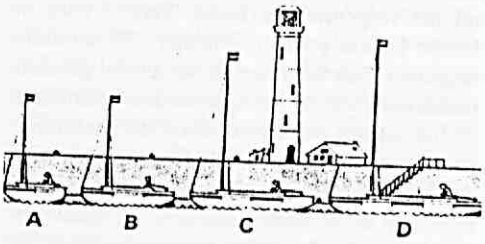
Opvallend is met name het verschil in aantallen gebiaste items tussen de toetsonderdelen (Taal, Rekenen en Informatieverwerking, elk bestaande uit 60 items) en tussen de onderzochte etnische minderheidsgroepen. Bij Rekenen zijn relatief nog de minste items negatief gebiast: voor de Turken en Creolen gaat het daarbij om respectievelijk 14 en 13 items, voor Marokkanen en Hindoestanen om respectievelijk 8 en 6. Bij het onderdeel Informatieverwerking zijn de aantallen (negatief) gebiaste items 17 (Turken), 18 (Marokkanen), 10 (Creolen) en 4 (Hindoestanen). Met uitzondering van de Creolen (voor wie Rekenen de meeste bias laat zien) bevat het onderdeel Taal de meeste bias voor de etnische minderheidsgroepen: voor Turken, Marokkanen, Creolen en Hindoestanen gaat het daarbij om respectievelijk 22, 25, 9 en 15 (van de 60) items. Niet alleen tussen, maar ook binnen de onderscheiden subgroepen zijn er verschillen, zoals blijkt uit het volgende voorbeeld. Terwijl voor de totale Turkse groep deelnemers 22 taalitems negatieve bias vertonen, is het aantal gebiaste taalitems voor de 'hoogscoorders' (minstens 29 items goed beantwoord) op het taalonderdeel binnen deze minderheidsgroep 19, voor de 'laagscoorders' 8. Soortgelijke diversiteit treedt op bij de andere etnische minderheidsgroepen en bij de andere toetsonderdelen. De bovenstaande aantallen negatief gebiaste items moeten nadrukkelijk met de nodige voorzichtigheid in ogenschouw worden genomen, niet alleen omdat de aantallen verschillen per subgroep maar ook omdat ze variëren afhankelijk van het aantal items dat binnen de statistische procedure wordt gebruikt voor het berekenen van de gezuiverde totaalscore. In verband met het dekkingsprobleem is in het onderhavige onderzoek voor de berekening van de totaalscore ernaar gestreefd steeds ten minste tweederde deel van de oorspronkelijke items te handhaven.

Hoewel de name van bias (z-waarde) verschilt per etnische groep, is de rangorde van met name de items met de meeste bias opvallend constant. Dat geldt vooral voor het onderdeel Taal: item 11 vertoont bij alle groepen de meeste bias, item 31 staat steeds op plaats twee of drie en ook 14 en 26 staan bij alle vier de etnische minderheidsgroepen bij de vijf meest gebiaste items. Bij Rekenen en Informatieverwerking is de volgorde minder constant, maar ook bij deze onderdelen zijn in grote lijnen dezelfde items in de top vijf vertegenwoordigd (Uiterwijk 1990).
Ter illustratie wordt hieronder het bij ieder onderdeel meest gebiaste item weergegeven.

Taal-item 11:

Wat kun je het beste doen met: *Gelukkig was het goed met de wind...* (r.4, 5)?
A. Zo laten staan
B. Vervangen door: Gelukkig hadden we de wind mee, ...
C. Vervangen door: Gelukkig hadden we goede wind, ...
D. Vervangen door: Gelukkig was de wind met ons, ...

Reken-item 29:



Welk bootje heeft in verhouding tot zijn lengte de langste mast?

Info-item 31:

Anke heeft één van de onderstaande mensen opgebeld. Toen ze het telefoonnummer opzocht, heeft ze het briefje laten zitten. Wie heeft ze gebeld?

A. W. van de Berg, Acaciaplein 6, Burger
B. L. van Dalen, Weegbree 27, Burger
C. A. Warringa, Boslaan 1, Dalen
D. K. Dootjes, Harkstee 7, Wezup

Er blijken ook enkele items statistisch gebiast te zijn ten voordele van minderheidsgroepen ($z < -2.58$), en derhalve ten nadele van qua kennis en vaardigheden vergelijkbare autochtone subgroepen ($n = 4969$). Bij Rekenen en Informatieverwerking gaat het in dat opzicht om kleine aantallen items; bij het onderdeel Taal zijn het er wat meer. Daarbij gaat het met name om spellingitems (zie 2.2).

2.2 *Inhoudelijke biasanalyse van de Eindtoets 1987*

Zoals reeds onder paragraaf 1 is aangegeven, is door De Jong en Vallen (1989) op basis van literatuuronderzoek een aantal veronderstellingen geformuleerd met betrekking tot potentiële linguïstische en culturele bronnen van itembias. De veronderstelde linguïstische bronnen zijn grofweg onder te brengen in factoren op woord-, zins- en tekstniveau. In het eerste geval wordt verondersteld dat ambigue of infrequente woorden in een leestekst, een instructie of in afleiders itembias kunnen veroorzaken, indien ze tenminste niet tot de vaardigheid behoren die gemeten dient te worden. Op zinsniveau wordt verondersteld dat ambigue of impliciete zinsconstructies een bron van itembias kunnen zijn. Op tekstniveau ten slotte wordt verondersteld dat moeilijke verwijzingen (bijvoorbeeld over langere tekstpassages heen) of ambigue verwijzingen (bijvoorbeeld het verwijswoord 'ze' kan een enkelvoudig of een meervoudig antecedent hebben) tot bias kunnen leiden, opnieuw indien meting ervan niet beoogd wordt.

Zowel vanwege de centrale rol die woordenschat ten aanzien van itembias lijkt te spelen als vanwege het gegeven dat, gezien de wetenschappelijke stand van zaken, biasonderzoek op zins- en tekstniveau vooralsnog wellicht weinig kans op succes heeft, werd het zinvol geacht in het onderhavige onderzoek vooral (maar niet exclusief) een accent op vocabulaire te leggen. Nadere inhoudelijke inspectie van de statistisch gebiaste items lijkt deze beslissing te ondersteunen. De 5 'zwaarst' gebiaste taalitems voor Turken in 1987 bijvoorbeeld hebben allemaal betrekking op vocabulaire. In totaal hebben van de 22 voor Turken gebiaste taalitems er tien tot doel woordkennis te meten. Van de in totaal 8 items die woordkeus (vorm en inhoud) willen toetsen zijn er 6 gebiast voor Turken, en verder zijn alle 4 de items die kennis van de betekenis van woorden en uitdrukkingen willen meten gebiast.

Het merendeel van de taalitems (10 van de 12) die een bias ten gunste van de Turkse leerlingen vertonen heeft betrekking op spelling. Dat betreft de helft van alle in de toets aanwezige spellingitems. Spelling is kennelijk een specifieke vaardigheid die als geheel in vergelijking met andere taalvaardigheden blijkbaar voor Turkse leerlingen relatief gemakkelijker is dan voor (qua kennis en vaardigheden vergelijkbare) autochtone leerlingen. Echter ook bij spelling lijkt het blijkens nadere analyses niet om een unidimensionele vaardigheid te gaan. Wanneer namelijk binnen de rubriek spelling wordt gekeken naar leerlingen met gelijke kennis en vaardigheden, blijkt de (sterk regelgebonden) spelling van werkwoorden een andere vaardigheid dan de (aanzienlijk minder regelgebonden) spelling van niet-werkwoorden. Eerstgenoemde blijkt

voor allochtonen (i.c. Turken) een positieve bias in vergelijking tot de autochtonen op te leveren, terwijl het beeld bij de spelling van niet-werkwoorden tegenovergesteld is (positieve bias voor autochtonen).
Een mogelijke verklaring voor de relatief goede beheersing van de spellingregels door allochtonen is dat deze regels pas op school worden geleerd. Voor een aantal andere vaardigheden geldt dat niet: het leren van woorden bijvoorbeeld begint lang voordat taalleerders naar school gaan. Tweede-taalleerders die thuis nauwelijk met Nederlands in aanraking komen, lopen het gevaar wat woordkennis betreft met een achterstand ten opzichte van eerste-taallleerders te beginnen. Spelling echter is voor zowel T1- als T2-leerders een nieuwe, op school te leren vaardigheid. Bovendien zijn spellingregels relatief gemakkelijk te verwerven, zijn ze eindig in aantal en krijgen veel aandacht op school. Ook kan het zijn dat T2-leerders meer afhankelijk zijn van en daardoor meer gespitst zijn op spellingregels, omdat ze minder dan autochtonen (kunnen) vertrouwen op fonetische woordkenmerken. Tot slot kan nog worden opgemerkt dat de context die bij spellingopgaven in de Eindtoets gehanteerd wordt altijd uit één losse zin bestaat, zodat er relatief weinig kans is op een niet-communale context die talige of culturele bias kan veroorzaken.

Bij culturele biases kan een onderscheid worden gemaakt tussen biases ten gevolge van de culturele lading van teksten/tekstfragmenten en biases ten gevolge van onbekendheid van etnische minderheidsgroepen met (bepaalde typen) toetsen, specifieke taken en/of specifieke oplossingstrategieën. Om verschillende redenen is het op dit moment uitermate problematisch inhoudelijk goed onderbouwde uitspraken te doen over de culturele lading van teksten en toetsitems en over cultureel bepaalde toetservaring. In De Jong en Vallen (1989, pp. 396-400) is daarop uitvoerig ingegaan. Ook nadere inspectie van de concrete Eindtoetsitems leverde in dit verband nauwelijks iets op. Enerzijds bleken geen plausibele, cultureel bepaalde oorzaken voor itembias aanwezig bij items waar in statistische zin itembias was vastgesteld, anderzijds bleken items waarbij om inhoudelijke redenen cultureel bepaalde verklaringen voor itembias plausibel konden worden gegeven, statistisch geen bias te vertonen. Zelfs duidelijk cultuurbepaalde items (bijvoorbeeld over wintersport) leverden in dit opzicht niets op. Waarschijnlijk zijn voor de hand liggende cultuurverschillen ook voor de kinderen zeer inzichtelijk en zullen subtiele verschillen zich waarschijnlijk zowel aan de waarneming van de toetsdeelnemers als aan die van de onderzoekers onttrekken. Verder moet worden geconstateerd dat bij een professioneel vervaardigd instrument als de Eindtoets er waarschijnlijk weinig kans op cultuurbepaalde 'uitglijders' bestaat.

Zolang niet volledig duidelijk is welke vaardigheden er in de verschillende Eindtoetsonderdelen precies gemeten (moeten) worden, blijft de relatie tussen itemmoeilijkheid en itembias vooral bij taalitems zodanig complex dat besloten is in eerste instantie voornamelijk op zoek te gaan naar oorzaken van talige biases bij items die geen taalvaardigheid pretenderen te meten (Rekenen en Informatieverwerking). Ook leek het van belang de invloed van onduidelijkheden in afbeeldingen bij items te onderzoeken. Wat het onderdeel Taal betreft leek het vanwege de centrale rol die het vocabulaire vervult toch zinvol enige aandacht aan woordenschatitems te besteden. Een eerste stap in deze zoektocht wordt gevormd door een hardop-denken-experiment op bescheiden schaal waarin aan de hand van een aantal statistisch sterk gebiaste items bij leerlingen onderzocht wordt of de door de onderzoekers veronderstelde biasoorzaken realiteitswaarde hebben.

3 *Het hardop-denken-experiment*

3.1 *Doelstelling*

Het experiment heeft tot doel na te gaan of en hoe vaak bij een aantal oorspronkelijke (en sterk gebiaste) items van de Eindtoets 1987 een fout antwoord wordt gegeven ten gevolge van een inhoudelijk voorspelde biasbron. In zoverre dat mogelijk is wordt ook onderzocht of en hoe vaak bij gemanipuleerde items ten gevolge van de manipulatie (= verwijdering veronderstelde biasbron) een goed antwoord wordt gegeven. Met gebruikmaking van een hardop-denk-procedure wordt geprobeerd in de redeneringen van de kinderen aanwijzingen te vinden die de geformuleerde verwachtingen omtrent de oorzaken van itembias bevestigen dan wel ontkennen. Wellicht treden er ook

nieuwe, niet voorspelde 'biasverdachte' aspecten aan het daglicht.

Bij de itemmanipulatie is geprobeerd de items zodanig te bewerken dat veronderstelde talige en visuele biases verwijderd zijn. Verwacht wordt dat vervolgens de gemanipuleerde items bij allochtonen en autochtonen meer dan de oorspronkelijke items een zuiverder beroep doen op beoogde vaardigheden. Als een dergelijke manipulatie juist gebeurd is, zou voorspeld kunnen worden dat bij de gemanipuleerde items minder verschillen in foutscores tussen allochtonen en autochtonen optreden dan bij de oorspronkelijke items.

3.2 *Biasbronnen*

In het kader van het experiment is een poging ondernomen om mogelijke biases uit items voor Rekenen en Informatieverwerking (exclusief de rubriek Lezen van teksten) van de Eindtoets 1987 te verwijderen. Zoals opgemerkt in 2.2 zou bij deze onderdelen wellicht iets gemakkelijker dan bij het onderdeel Taal een onderscheid gemaakt kunnen worden tussen terecht en onterecht gemeten vaardigheden. Bij opgaven die rekenvaardigheden en vaardigheden in het hanteren van informatiebronnen, kaarten, tabellen en grafieken willen meten, staan talige vaardigheden immers niet expliciet in de doelstelling. Er mag derhalve verondersteld worden dat het CITO ervan uitgaat dat de talige vaardigheden die voor het maken van deze opgaven vereist zijn, communaal zijn en niet discrimeren tussen de diverse (sub)groepen. In ieder geval zouden ze niet van invloed mogen zijn op de hoogte van de score. Verondersteld zou nu mogen worden dat de talige biases in deze opgaven in veel gevallen bestaan uit onnodig moeilijke woorden en uit complexe grammaticale en/of ambigue constructies en/of uit impliciete zins- en tekststructuren. Verder blijken veel opgaven die sterk statistisch gebiast zijn onduidelijkheden te bevatten in tekeningen, kaartjes, tabellen of grafieken, die soms cruciaal zijn voor het vinden van het goede antwoord. Tenslotte zijn de meeste sterk gebiaste reken- en informatieverwerkingsopgaven nogal complex van aard in die zin, dat er achtereenvolgens en in de goede volgorde meerdere stappen gemaakt moeten worden om tot het goede antwoord te kunnen komen. Complexiteit lijkt weliswaar een belangrijke factor, maar het is geen noodzakelijke noch voldoende 'voorwaarden' voor bias. Enerzijds zijn er opgaven die niet complex maar toch sterk gebiast zijn (zie bijvoorbeeld voorbeeld 1 in paragraaf 4), anderzijds zijn er ook zeer complexe opgaven die niet gebiast zijn. Blijkbaar wordt bias vaak veroorzaakt door een combinatie van complexiteit en talige en/of visuele onduidelijkheden. Een grote talige moeilijkheid alleen kan evenwel ook al tot bias leiden. Uit het Eindtoetsonderdeel Taal zijn, zoals al vermeld, alleen opgaven opgenomen die volgens hun doelstelling woordenschat moeten meten.

3.3 *'Fouten' in contexten*

In sommige tekstpassages van het taalonderdeel zijn voor het meten van correct taalgebruik opzettelijk allerlei soorten 'fouten' aangebracht, zoals inconsistenties, weggelaten of overbodige informatie en ongebruikelijke formuleringen. Daarover worden dan vervolgens vragen gesteld. Hoewel de leerlingen hierop aan het begin van de taak wel kort worden geattendeerd en er op school in veel gevallen enige voorbereiding op de toets heeft plaatsgevonden, kan dit niettemin zeer verwarrend zijn, zeker wanneer er onvoldoende ervaring met dit soort taken is opgedaan. Verder kunnen deze opzettelijk aangebrachte 'fouten' zeer belemmerend werken bij het opbouwen van een samenhangende tekstrepresentatie in het geheugen tijdens het lezen. Dit zou met name nadelig kunnen zijn voor tweede-taalleerders, die vanwege hun vaak kleinere woordenschat toch al meer moeite kunnen hebben met het begrijpen van een tekst.

Voor deze veronderstelling zijn enkele aanwijzingen te vinden in tekstverwerkingsonderzoek. Uit een onderzoek van Boonman en Kok (1986, p. 270 e.v.) blijkt dat de cognitieve verwerkingsactiviteiten van leerlingen uit groep 8 primair gericht waren op het zo letterlijk mogelijk opnemen van informatie zonder veel bewerkingen uit te voeren. Inconsistenties bijvoorbeeld werden, met name als ze niet erg expliciet waren, nauwelijks opgemerkt. Een kwart van de leerlingen die de inconsistenties tijdens het lezen niet opmerkten, ontdekte deze ook niet toen ze er in een gesprek achteraf expliciet mee geconfronteerd werden. Stanovich en West (1981; besproken in Boonman en Kok, 1986, p. 53 e.v.) hebben de invloed van de context op tekstverwerking onderzocht en constateerden dat de context een relatief grotere invloed had bij lezers met een kleine woordenschat dan bij lezers met een grote woordenschat. De laatsten hadden die

context vaak niet meer nodig, omdat de woorden ook zonder context snel herkend werden. De groep lezers met weinig woordkennis, waartoe veel T2-leerders behoren, had echter relatief veel tijd en aandacht nodig voor woordherkenning. Een passende context kan dit proces versnellen. Een niet passende context heeft bij lezers met een grote woordenschat weinig invloed, omdat de woordherkenning vrijwel geautomatiseerd is; bij lezers met een kleine woordenschat echter zal dit de woordherkenning, en dus ook het leesproces, nog verder vertragen.

Deze grotere invloed van de context bij lezers met weinig woordkennis wordt ook gevonden door Hacquebord (1989). Zij laat zien dat Turkse MAVO-leerlingen lager scoorden dan Nederlandse LBO-leerlingen op woordherkennings- en grammaticataken, terwijl ze wel telkens hoger scoorden op tekstbegriptoetsen. Blijkbaar gebruiken de Turkse leerlingen compenserende globale leesstrategieën om hun gebrek aan lexicale en grammaticale kennis op te vangen. Als echter de itemcontext, die vooral T2-leerders tot steun kan zijn, 'fouten' bevat, worden deze leerlingen ten onrechte op het verkeerde been gezet. Waarschijnlijk geldt een analoge redenering voor relatief impliciete contexten. Explicitering kan dan helpen om betere tekstrepresentaties op te bouwen.

3.4 *Geselecteerde opgaven*

Er zijn uit genoemde Eindtoetsonderdelen van 1987 alleen opgaven geselecteerd die een grote statistische bias vertonen voor allochtone leerlingen. Hierbij zijn twee selectiecriteria gebruikt. Enerzijds moest een opgave voor minstens twee, maar liefst voor drie etnische minderheidsgroepen (Turken, Marokkanen, Surinamers) gebiast zijn, anderzijds moest het mogelijk zijn een (minimale) manipulatie toe te passen waarvan enig effect verwacht kon worden. In totaal zijn uiteindelijk 17 opgaven gekozen die middels versie A in hun oorspronkelijke vorm en middels versie B in een gemanipuleerde vorm zijn getoetst: acht rekenopgaven, vier informatieverwerkingsopgaven en vijf taalopgaven. In alle gevallen is er in de oorspronkelijke versie sprake van een of meerdere globale biasbronnen: talige moeilijkheid, visuele moeilijkheid, complexiteit, ambiguïteit, 'fout' in de context.

3.5 *Informanten*

De 44 deelnemende leerlingen waren afkomstig uit groep 8 van vijf verschillende basisscholen. Het design komt naar voren uit Tabel 1.

Tabel 1 *Verdeling van de informanten over de beide toetsversies (A, B)*

	allochtonen	autochtonen	totaal
A: ongemanipuleerde versie	11	11	22
B: gemanipuleerde versie	11	11	22
	22	22	44

Om het effect van de manipulaties goed te kunnen nagaan zouden in principe beide toetsversies aan dezelfde groep leerlingen voorgelegd moeten worden, uiteraard met een interval van enkele maanden. Het nadeel van een dergelijke opzet is echter dat niet duidelijk is hoe groot de rol van de geheugenfactor is; er bestaat een grote kans dat leerlingen zich bepaalde opgaven herinneren. Vanwege dit risico werd gekozen voor afname bij twee verschillende, maar zoveel mogelijk vergelijkbare groepen leerlingen. De A- en B-versie zijn dus niet door dezelfde leerlingen gemaakt. Bij de interpretatie van de resultaten moet er derhalve rekening mee worden gehouden dat de vier cellen niet volledig vergelijkbaar zijn.

In overleg met de leerkrachten van de leerlingen werden steeds paren allochtone en autochtone leerlingen geselecteerd. Alle tweede-taalleerders moesten aan twee criteria voldoen. Ze moesten thuis (ook) hun eerste taal spreken en ze moesten minimaal vanaf het begin van groep 3 in het Nederlandse (basis)onderwijssysteem participeren. Dit was enerzijds nodig om te voorkomen dat leerlingen nog volop in hun tweede-taalleerproces zouden zitten en daardoor veel opgaven niet zouden begrijpen en anderzijds om te voorkomen dat de toch al heterogene groep allochtonen nog heterogener zou worden door grote verschillen in taalvaardigheid Nederlands. Bij iedere gekozen allochtone leerling zocht de leerkracht een autochtone leerling die naar zijn oordeel zoveel mogelijk vergelijkbaar was op een aantal factoren die van belang zijn voor schoolsucces, zoals sociaal-economische achtergrond, taalvaardigheid Nederlands, rekenvaardigheid, motivatie, en meer objectieve

factoren als doubleergeschiedenis, Eindtoetsscore en advies voor het voortgezet onderwijs. Uiteraard kregen beide leerlingen van de 22 op deze wijze geselecteerde paren ofwel de oorspronkelijke ofwel de gemanipuleerde toetsversie.

Om het effect van de manipulaties te kunnen onderzoeken zijn ook de groep die de oorspronkelijke versie en de groep die de gemanipuleerde versie ging maken zoveel mogelijk vergelijkbaar gehouden qua prestatieniveau. In beide groepen zitten ongeveer evenveel leerlingen met een LBO-advies, hetgeen ook geldt voor MAVO- en HAVO-advies. Zowel bij de allochtonen als autochtonen zijn er precies evenveel meisjes als jongens. De groep allochtonen is samengesteld uit voornamelijk Turken (11) en Marokkanen (8), verder een Chinese, een Antilliaanse en een Braziliaanse leerling. De Turkse en Marokkaanse leerlingen waren nagenoeg gelijk verdeeld over de twee toetsversies. De vaders van de meeste leerlingen vallen in de beroepsklassen van geschoolde of ongeschoolde arbeiders.

3.6 *Procedure*

De kinderen kregen de opdracht iedere opgave eerst te bestuderen, vervolgens het naar hun oordeel goede antwoord aan te kruisen en daarna zo uitgebreid en precies mogelijk te vertellen hoe ze de opgave opgelost hadden. Hierbij werd geen tijdslimiet gesteld. De gesprekken werden op audioband opgenomen. Als een kind een fout maakte of een verkeerde of onduidelijke toelichting gaf, werd eerst verder gevraagd om er achter te komen of het de bedoelde vaardigheden beheerste. Vervolgens werd nagegaan of de veronderstelde biasbronnen inderdaad problematisch waren geweest.

4 *Resultaten*

4.1 *Gemiddelden*

De allochtone leerlingen die de ongemanipuleerde versie (versie A) hebben gemaakt, hebben gemiddeld 7,9 van de 17 opgaven goed. De autochtonen zitten daar met 11,7 opgaven goed bijna 4 opgaven boven. Met uitzondering van de leerlingen met een HAVO-advies verschillen de autochtone en allochtone leerlingen significant van elkaar (de leerlingen met MAVO-advies op 5%-niveau, de leerlingen met LBO-advies op 1%-niveau). Uiteraard gaat het hier telkens om nogal kleine groepjes leerlingen waardoor de kans op toevalsfouten groot is. De constatering van Uiterwijk (1990) dat allochtonen ondanks lagere scores bij het advies voor het voortgezet onderwijs vaak het voordeel van de twijfel krijgen, wordt echter zeker niet tegengesproken.

De allochtonen die de gemanipuleerde versie (versie B) hebben gemaakt, maken gemiddeld 2,6 opgaven meer goed dan de allochtonen die de ongemanipuleerde versie hebben gemaakt. Bij de autochtonen is het verschil daarentegen slechts 0,7 opgave. Het scoreverschil tussen de allochtonen en autochtonen bij de gemanipuleerde versie (bijna 2 opgaven) is gehalveerd in vergelijking met het verschil tussen de groepen die de ongemanipuleerde versie maakten (bijna 4 opgaven). Opvallend is dat bij de gemanipuleerde versie de verschillen tussen allochtonen en autochtonen niet alleen bij leerlingen met een HAVO- maar ook bij die met een MAVO-advies verwaarloosbaar klein zijn. Bij de leerlingen met een LBO-advies is het verschil echter tamelijk groot (gemiddeld 4,2 opgaven; significant op 5%-niveau). Toch scoren ook de allochtone LBO-leerlingen die de B-versie maakten 3,8 opgaven meer goed in vergelijking met de allochtonen van de A-versie (bij de autochtonen is het verschil 1,7 opgaven).

Bij het onderdeel Informatieverwerking blijken alle vier de opgaven door de allochtonen in de B-versie beter gemaakt te worden dan in de A-versie. Ook drie van de acht rekenopgaven zijn in de gemanipuleerde versie door veel meer allochtonen goed gemaakt dan in de ongemanipuleerde versie. Bij de rest van de rekenopgaven zijn er nauwelijks verschillen. Bij drie van de vijf taalopgaven wordt eveneens de B-versie beter gemaakt, maar bij de andere twee wordt de B-versie slechter gemaakt.

Opvallend is dat de autochtonen van de B-versie in vergelijking met die van de A-versie alleen op Taal enigszins hoger scoren (gemiddeld 0,45 opgave meer goed = 9 procent van alle taalitems; zie Tabel 2). Ook bij de allochtonen is er bij Taal een scoreverschil van 9 procent, maar de verschillen zijn groter bij Rekenen (gemiddeld 1,36 opgave meer goed = 17 procent van de rekenopgaven) en Informatieverwerking (gemiddeld 0,8 opgave meer goed = 20 procent van de informatieverwer-

kingsitems), ten gunste van de B-versie. Blijkbaar hebben de manipulaties bij Taal voor beide groepen evenveel effect gehad, en die bij Rekenen en Informatieverwerking alleen voor de allochtonen. Dit komt overeen met de verwachting dat taalopgaven veel moeilijker te manipuleren zouden zijn zonder de bedoelde vaardigheden te veranderen. Het lijkt erop dat de gemanipuleerde taalopgaven gewoon gemakkelijker zijn geworden voor iedereen. De manipulaties bij Rekenen en Informatieverwerking lijken beter geslaagd, immers alleen de allochtonen hebben er zichtbaar baat bij gehad. Hoewel de cijfers vanwege de kleine aantallen leerlingen per cel voorzichtig geïnterpreteerd moeten worden, zijn dit toch opvallende verschillen. In de volgende paragraaf zal worden nagegaan of deze resultaten bevestigd worden in de hardop-denk-protocollen.

Tabel 2 *Gemiddelden goede antwoorden (in percentages), uitgesplitst naar toetsversie (A, B), toetsonderdelen en allochtonen/autochtone leerlingen (k = aantal items)*

	allochtonen	autochtonen
Rekenen		
A (k = 8)	36.4 (n = 11)	55.7 (n = 11)
B (k = 8)	53.4 (n = 11)	58.0 (n = 11)
B-A	17.0	2.3
Informatie-verwerking		
A (k = 4)	52.3 (n = 11)	84.1 (n = 11)
B (k = 4)	72.7 (n = 11)	84.1 (n = 11)
B-A	20.4	0
Taal		
A (k = 5)	58.2 (n = 11)	78.2 (n = 11)
B (k = 5)	67.3 (n = 11)	87.3 (n = 11)
B-A	9.1	9.1
totaal		
A (k = 17)	46.5 (n = 11)	69.0 (n = 11)
B (k = 17)	62.0 (n = 11)	72.7 (n = 11)
B-A	15.5	3.7

4.2 *Inhoudelijke protocol-analyses*

Zoals eerder opgemerkt, worden talige of visuele itemkenmerken pas van een biaslabel voorzien als ze tot fouten leiden, terwijl de vaardigheden die ter meting beoogd worden wel beheerst worden. Van zowel Taal, Rekenen als Informatieverwerking wordt in het volgende telkens één opgave uitgebreid gepresenteerd, waarbij zowel de scores als de belangrijkste resultaten uit de protocolanalyses worden vermeld. Van de overige opgaven worden alleen korte samenvattingen van de resultaten gegeven.

Voorbeeld 1: Rekenen

(A-versie)	(B-versie)
"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.?'	idem, met manipulatie van de laatste zin: "Wat moet vader voor de naaimachine betalen met B.T.W.?'
a. *f* 160,–	c. *f* 820,–
b. *f* 640,–	d. *f* 960,–

De ongemanipuleerde versie wordt fout beantwoord door 6 van de 11 allochtonen en 2 van de 11 autochtonen. Bij 2 allochtonen en bij 1 autochtoon is het woord *inclusief* de foutoorzaak, dus biasbron. Een andere autochtoon heeft ook problemen met inclusief, maar nog meer met het berekenen van procenten. Als i.p.v. inclusief de aanduiding *met* zou zijn gebruikt, maakt dat (naar het eigen oordeel van de leerlingen) voor 2 allochtonen niet uit en zou dat voor 4 andere allochtonen en de 2 autochtonen gemakkelijker zijn geweest. Twee andere allochtonen geven een fout antwoord vanwege de impliciete vraag; de ene leerling ziet uit zichzelf dat ze fout zit, de andere pas na explicitering met *voor de naaimachine*.

Bij de gemanipuleerde versie geven 2 allochtonen en 1 autochtoon een fout antwoord.

Bij de overige rekenopgaven levert eenmaal een *tekening* van bootjes die nogal klein en onduidelijk zijn problemen op. Verder zijn vooral talige aspecten problematisch, bijvoorbeeld bij *welk bootje heeft in verhouding tot zijn lengte de langste mast?* (rekenitem 29, zie paragraaf 2.1) wordt *zijn lengte* vaak geïnterpreteerd als de lengte van de mast. De aanduiding *overige onkosten* is niet bij iedereen bekend, en de aanduiding *half procent* wordt begripsmatig verward met *de helft*.

Voorbeeld 2: Informatieverwerking

(Er staan 4 grafieken getekend: land a, b, c en d met steeds een geboorte- en sterftelijn die stijgt of daalt)

(A-versie)	(B-versie)
Van welk land kan men zeggen dat het aantal geboorten toeneemt en het aantal sterfgevallen afneemt?	In welk land is tussen 1960 en 1985 het aantal geboorten gestegen en het aantal sterfgevallen gedaald?

Bij de ongemanipuleerde versie geven 5 allochtonen en 0 autochtonen een fout antwoord. Twee allochtonen hebben moeite met de *legenda* die erg dicht op de grafieken staat; ze merken hem niet meteen op. Met de worden *toenemen* en *afnemen* hebben 4 allochtonen problemen. Een van deze leerlingen, die denkt dat afnemen *iemand zijn spullen afpakken* betekent en toenemen *pakken*, komt via de redenering dat de andere grafieken 'raar' zijn toch op het goede antwoord uit. De gehanteerde betekenis van de werkwoorden toenemen en afnemen is voor 1 allochtoon waarschijnlijk de belangrijkste foutenbron en voor een andere mede oorzaak van de fout. Alle vier de leerlingen die deze aanduidingen problematisch vonden, begrepen de combinatie *stijgen/dalen* wel.

Bij de gemanipuleerde versie geven 2 allochtonen en 1 autochtoon een fout antwoord.

Bij de overige opgaven spelen vooral *visuele onduidelijkheden* een rol. Bij een plattegrond bijvoorbeeld wordt over splitsingen gesproken die niet duidelijk als splitsingen in het kaartje zijn aangegeven. Bij een opgave over een telefoongids wordt verwezen naar een onduidelijk aangegeven briefje en zijn de cruciale plaatsnamen op de voorkant van de gids dermate klein gedrukt dat een aantal leerlingen er overheen leest.

Voorbeeld 3: Taal

(A-versie)	(B-versie)
"...Ze komen aan de voet van een zandberg neer. De bemanning is er zonder kleerscheuren afgelopen. ...' Wat kun je het beste doen met "afgelopen"?	"De autobestuurder was er bij het ongeluk gelukkig zonder kleerscheuren ...' Welk woord past hier het beste?
a. Zo laten staan. *b.* Vervangen door: afgekomen *c.* Vervangen door: afgeraakt *d.* Vervangen door: afgevallen	*a.* afgekomen *b.* afgelopen *c.* afgeraakt *d.* afgevallen

Bij de ongemanipuleerde versie geven 3 allochtonen en 2 autochtonen een fout antwoord. Drie leerlingen kiezen voor a (*afgelopen*, welk alternatief in zekere zin misleidend is omdat het door de vorige zin opgeroepen wordt).

Bij de gemanipuleerde versie geven 1 allochtoon en 2 autochtonen een fout antwoord.

Bij de andere taalopgaven die ook allemaal tot doel hebben woordkennis te meten, wordt bij twee opgaven in de B-versie wellicht geen beroep meer op dezelfde vaardigheid gedaan: in de ene opgave is de infrequente uitdrukking *wonderlijk genoeg* vervangen door *vreemd genoeg*, bij de andere kan de (gevraagde) betekenis van *bijtijds* uit de context afgeleid worden. Vreemd genoeg is bij twee opgaven de B-versie moeilijker dan de A-versie.

5 *Conclusies en discussie*

Interpretatieproblemen

Het besproken experiment heeft een inhoudelijke toetsing van veronderstelde itembiasbronnen tot doel. Er is met behulp van hardop-denk-protocollen nagegaan hoe vaak het voorkwam dat leerlingen wel de bedoelde vaardigheden beheersten maar toch fouten maakten vanwege onbedoelde moeilijkheden. Vervolgens is getracht te abstraheren van concrete biasbronnen en gepoogd tot generalisa-

ties te komen. Bij deze werkwijze doemden drie belangrijke problemen op.

Op de eerste plaats zal om na te kunnen gaan welke vaardigheden onterecht gemeten worden, eerst duidelijk moeten zijn welke vaardigheden een bepaald item beoogt te meten. Het beschikbare Doelenboek (CITO 1986) biedt in dit opzicht wel informatie over de kerndoelstelling, maar het is onduidelijk hoe deze zich verhoudt tot de itemcontext. Bij itemmanipulaties is het derhalve niet altijd duidelijk of nog wel op dezelfde vaardigheden een beroep wordt gedaan als in de oorspronkelijke opgave.

Ten tweede kon tijdens de gesprekken niet altijd goed nagegaan worden of de kinderen de bedoelde vaardigheden beheersten. Bij opgaven die in afzonderlijke concrete stappen opgelost moesten worden, konden deze vaardigheden apart bevraagd worden, maar dat was minder eenvoudig bij opgaven waarbij dat niet het geval was.

Ten derde gaat het bij de ongemanipuleerde en gemanipuleerde versie, zoals reeds benadrukt, om andere leerlingen. Ze zijn weliswaar zo goed mogelijk geselecteerd op vergelijkbare capaciteiten e.d. maar volledig vergelijkbare groepen zijn uiteraard onmogelijk. Bovendien zijn het kleine groepen met daardoor een verhoogde kans op toevalsfouten. De biasbronnen die nu alleen bij allochtonen gevonden zijn zouden bij een grotere groep ook bij autochtonen kunnen voorkomen.

Voornaamste biasbronnen

Zoals verwacht liggen de biasbronnen voor een groot deel op het gebied van woordgebruik en impliciete zins- en tekstverbanden. Alle infrequente woorden en uitdrukkingen die gemanipuleerd zijn, leken in de ongemanipuleerde versie inderdaad bias te veroorzaken en bijna alle manipulaties op dit gebied leken effect te hebben. Daarnaast bleken een ongebruikelijke uitdrukking (*niet te laat*) en een woordvormgelijkenis (*half procent/helft*) tot problemen te leiden. Waarschijnlijk kan 'woordgebruik' als belangrijke biasbron verklaard worden door het feit dat er voor T2-leerders veel minder gunstige omstandigheden aanwezig zijn dan voor T1-leerders om een grote Nederlandse woordenschat op te bouwen. Ook het feit dat ambiguïteit en impliciete relaties voor deze leerlingen moeilijker zijn op te lossen is waarschijnlijk te wijten aan hun zwakkere Nederlandse taalvaardigheid.

De gevonden problematische visuele aspecten berusten bijna allemaal op onduidelijkheden die verwarrend kunnen werken. Ze kunnen meestal gemakkelijk verholpen worden zonder dat cruciale vaardigheden aangetast worden.

Ten slotte lijkt complexiteit van de opgaven een rol te spelen. Niet complexe opgaven behoren meestal tot domeinen, waarvan de doelstellingen relatief gemakkelijk te operationaliseren zijn. Bij hoofdrekenen of spelling bijvoorbeeld is het relatief eenvoudig opgaven te maken die het domein goed representeren, omdat deze domeinen uit eindige lijsten van doelstellingen bestaan. Domeinen als 'rekenvraagstukjes' en 'het trekken van conclusies op basis van combinatie van gegevens uit de tabel of grafiek met algemene kennis van de wereld' (CITO 1986) zijn veel minder eenduidig gedefinieerd waardoor de kans op binnensluipen van niet beoogde vaardigheden groter is. Bovendien vereisen dit soort opgaven veel context met alle kans dat gedeelten daarvan niet communaal zijn.

Een andere, vooralsnog nogal speculatieve, verklaring zou gevonden kunnen worden in het gedifferentieerde onderwijsaanbod dat op veel scholen gehanteerd wordt: de 'verborgen agenda' (Jungbluth, 1985). In dit licht lijkt het aannemelijk dat alle leerlingen op school de toepassing van relatief eenvoudige regels aangeleerd krijgen. Daarvan uitgaande hebben allochtone kinderen evenveel kansen en mogelijkheden als autochtonen om regels voor bijvoorbeeld hoofdrekenen en spelling te leren. Ze lijken echter in mindere mate toe te komen aan vaardigheden, die niet in eenvoudige regels te vangen zijn. Een mogelijke oorzaak daarvoor zou kunnen zijn dat leerkrachten (onbewust) a priori hun verwachtingen over de prestaties van allochtonen met name bij talige onderwijsonderdelen op een lager plan stellen. Allochtonen krijgen daardoor minder oefening in complexe taken aangeboden en blijken deze bijgevolg in de Eindtoets minder goed te beheersen.

Andere verklaringen dan de bovenstaande zijn natuurlijk niet uitgesloten. Daarbij kan gedacht worden aan factoren als attitude, motivatie en faalangst. In verband met het laatste is uit onderzoek (zie Hermans, 1975, p. 93) gebleken dat het effect van faalangst op prestatiegedrag afhankelijk is van de aard van de taak. Bij moeilijke taken leveren mensen met veel faalangst slechtere prestaties dan

mensen met weinig faalangst, maar bij gemakkelijke taken geldt het omgekeerde.

Literatuur

Boonman, J. & W. Kok, *Kennis verwerven uit teksten*. Utrecht: Vakgroep Onderwijskunde RUU, 1986.

Centraal Instituut voor Toetsontwikkeling, *Doelenboek, inhoudsverantwoording van de Eindtoets Basisonderwijs*. Arnhem: CITO, 1986.

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

Jungbluth, P., *Verborgen differentiatie, leerlingbeeld en onderwijsaanbod op de basisschool*. Nijmegen: ITS, 1985.

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

Hermans, H., *Prestatiemotief en faalangst in gezin en onderwijs*. Amsterdam: Swets en Zeitlinger, 1975.

Holland, P. & D. Thayer, *Differential item performance and the Mantel-Haenszel procedure*. Paper presented at the American Educational Research Association Annual Meeting, San Francisco, april 1986.

Kok, F., *Vraagpartijdigheid*. Amsterdam: SCO (uitgave in eigen beheer), 1988.

Scheunemann, J., Item bias and individual differences. In: S. Irvine (Ed.), *Human assessment in computer context*. Den Haag: Nijhoff, 1988.

Tobi, H., *DIF-onderzoek met behulp van de Mantel-Haenszel procedure bij de Eindtoets Basisonderwijs 1987*. Interne documentatie nr. 303. Arnhem: CITO, 1987.

Uiterwijk, J., *Item- en testbias in de Eindtoets Basisonderwijs 1987*. Onderzoeksrapporten basis- en speciaal onderwijs 1. Arnhem: CITO, 1990.

Verhelst, N., *De Mantel-Haenszel-toetsen*. Interne documentatie nr. 271. Arnhem: CITO, 1988.

Curricula vitae

M. Coenen (1963) specialiseerde zich tijdens haar studie Taal- en Literatuurwetenschap aan de KUB op het terrein van Taal & Minderheden. Voor en na haar afstuderen (in 1988) werkte ze bij het Werkverband Taal & Minderheden van de Letterenfaculteit van de KUB aan enkele projecten op het gebied van woordenschatverwerving door allochtone kinderen en deed daarna onderzoek naar linguïstische en culturele bronnen van itembias in de Eindtoets. Momenteel is ze werkzaam bij het Projectbureau OVB te Rotterdam ten behoeve van leermiddelenontwikkeling voor allochtone kinderen.

T. Vallen (1946) studeerde taalwetenschap, was tot 1981 werkzaam aan de Letterenfaculteit van de KUN en promoveerde in hetzelfde jaar aan dezelfde universiteit op een samen met S. Stijnen geschreven proefschrift over de onderwijsproblemen van dialectsprekende basisschoolleerlingen. Is sinds 1981 verbonden aan het Werkverband Taal & Minderheden van de Letterenfaculteit van de KUB in Tilburg en doet daar onderzoek naar en publiceert op de terreinen van sociolinguïstiek, taalbeleid, taalvariatie, taalattitudes en taaltoetsing, met name in relatie tot het eerste- en tweede-taalonderwijs voor etnische minderheden.

Adres: Katholieke Universiteit Brabant, Faculteit der Letteren, Postbus 90153, 5000 LE Tilburg

Manuscript aanvaard 22-8-'90

Summary

Coenen, M. & T. Vallen. 'Itembias in the CITO final primary schooltests.' *Pedagogische Studiën*, 1991, *68*, 15-26.

De Jong and Vallen (1989) reported on the most important theoretical points of view with respect to possible linguistic and cultural sources of item bias in the final primary school tests of het National Institute for Educational Measurement (CITO). Elaborating upon these viewpoints, this contribution focuses on an experiment in which the central question was whether linguistic manipulation of a number of statistically strongly biased items in the 1987 CITO final primary school tests yields results other than the results on those items in their original form. In consultation with the teachers, 22 pairs of 'comparable' indigenous and non-indigenous (mainly Turkish and Moroccan) 8th graders were selected to take part in the experiment. Half the pairs took the original and half took the manipulated version of the items in question. The children had to give an extensive oral explanation of how they solved the problems. The non-indigenous pupils performed better on the manipulated items than on the original non-manipulated ones, while the indigenous pupils hardly showed any differences between the two item versions. In spite of a number of interpretation problems, the experiment indicated that word choice, implicit sentence and text relationships, and item complexity were the main sources of item bias for non-indigenous children.