

Samenvatting van het proefschrift van Tamara van Schilt-Mol: Differential Item Functioning en Itembias in de Cito-Eindtoets Basisonderwijs.

Het proefschrift is uitgegeven door Aksant Academic Publishers, Amsterdam (ISBN 978 90 5260 260 8) en kost € 29,50.

Inleiding

Hoewel de onderwijsresultaten van Turkse en Marokkaanse leerlingen de laatste jaren zijn verbeterd, is er aan het einde van de basisschool nog steeds sprake van een structurele achterstand ten opzichte van vergelijkbare Nederlandse leerlingen (o.a. Dagevos, Gijsberts & Van Praag, 2003; Distelbrink & Hooghiemstra, 2005; Van Praag, 2006).

Tal van onderzoeken zijn uitgevoerd om na te gaan in hoeverre factoren als sociaaleconomische achtergrond (o.a. Van der Hoek, 1994; Van der Veen, 2001), schoolkenmerken (o.a. Jungbluth, 2003) of etnische/culturele achtergrond (o.a. Driessen, 1997; Nijsten, 1998; Crul, 2000) daarvan de oorzaak zijn. Opmerkelijk is echter dat er in Nederland weinig onderzoek is uitgevoerd naar de vraag in hoeverre deze scoreverschillen enkel zijn toe te schrijven aan aspecten als thuistaal en schoolkenmerken of aan eventuele andere factoren. Uiterwijk (1994: 1) merkt in dit verband het volgende op:

“Vooraf wanneer de gemiddelde toetscores van onderscheiden groepen, zoals allochtone en autochtone leerlingen, aanzienlijk verschillen, kan de onderzoeker zich immers afvragen of die verschillen toe te schrijven zijn aan verschillen in de te meten vaardigheden of dat ze een artefact zijn van de gehanteerde meetprocedure.”

De noodzaak om deze vraag te stellen wordt groter naarmate het belang dat aan een toets gehecht wordt toeneemt. Het is dan ook niet verwonderlijk dat de weinige onderzoeken naar de vraag of een toets in gelijke mate bruikbaar is voor verschillende subgroepen, voornamelijk gericht zijn op intelligentietests en op schoolvorderingstoetsen als de Cito-Eindtoets Basisonderwijs. De resultaten op dergelijke instrumenten kunnen immers grote implicaties hebben voor de (onderwijs)toekomst van de toetsdeelnemers.

Als (schoolvorderings-)toetsen bedoeld zijn om te differentiëren tussen ‘goede’ en ‘slechte’ leerlingen, moeten docenten er vanuit kunnen gaan dat de scores die door verschillende subgroepen leerlingen behaald zijn op eenzelfde manier geïnterpreteerd kunnen worden. Onderzoek heeft echter aangetoond dat bepaalde toetsen opgaven bevatten die verschillend functioneren voor verschillende subgroepen leerlingen, zelfs wanneer deze leerlingen een vergelijkbaar prestatieniveau hebben (o.a. Uiterwijk, 1994). Dit verschijnsel wordt aangeduid als Differential Item Functioning (DIF), ook wel omschreven als ‘onbedoelde moeilijkheden’. Er is sprake van onbedoelde moeilijkheden als leerlingen uit verschillende subgroepen met een vergelijkbaar prestatieniveau een ongelijke kans hebben om een toetsopgave juist te beantwoorden. Wanneer de oorzaak van deze onbedoelde moeilijkheid niet behoort tot het construct dat de opgave beoogt te meten is er sprake van itembias.

Dit is bijvoorbeeld het geval wanneer additionele vaardigheden of kennis nodig zijn om een opgave juist te kunnen beantwoorden. Een voorbeeld van itembias is een opgave uit de Eindtoets Basisonderwijs 1987 (onderdeel Rekenen, opgave 57).

Figuur 1: Opgave 57 – CITO Eindtoetsbasisonderwijs 1987

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,-	C	f 820,-
B	f 640,-	D	f 960,-

Uiterwijk (1994) heeft door middel van een hardopdenkprocedure bij leerlingen getracht te achterhalen wat de oorzaak is van het verschillend functioneren van deze opgave. Uit zijn onderzoek blijkt dat de oorzaak van DIF hoogstwaarschijnlijk het woord ‘inclusief’ is. Bij deze opgave is niet het doel om na te gaan of de leerlingen weten wat de betekenis van het woord ‘inclusief’ is, maar om te meten of leerlingen in staat zijn te rekenen met percentages. Als ervan uitgegaan wordt dat het begrip ‘inclusief’ niet behoort tot het domein Rekenen-Wiskunde, en niet als algemene kennis van groep 8 van het basisonderwijs kan worden beschouwd, kan dus gesteld worden dat hier sprake is van het toetsen van additionele kennis en vaardigheden, anders dan die beoogd worden te meten. Bij deze specifieke opgave is er, met andere woorden, dus niet alleen sprake van DIF, maar ook van itembias.

Daarnaast is het echter ook mogelijk dat een opgave wel DIF bevat, maar dat deze geen itembias-opgave is. Zo is het bijvoorbeeld mogelijk dat de bron van DIF wel behoort tot het kennisdomein dat de opgave beoogt te meten. Een voorbeeld van een opgave waarbij wel sprake is van DIF, maar niet van itembias, is de volgende, door Uiterwijk (1994) onderzochte rekenopgave.

Figuur 2: Opgave 41 – CITO Eindtoetsbasisonderwijs 1987

Wat is het gemiddelde van de volgende getallen:

1 ; 2 ; 3 ; 99 ; 98 ; 97

A	50	C	150
B	100	D	300

Uit het onderzoek van Uiterwijk komt naar voren dat de vermoedelijke oorzaak van DIF in deze opgave het woord 'gemiddelde' is. Omdat dit rekenbegrip behoort tot de (reken)vaktaal zoals die in groep 8 van de basisschool aan de orde is, mag verondersteld worden dat kennis van dit begrip bij alle groep 8 leerlingen aanwezig zou moeten zijn. Hoewel het begrip hoogstwaarschijnlijk verantwoordelijk is voor het verschillend functioneren van deze opgave voor uiteenlopende groepen leerlingen, is er geen sprake van itembias.

Onderzoeksvragen

In dit onderzoek worden de volgende onderzoeksvragen beantwoord:

1. Welke kenmerken van de (talig-culturele) gezinssocialisatie en de schoolse context van leerlingen zijn van invloed op de scores op de Eindtoets Basisonderwijs?
2. In welke mate bevat de Eindtoets Basisonderwijs 1997 (EB97) DIF voor leerlingen met een Turkse of Arabisch/Berberse thuistaalachtergrond?
3. Wat zijn de bronnen van DIF in de opgespoorde DIF-opgaven?
4. Is er bij de opgaven die DIF bevatten ook sprake van itembias?

Om deze vragen te beantwoorden zijn in vier opeenvolgende onderzoeksfases verschillende (deel)onderzoeken uitgevoerd waaraan uiteenlopende informantgroepen hebben deelgenomen. Daarnaast zijn verschillende statistische procedures toegepast om (1) DIF-opgaven op te sporen in de Cito-Eindtoets Basisonderwijs 1997, (2) na te gaan welke relaties er bestaan tussen specifieke leerlingkenmerken en de toetsscores en (3) te toetsen of de opgestelde verwachtingen over de oorzaken van DIF al dan niet bevestigd worden. Tabel 1 geeft een overzicht van de vier onderzoeksfases, de daartoe behorende (deel)onderzoeken en de daarbij uitgevoerde statistische analyses, en de diverse informantgroepen.

Tabel 1: Overzicht (deel)onderzoeken en informantgroepen per onderzoeksfase.

Deelonderzoek	Informanten	N
Fase 1. Opsporen DIF-opgaven		
Leerlingvragenlijst	Brugklasleerlingen	5026
Statistische analyse opgaven EB97		
Fase 2. Achterhalen potentiële bronnen van DIF		
Deelonderzoek 1: Plus-en-minmethode	Leerlingen groep 8	16
Deelonderzoek 2: Plus-en-minmethode	Pabostudenten	19
Deelonderzoek 3: Hardopdenkprocedure	Leerlingen groep 8	16
Deelonderzoek 4: Herschrijfprocedure	Pabo-studenten	19
Deelonderzoek 5: Gecombineerde plus-en-minmethode en herschrijfprocedure	Leerkrachten groep 8, toetsconstructeurs en taalkundigen	26 5 2
Fase 3. Toetsen potentiële bronnen van DIF		
Manipulatie-exercitie	Leerlingen groep 8	2526
Statistische analyse gemanipuleerde opgaven		
Fase 4. DIF vs. itembias		
Inhoudelijke analyse DIF-opgaven EB97 en gemanipuleerde opgaven		
Enquête	Leerkrachten groep 8, en experts	13 4

In deze samenvatting zullen de antwoorden op de gestelde onderzoeksvragen afzonderlijk worden besproken.

Onderzoeksvraag 1:

Welke kenmerken van de (talig-culturele) gezinssocialisatie en de schoolse context van leerlingen zijn van invloed op de scores op de Eindtoets Basisonderwijs?

In fase 1 is een vragenlijst afgenomen bij 5026 leerlingen die in 1997 aan de Eindtoets Basisonderwijs hebben deelgenomen. De uitgevoerde statistische analyse van de resultaten van deze vragenlijst laat zien dat er grote verschillen bestaan in de mate waarin kenmerken van leerlingen samenhang vertonen met de behaalde scores op de onderdelen van de EB97 door Nederlandse, Turkse en Marokkaanse leerlingen. Bij slechts één leerlingachtergrondkenmerk is er sprake van grote effecten. Dit is het geval bij de door de ouders onderling gesproken thuistaal. De verschillen in de behaalde scores tussen leerlingen van wie de ouders onderling Nederlands, Turks of Arabisch/Berbers spreken zijn significant voor alle toetsonderdelen. Voor alle leerlingen is er sprake van een groot effect van de gesproken thuistaal op de toetsonderdelen Taal, Informatieverwerking en Wereldoriëntatie. De thuistaal heeft een middelgroot effect op het onderdeel Rekenen.

Naast het middelgrote effect van de door de ouders gesproken thuistaal op het onderdeel Rekenen is er ook bij zeven andere leerlingkenmerken sprake van middelgrote significant effecten op de Eindtoetscores.

1. **Sekse.** Voor alle leerlingen is er sprake van een middelgroot effect op het onderdeel Wereldoriëntatie. Voor de Turkse en Marokkaanse leerlingen geldt daarnaast dat sekse een middelgroot effect heeft op de scores van het onderdeel Rekenen. In al deze gevallen behalen jongens een significant hogere score dan meisjes.
2. **Opleidingsniveau van de vader.** Voor alle leerlingen heeft een hoog opleidingsniveau van de vader positieve middelgrote effecten op de toetsresultaten van het onderdeel Informatieverwerking. Bij de Nederlandse leerlingen is dit ook bij de onderdelen Taal en Rekenen het geval.
3. **Mate van spreken over actualiteiten.** Voor de Turkse leerlingen heeft het regelmatig met de vader spreken over actualiteiten een positief middelgroot effect op de behaalde score bij het onderdeel Rekenen.
4. **Mate van computergebruik.** Voor de Turkse en Marokkaanse leerlingen geldt dat de mate waarin zij de computer gebruiken een middelgroot effect heeft op de scores op het onderdeel Wereldoriëntatie. Daarnaast is er voor de Turkse leerlingen ook sprake van middelgrote effecten op de scores bij de onderdelen Rekenen en Informatieverwerking. In al deze gevallen behalen de leerlingen significant hogere scores wanneer zij vaak achter de computer zitten.
5. **Gebruiksdoeleinden computer.** Voor Turkse jongens geldt dat het spelen van spelletjes op de computer een positief middelgroot effect heeft op de behaalde scores bij de onderdelen Informatieverwerking en Wereldoriëntatie.
6. **Lidmaatschap van een club.** Er is sprake van positieve middelgrote effecten op de scores van de Nederlandse leerlingen bij de onderdelen Rekenen en Informatieverwerking als de leerlingen lid zijn geweest van een club.
7. **Mate van televisiekijken.** De mate waarin Nederlandse leerlingen televisiekijken heeft een middelgroot effect op de behaalde scores bij de onderdelen Taal, Rekenen en Informatieverwerking. Voor deze leerlingen geldt dat zij significant hogere scores behalen wanneer zij weinig televisiekijken.

Onderzoeksvraag 2:

In welke mate bevat de Eindtoets 1997 DIF voor leerlingen met een Turkse of Arabisch/Berberse thuistaalachtergrond?

Uit eerder uitgevoerd onderzoek (o.a. Uiterwijk, 1994; Uiterwijk & Vallen, 1997, 2005) is gebleken dat de Cito Eindtoets Basisonderwijs opgaven bevat waarbij sprake is Differential Item Functioning (DIF).

Omdat er bij deze opgaven bovendien sprake kan zijn van itembias is het van groot belang de oorzaken van DIF te achterhalen, zodat deze in toekomstige Eindtoetsen voorkomen kunnen worden. In dit onderzoek is er voor gekozen DIF-opgaven op te sporen met behulp van twee statistische procedures: Mantel-Haenszel (MH, Mantel & Haenszel, 1959), gebaseerd op geobserveerde toetsscores, en het One Parameter Logistic Model (OPLM, Verhelst, 1992), gebaseerd op de itemresponsetheorie.

Uit de DIF-analyses komt naar voren dat de Cito-Eindtoets Basisonderwijs 1997, die bestaat uit 240 opgaven, in totaal 32 DIF-opgaven bevat. Hiervan vertonen er 21 DIF in het nadeel en 11 in het voordeel van Turkse/Marokkaanse leerlingen. Opvallend is dat het merendeel van deze opgaven zich bevindt in de toetsonderdelen Taal en Wereldoriëntatie. Van 50% van de opgespoorde opgaven geldt dat deze door beide procedures zijn gedetecteerd als DIF-opgave. Voor 13 opgaven geldt dat deze uitsluitend door MH zijn opgespoord. Drie opgaven zijn enkel door OPLM gedetecteerd. In onderstaande tabel is weergegeven hoeveel opgaven er per toetsonderdeel DIF vertonen.

Tabel 2: aantal opgaven dat DIF vertoont op toetsonderdeel, onderverdeeld naar de statistische procedure waarmee de opgaven zijn opgespoord.

	MH	OPLM	MH & OPLM
Taal nadeel	4	–	5
Taal voordeel	1	–	3
Rekenen nadeel	1	–	–
Rekenen voordeel	1	–	–
Informatieverwerking nadeel	4	–	–
Informatieverwerking voordeel	–	1	–
Wereldoriëntatie nadeel	2	1	4
Wereldoriëntatie voordeel	–	1	4

Onderzoeksvraag 3:

Wat zijn de bronnen van DIF bij de opgespoorde DIF-opgaven?

Uit de eerste onderzoeksfase is gebleken dat 32 opgaven DIF vertonen, waarvan 11 in het voordeel en 21 in het nadeel van Turkse/Marokkaanse leerlingen. Om inzicht te krijgen in de mogelijke oorzaken van DIF in deze opgaven zijn in de tweede onderzoeksfase vijf deelonderzoeken uitgevoerd (zie Tabel 1).

Op basis van de resultaten van deze deelonderzoeken zijn vijf verwachtingen geformuleerd over de mogelijke bronnen van DIF in het nadeel van Turkse/ Marokkaanse leerlingen, drie verwachtingen over mogelijke bronnen van DIF in het voordeel van deze leerlingen en twee aanvullende verwachtingen.

Verwachtingen over mogelijke bronnen van DIF in het nadeel van Turkse/Marokkaanse leerlingen:

1. Aandacht voor specifieke opgave-kenmerken

Uit de deelonderzoeken is naar voren gekomen dat Turkse en Marokkaanse leerlingen meer aandacht besteden aan een aan de opgave toegevoegde illustratie en aan een aan de opgave toegevoegde tekstuele context dan Nederlandse leerlingen. Dit doen zij vooral wanneer zij menen het juiste antwoord op de gestelde vraag niet te weten. De nadruk die door de Turkse en Marokkaanse leerlingen tijdens het beantwoorden van de vraag op de voor hen mogelijk ambigue elementen, de illustratie, het woordgebruik in de antwoordmogelijkheden en de tekstuele context wordt gelegd, kan ertoe leiden dat de leerlingen gebruik gaan maken van zogenaamde verticale relaties. Dit houdt in dat leerlingen verbanden proberen te leggen tussen de vraag of de illustratie enerzijds en de antwoordmogelijkheden anderzijds. Omdat deze verbanden vaak ten onrechte blijken te zijn, leidt dit regelmatig tot een fout antwoord.

2. Standaardantwoordmogelijkheden

Uit de analyses van de deelonderzoeken is gebleken dat de standaard antwoordmogelijkheden 'zo laten staan' en 'geen fout' verschillend gewaardeerd worden door de Turkse, Marokkaanse en Nederlandse leerlingen. Hieruit kunnen twee deelverwachtingen worden afgeleid:

- Voor de standaard antwoordmogelijkheid 'zo laten staan' geldt dat deze door de Turkse en Marokkaanse leerlingen niet correct geïnterpreteerd wordt. Uit de analyses van de deelonderzoeken blijkt dat deze standaardantwoordmogelijkheid door de Turkse en Marokkaanse leerlingen beschouwd wordt als een antwoordmogelijkheid die, naast de door Cito als correct beschouwde antwoordmogelijkheid, ook goed is. In geval van twijfel zijn deze leerlingen daardoor eerder geneigd te kiezen voor 'zo laten staan'.
- Uit de analyse van de plus-en-minmethode komt naar voren dat Turkse leerlingen, en ook de leerkrachten en Pabostudenten, negatief oordelen over de standaardantwoordmogelijkheid 'geen fout' bij de spellingopgaven. Volgens de leerlingen leidt de toevoeging van deze antwoordmogelijkheid ertoe dat leerlingen die het juiste antwoord niet weten zullen kiezen voor 'geen fout' om zo het maken van een keuze tussen de andere antwoordmogelijkheden te vermijden. Toevoeging van een andere afleider de leerlingen zou hen dwingen om een keuze te maken.

3. Taalgebruik

Uit de analyses van de deelonderzoeken is gebleken dat specifieke taalgebruiksaspecten, zowel in de vraag en antwoordmogelijkheden als in de tekst waarover een vraag gesteld wordt, de oorzaak kunnen zijn van DIF in het nadeel van Turkse en Marokkaanse leerlingen. Het gaat hierbij om de volgende aspecten:

- het gebruik van laagfrequente en/of volgens leerlingen te moeilijke woorden;
- het gebruik van laagfrequente en/of volgens leerlingen te moeilijke uitdrukkingen;
- het gebruik van complexe zinnen;
- het gebruik van complexe verwijzingen.

4. Afleiders

Uit de analyses van de deelonderzoeken is naar voren gekomen dat de wijze waarop een cluster van vier antwoordmogelijkheden is samengesteld, in een aantal gevallen de oorzaak kan zijn van DIF in het nadeel van Turkse en Marokkaanse leerlingen. Zowel door de leerlingen uit groep 8 als door de leerkrachten en Pabostudenten wordt aangegeven dat voor een aantal opgaven geldt dat meer dan één van de antwoordmogelijkheden aangeduid kan worden als correct. Uit de analyses van de deelonderzoeken blijkt dat de Turkse en Marokkaanse leerlingen in geval van twijfel tussen twee of drie juiste antwoordmogelijkheden, volgens leerlingen en leerkrachten geneigd zijn te kiezen voor de antwoordmogelijkheid waarin de meest frequente woorden of uitdrukkingen voorkomen.

5. Herlezen van tekstpassages

Leerlingen uit groep 8, leerkrachten en Pabostudenten geven aan dat opgaven waarbij de leerlingen (lange) tekstpassages moeten herlezen om de vraag te kunnen beantwoorden een groter beroep doen op het geheugen van de leerling dan op de taalvaardigheid. Daarnaast vinden zij dat deze specifieke opgaven door de extra vaardigheid die nodig is om de opgave te kunnen beantwoorden, moeilijk zijn voor taalzwakke leerlingen. In het verlengde hiervan is de verwachting geformuleerd dat het moeten herlezen van (lange) tekstpassages om de vraag te kunnen beantwoorden mogelijk de oorzaak is van DIF in het nadeel van Turkse/Marokkaanse leerlingen.

Verwachtingen over mogelijke bronnen van DIF in het voordeel van Turkse/Marokkaanse leerlingen:

6. Spellingfouten in werkwoordsvormen

In de literatuur wordt met regelmaat opgemerkt dat opgaven die vragen naar spellingfouten in werkwoorden in het voordeel kunnen werken van allochtone leerlingen (o.a. Uiterwijk & Vallen, 1997, 2005). Uit de statistische analyses blijkt dat dit ook geldt voor drie van de tien opgenomen opgaven in de Eindtoets Basisonderwijs 1997. In deze drie gevallen gaat om opgaven die de vragen naar de spelling van werkwoorden in de onvoltooid verleden tijd enkelvoud waarbij sprake is van een verdubbeling van de /t/, zoals bij 'barstte' en 'stortte'. Op basis hiervan is de verwachting geformuleerd dat taalopgaven die de vervoegingregels voor verledentijdsvormen bevragen de oorzaak kunnen zijn van DIF in het voordeel van Turkse en Marokkaanse leerlingen.

7. spellingfouten in niet-werkwoorden

Uit de DIF-analyses blijkt dat voor drie opgaven uit de opgavenrubriek 'spellen van niet-werkwoorden' geldt dat deze DIF vertonen. Hiervan blijkt echter slechts één opgave DIF

te vertonen in het voordeel van Turkse/Marokkaanse leerlingen. Bestudering van de betreffende opgaven wijst uit dat bij de voordeel-opgave de spelling bevestigd wordt van een woord waarbij de overtreffende trap wordt toegepast (*vreemste*), terwijl bij de andere opgaven woorden bevestigd worden waarvan de uitspraak niet overeenkomt met de correcte spelling (*felicitatie* en *korsje*). Op basis van deze analyse kunnen de volgende twee deelverwachtingen geformuleerd worden:

- het bevragen van woorden waarbij sprake is van een 'leerbaar proces', zoals de superlatief, kan net als de opgaven die 'leerbare' spellingregels zoals de toevoeging van /te/ aan de werkwoordstam in de onvoltooid verleden tijd bevragen, de oorzaak zijn van DIF in het *voordeel* van Turkse en Marokkaanse leerlingen;
- het bevragen van de spelling van woorden waarvan de gebruikelijke uitspraak niet overeenkomt met de correcte spelling, kan de oorzaak zijn van DIF in het *nadeel* van Turkse en Marokkaanse leerlingen.

8. Religie

Uit de statistische analyses blijkt dat alle opgaven die vallen binnen de opgavenrubriek 'Maatschappelijke verhoudingen en geestelijke stromingen' en die een met religie samenhangend onderwerp bevragen, DIF vertonen in het voordeel van Turkse en Marokkaanse leerlingen. Op basis hiervan is de verwachting geformuleerd dat het bevragen van een aan religie verwant onderwerp, ongeacht de godsdienst, de oorzaak kan zijn van DIF in het voordeel van Turkse/Marokkaanse leerlingen.

Aanvullende verwachtingen:

9. Schattend rekenen

Uit de analyses van de deelonderzoeken komt geen informatie naar voren over een mogelijke oorzaak van DIF (in het voordeel van Turkse/ Marokkaanse leerlingen) bij een rekenopgave die behoort tot de opgavenrubriek 'Schattend rekenen'. Uit de inhoudelijke en DIF-analyses van alle in 1997 opgenomen opgaven die dit tot deze opgavenrubriek behoren, is geen onderscheid aan te tonen tussen de rekenopgaven die zijn opgenomen in onderdeel 2 (hoofdrekenen) en onderdeel 3 (rekenen met uitrekenpapier) en tussen de rekenopgaven waarbij wel en geen illustratie is opgenomen. Wellicht zou de oorzaak van DIF kunnen liggen in het soort opgavenrubriek ('Schattend rekenen'). Nader onderzoek naar de potentiële bron is hier echter noodzakelijk.

10. Levende en niet-levende natuur

Bestudering van de DIF-opgaven uit het onderdeel Wereldoriëntatie laat zien dat relatief veel opgaven uit de opgavenrubrieken 'Levende natuur' en 'Niet-levende natuur' DIF vertonen in het nadeel van Turkse en Marokkaanse leerlingen (4 van de 20 opgaven). Hoewel bij de deelonderzoeken per opgave specifieke mogelijke oorzaken aangewezen worden, zou wellicht verondersteld kunnen worden dat opgaven die binnen de genoemde opgavenrubrieken vallen over het algemeen eerder DIF vertonen.

Om deze verwachtingen in de derde onderzoeksfase te toetsen zijn op de eerste plaats de oorspronkelijke DIF-opgaven uit de Eindtoets 1997 gemanipuleerd. Bij deze manipulaties zijn die opgave-aspecten veranderd waarvan op basis van de resultaten van de deelonderzoeken verwacht werd dat deze de oorzaak zouden zijn van DIF. Verwacht werd dan ook dat de gemanipuleerde opgaven DIF-vrij zouden zijn. Daarnaast zijn enkele niet eerder onderzochte opgaven uit de Eindtoetsen 2002 en 2003 geselecteerd waarvan op basis van de geformuleerde verwachtingen verondersteld werd dat deze wellicht ook DIF zouden kunnen vertonen. De gemanipuleerde opgaven zijn, samen met de geselecteerde opgaven uit de Eindtoetsen 2002 en 2003, voorgelegd aan twee groepen leerlingen, ieder bestaande uit Nederlandse, Turkse en Marokkaanse leerlingen uit groep 8.

Na afname van de gemanipuleerde en toegevoegde opgaven zijn deze statistisch geanalyseerd. Deze analyses laten zien dat vijf van de opgestelde verwachtingen naar alle waarschijnlijkheid terecht zijn. Tabel 3 geeft een overzicht van de opgestelde verwachtingen. Hierin wordt ook aangegeven of deze al dan niet bevestigd konden worden.

In de vierde onderzoeksfase, tot slot, is een enquête afgenomen onder leerkrachten van groep 8, een taalkundige, toetsconstructeur, onderwijskundige en Paboleerkracht, om een eerste antwoord te kunnen geven op de vraag of er bij de onderzochte opgaven behalve DIF ook sprake is van itembias.

Tabel 3: Opgestelde verwachtingen inzake potentiële bronnen van DIF in het nadeel en voordeel van Turkse/Marokkaanse leerlingen en eventuele bevestiging hiervan.

Verwachting		Bevestigd
Verwachtingen potentiële bronnen van DIF in het nadeel van Turkse/Marokkaanse leerlingen		
Verwachting 1	Aandacht voor specifieke opgave-kenmerken	Ja
Verwachting 2	Standaardantwoordmogelijkheden	Ja
Verwachting 3	Taalgebruik	Ja
Verwachting 4	Afleidings	Vooralsnog niet duidelijk
Verwachting 5	Herlezen van tekstpassages	Nee
Verwachtingen potentiële bronnen van DIF in het voordeel van Turkse/Marokkaanse leerlingen		
Verwachting 6	Spellingfouten in werkwoorden	Ja
Verwachting 7	Spellingfouten in niet-werkwoorden	Vooralsnog niet duidelijk
Verwachting 8	Religie	Ja
Aanvullende verwachtingen		
Verwachting 9	Schattend rekenen	Nee
Verwachting 10	Levende en niet-levende natuur	Nee

Onderzoeksvraag 4:

Is er bij de opgaven die DIF bevatten ook sprake van itembias?

In de vierde onderzoeksfase is geprobeerd een antwoord te geven op de vraag of er bij de 32 DIF-opgaven behalve DIF ook sprake is van itembias. Er is sprake van itembias wanneer de bron van DIF veroorzaakt wordt door construct-irrelevante opgavefactoren. De bron van DIF behoort in dit geval, met andere woorden, niet tot het kennisdomein dat de opgave beoogt te meten. Hierdoor spelen additionele kennis en vaardigheden, die niet iedere leerling noodzakelijkerwijs in dezelfde mate bezit, een rol bij het beantwoorden van de opgave. Ook kan het bij construct-irrelevante factoren gaan om onbedoelde 'aantrekkingskracht' van een of meerdere afleidings waardoor niet vanuit het kennisdomein gestuurde antwoorden kunnen worden gegeven.

Om antwoord te geven op de gestelde vraag moet dus op de eerste plaats worden nagegaan of de oorzaak van het verschillend functioneren van de opgave te wijten is aan construct-irrelevante factoren. In dit onderzoek is dit gedaan door, zoals aanbevolen door ETS (2004), na te gaan of er sprake is van een verschil tussen de kennis die het construct beoogt te meten en de kennis die daadwerkelijk nodig is om de opgave te kunnen beantwoorden. Wanneer dit het geval is en, met andere woorden, additionele kennis nodig is om de opgave te kunnen beantwoorden, is er sprake van construct-irrelevante factoren en vertoont een opgave dus itembias.

Wanneer er geen sprake is van construct-irrelevante factoren, moet worden nagegaan of de leerstof die in de opgave bevestigd wordt, behoort tot de leerstof die de overgrote meerderheid van de leerlingen uit groep 8 zou moeten beheersen. Om hier een antwoord op te kunnen geven, zijn 13 leerkrachten en 4 experts benaderd om hun mening te geven over de geschiktheid van het te meten construct van de 32 DIF-opgaven voor leerlingen van groep 8. De volgende vraag is hen daartoe voorgelegd:

"Heeft de leerstof die in de opgave aan de orde gesteld wordt, naar uw mening betrekking op de basisschoolleerstof die de overgrote meerderheid van de leerlingen uit groep 8 zou moeten beheersen?"

Wanneer de leerstof niet behoort tot de leerstof van groep 8, is de opgave ten onrechte opgenomen in de toets. Ook in dit geval is er sprake van itembias. Wanneer de leerstof echter wel behoort tot de leerstof van groep 8, is er enkel sprake van DIF.

Uit de analyse van de resultaten van de verschillende onderzoeken is echter gebleken dat er verschillen bestaan tussen de DIF-opgaven. In een aantal gevallen lijkt er sprake te zijn van zogenaamde 'valkuilen' in een DIF-opgave. Deze 'valkuilen' zijn opgave-elementen die bijvoorbeeld mogelijk ambigu, verwarrend of moeilijk zijn. Hierdoor kunnen zij de manier waarop leerlingen een opgave beantwoorden beïnvloeden. Daarnaast geldt voor deze 'valkuilen' dat de manier waarop deze vormgegeven zijn, geen noodzakelijke voorwaarde is om het beoogde construct te meten. Deze opgave-elementen kunnen, met andere woorden, aangepast worden zonder dat het te meten construct verandert. Hoewel het bij dergelijke opgaven, gezien het feit dat er alleen sprake is van DIF, niet noodzakelijk is de opgave aan te passen, wordt dat op

basis van de resultaten van de deelonderzoeken en de resultaten van de manipulatie-exercitie hierbij wel aanbevolen.

Voor DIF geldt, met andere woorden, dat er op basis van de resultaten van de deelonderzoeken en de manipulatie-exercitie een onderscheid gemaakt zou moeten worden tussen DIF waarbij aanpassing niet nodig is en DIF waarbij aanpassing wenselijk is.

Om in dit onderzoek een eerste antwoord te kunnen geven op de vraag of er behalve DIF ook sprake is van itembias is op de eerste plaats nagegaan of construct-irrelevante factoren de oorzaak zijn van DIF in de betreffende opgaven. Dit is gedaan door allereerst aan de hand van de kennisdomeinen vast te stellen welke kennis en vaardigheden met de specifieke opgaven beoogd worden te meten. Vervolgens is nagegaan of de bron van DIF al dan niet behoort tot deze kennis en vaardigheden. Wanneer dit niet het geval is en, met andere woorden, de daadwerkelijke benodigde additionele kennis en vaardigheden een verschil vertonen met de kennis en vaardigheden die de opgave beoogt te meten, is er sprake van construct-irrelevante factoren.

Bij de 26 opgaven waarvan de oorzaak is vastgesteld is vervolgens nagegaan of er sprake is van DIF of itembias. Bij negen van de 26 opgaven bleek er sprake te zijn van itembias (zie Tabel 4). Hiervan behoren drie opgaven tot het onderdeel Informatieverwerking, drie opgaven tot het onderdeel Wereldoriëntatie, twee opgaven tot het toetsonderdeel Taal en één opgave tot het onderdeel Rekenen.

Daarnaast zijn leerkrachten en experts gevraagd alle DIF-opgaven te beoordelen op de geschiktheid van de aan de orde gestelde leerstof voor leerlingen van groep 8. Daaruit komt naar voren dat in alle gevallen een meerderheid aangeeft dat de in de opgave aan de orde gestelde leerstof beheerst zou moeten worden door de leerlingen uit groep 8. Voor tien opgaven geldt dat de informanten het hier unaniem over eens zijn. Daarnaast is er bij drie opgaven sprake van een krappe meerderheid. Dit betekent dat voor de 17 opgaven waarbij geen sprake is van construct-irrelevante factoren geldt dat deze geen itembias vertonen.

Tot slot is op basis van de resultaten van de manipulatie-exercitie getracht een onderscheid aan te brengen tussen de 17 DIF-opgaven waarbij aanpassing niet nodig is en DIF-opgaven waarbij aanpassing wenselijk is (Tabel 4). Hieruit blijkt dat dit laatste het geval is voor vijf opgaven. Deze opgaven zijn afkomstig uit de toetsonderdelen Taal, Informatieverwerking en Wereldoriëntatie.

Tabel 4: Onderverdeling van de 26 opgaven in opgaven waarbij sprake is van DIF en opgaven waarbij sprake is van itembias.

Type DIF	DIF-items in het nadeel van Turkse/Marokkaanse leerlingen
DIF – aanpassing niet noodzakelijk	4
DIF – aanpassing wenselijk	5
Itembias	9
Nader onderzoek	3
Type DIF	DIF-items in het voordeel van Turkse/Marokkaanse leerlingen
DIF – aanpassing niet noodzakelijk	8
DIF – aanpassing wenselijk	
Itembias	
Nader onderzoek	3

Discussie

Een belangrijke vraag die op basis van de hierboven beschreven onderzoeksresultaten gesteld zou moeten worden is wat de implicaties zijn van de aanwezigheid van DIF en itembias in de Eindtoets Basisonderwijs, met name op de standardscore van de leerlingen.

Bij de berekening van de standardscore wordt de score die door de leerlingen is behaald op de onderdelen Taal, Rekenen en Informatieverwerking omgerekend naar een gestandaardiseerde score tussen 501 tot 550. Deze wordt jaarlijks gecorrigeerd voor de moeilijkheidsgraad van de toets en voor populatieverschillen met het voorgaande jaar (vanaf 2003 wordt deze score berekend op basis van de behaalde scores op de onderdelen Taal, Rekenen en Studievaardigheden). De standardscore kan vervolgens met behulp van onderzoek (het zogenaamde toelatings- en doorstroomonderzoek bij de Eindtoets Basisonderwijs) gekoppeld worden aan een bepaald type voortgezet onderwijs. Aan de hand

hiervan worden uitspraken gedaan over de kans van slagen van een leerling in een bepaalde onderwijstype.

Uit onderzoek van Cito blijkt dat 74% van de in het onderzoek onderzochte scholen voor voortgezet onderwijs bij de toelating gebruik maakt van vuistregels gebaseerd op de score op de Eindtoets Basisonderwijs (Van der Lubbe *et al.*, 2005). Wel hanteren de meeste scholen geen strikte norm; overleg over plaatsing lijkt bij de meeste scholen tot de mogelijkheid te behoren.

Niettemin bevestigt het hoge percentage scholen dat gebruik maakt van de standaardscore bij de plaatsing van de leerlingen niet alleen de grote waarde die door de scholen aan deze standaardscore toegekend wordt, maar ook de impact die de door de leerlingen behaalde score op hun onderwijstoekomst heeft. Het staat dan ook buiten kijf dat de behaalde standaardscore een goede afspiegeling zou moeten zijn van de daadwerkelijke capaciteiten van de leerlingen. Het gevolg van de aanwezigheid van itembias is dat de door de leerlingen behaalde score *geen* juiste afspiegeling is van hun capaciteiten. Dit komt doordat additionele kennis en vaardigheden die niet door alle leerlingen in gelijke mate beheerst worden een rol spelen bij het beantwoorden van opgaven met itembias.

Een hieruit voortvloeiende vraag is dan ook wat de daadwerkelijke gevolgen zijn van de aanwezigheid van itembias voor de advisering van leerlingen wat betreft de verschillende onderwijstypen in het voortgezet onderwijs en daarmee voor de voorspellende waarde van de Eindtoets. Binnen het uitgevoerde onderzoek is het echter niet mogelijk om de relatie na te gaan tussen Differential Item Functioning en itembias enerzijds en de voorspellende waarde van de toets anderzijds. Omdat de relatie tussen DIF en itembias enerzijds en voorspellende waarde van de toets anderzijds is van groot belang, met name bij toetsen waarvan de resultaten implicaties hebben voor de (onderwijs)toekomst van de toetsdeelnemers. Beslissingen worden immers genomen op basis van de resultaten op een toets en niet op basis van resultaten op afzonderlijke toetsopgaven. Nader onderzoek hiernaar wordt dan ook aanbevolen.

Aanbevelingen

Op basis van de resultaten van de verschillende (deel-) onderzoeken is het mogelijk enkele aanbevelingen te doen met betrekking tot het construeren van DIF-vrije toetsen.

Op de eerste plaats lijkt het goed mogelijk om met een tweetal kleinschalige onderzoeken, een plus-en-minmethode en een hardopdenkprocedure, voorafgaand aan de afname van een toets na te gaan of er bij een aantal opgaven wellicht sprake zou kunnen zijn van DIF. Uit de analyse van de resultaten van de in dit onderzoek uitgevoerde plus-en-minmethode en hardopdenkprocedure bleek dat, ondanks het beperkte aantal informanten, de informatie die door de informanten werd gegeven vrijwel gelijk was. Het bevragen van meer informanten zou, met andere woorden, hoogstwaarschijnlijk niet tot nieuwe of aanvullende informatie hebben geleid. Op basis van deze twee deelonderzoeken bleek het goed mogelijk mogelijke DIF-oorzaken te achterhalen. Bovendien bleken de Pabostudenten, nadat zij eerst zelf de opgaven kritisch hadden beoordeeld met behulp van de plus-en-minmethode en daarna een hardopdenkprocedure bij een Turkse of Marokkaanse leerling hadden afgenomen, goed in staat om uit alle beschikbare DIF-opgaven die opgaven te selecteren die DIF vertonen in het nadeel van Turkse/Marokkaanse leerlingen. Ook bleek het daarnaast mogelijk om op basis van de middels beide procedures opgedane kennis in niet eerder onderzochte Eindtoetsen opgaven aan te wijzen die, na statistische analyse, daadwerkelijk DIF bleken te vertonen. Dit was zowel mogelijk voor opgaven die DIF vertonen in het nadeel van Turkse en Marokkaanse leerlingen als voor opgaven die in hun voordeel werken. Op basis van het voorgaande wordt dan ook aanbevolen om opgaven waarvan op basis van de in dit onderzoek beschreven verwachtingen verondersteld kan worden dat deze mogelijkerwijs DIF vertonen, in een tweetal kleinschalige experimenten voor te leggen aan leerlingen van groep 8 (plus-en-minmethode en hardopdenkprocedure) en eventueel ook aan leerkrachten van groep 8 (plus-en-minmethode).

Op de tweede plaats is het aan de hand van de opgestelde en bevestigde verwachtingen mogelijk enkele **richtlijnen** te geven voor het construeren van zo veel mogelijk DIF-vrije toetsen.

- Op basis van de eerste verwachting, aandacht voor specifieke opgave-kenmerken, kan worden aanbevolen om tijdens het maken van opgaven extra aandacht en zorg te geven aan de tekstuele context van de opgaven en aan de illustraties die aan deze opgaven worden toegevoegd. Hoewel dit een voor de hand liggende aanbeveling lijkt, blijkt uit de praktijk dat met name illustraties weliswaar zorgvuldig worden uitgezocht, maar dat de impact die deze illustraties hebben wordt onderschat. Dit zou bij een kleine groep leerlingen, bijvoorbeeld met behulp van een hardopdenkprocedure, uitgeprobeerd moeten worden. Uit het onderzoek is immers gebleken dat het van groot belang is dat een opgave geen elementen bevat die mogelijk ambigu zijn. Illustraties en tekstuele contexten van opgaven dienen daarom zorgvuldig gecontroleerd te worden op elementen die zich kunnen lenen voor het door de leerlingen toepassen van verticale relaties en woordassociaties.
- De tweede verwachting, het gebruik van standaardantwoordmogelijkheden, leidt tot de aanbeveling dat, wanneer het construct van de vraag dit toelaat, afleiders als 'geen fout' en 'zo laten staan' vermeden zouden moeten worden. Uit de deelonderzoeken is gebleken dat deze antwoordmogelijkheden door Turkse en Marokkaanse leerlingen anders geïnterpreteerd worden dan door Nederlandse leerlingen. De eerstgenoemde groep leerlingen lijkt de standaardantwoordmogelijkheden te beschouwen als antwoorden die, naast het door Cito als correct beschouwde antwoord, *ook* goed zijn. Door deze standaardantwoordmogelijkheden te vermijden en te vervangen door een concreter antwoordalternatief wordt de leerling meer gedwongen een weloverwogen keuze te maken. Overigens moet hierbij vermeld worden dat de standaardantwoordmogelijkheid 'geen fout' sinds 2002 niet meer in de Eindtoets Basisonderwijs is opgenomen. Daarnaast wordt aanbevolen om informatie omtrent de wijze waarop Turkse en Marokkaanse leerlingen omgaan met dergelijke standaardantwoordmogelijkheden bekend te maken onder leerkrachten, zodat hier specifiek klassikaal aandacht aan besteed kan worden.
- Op basis van de derde verwachting kan gesteld worden dat het in een opgave gehanteerde taalgebruik toegankelijk dient te zijn voor alle leerlingen uit groep 8. Hoewel ook deze aanbeveling voor de hand liggend lijkt, is uit de manipulatie-exercitie gebleken dat 'te moeilijk' taalgebruik - als het niet tot het te meten construct behoort - in een aantal gevallen de oorzaak is van itembias. Aanbevolen wordt dan ook om te voorkomen dat een vraag, de antwoordmogelijkheden en/of de tekst die ter beantwoording van de opgave gelezen dient te worden, woorden en/of uitdrukkingen bevatten die niet noodzakelijkerwijs even bekend zijn bij alle leerlingen en die daardoor een belemmering kunnen vormen bij het beantwoorden van de opgave. Uiteraard is dit enkel het geval wanneer kennis van de woorden en/of uitdrukkingen niet behoren tot het kennisdomein dat de specifieke opgave beoogt te meten.

Voor de overige bevestigde verwachtingen, de verwachtingen 6 en 8 (respectievelijk 'spellingfouten in werkwoorden' en 'religie'), geldt dat deze enkel opgaven bevatten die DIF vertonen waarbij aanpassing niet noodzakelijk is. Het gaat hierbij om opgaven behorend tot de opgavenrubrieken "Spellen van werkwoorden" en "Maatschappelijke verhoudingen en geestelijke stromingen". Uit de analyses van de deelonderzoeken en de resultaten van de manipulatie-exercitie is gebleken dat de oorzaken van DIF bij de opgaven behorend tot deze rubrieken te wijten zijn aan daadwerkelijke verschillen in kennis en vaardigheden tussen de onderzochte subgroepen. Er is, met andere woorden, bij deze opgaven sprake van terechte differentiatie.

Bibliografie samenvatting

Crul, M.

2000 *De sleutel tot succes. Over hulp, keuzes en kansen in de schoolloopbanen van Turkse en Marokkaanse jongeren van de tweede generatie*. Amsterdam: Het Spinhuis.

Dagevos, J., M. Gijsberts & C. van Praag

2003 *Rapportage minderheden 2003. Onderwijs, arbeid en sociaal-culturele integratie*. Den Haag: Sociaal en Cultureel Planbureau.

Distelbrink, M. & E. Hooghiemstra

2005 *Allochtone gezinnen. Feiten en cijfers*. Den Haag: Nederlandse Gezinsraad.

- Driessen, G.
1997 Indicatoren van etniciteit in relatie tot predictoren van taalvaardigheid in het basisonderwijs. *Toegepaste Taalwetenschap in Artikelen*, 56 (1), 89-103.
- Educational Testing Service
2004 ETS international principles for fairness review of assessments. Princeton, NJ: Author.
http://www.ets.org/Media/About_ETS/pdf/frintl.pdf.
- Hoek, J. van der
1994 *Socialisatie in migrantengezinnen*. Utrecht: De Tijdstroom.
- Jungbluth, P.
2003 *De ongelijke basisschool: etniciteit, sociaal milieu, sekse, verborgen differentiatie, segregatie, onderwijskansen en schooleffectiviteit*. Nijmegen: ITS
- Lubbe, M. van der, N. Verhelst, T. Heuvelmans & G. Staphorsius
2005 *Verslag van een onderzoek naar de toelating van leerlingen in het voortgezet onderwijs*. Arnhem: Cito.
- Mantel, N. & W. Haenszel
1959 Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22, 719-748.
- Nijsten, C.
1998 *Opvoeding in Turkse gezinnen in Nederland*. Assen: Van Gorcum.
- Praag, C. van
2006 Opleiding en onderwijs. In: C. van Praag (Red.), *Marokkanen in Nederland: een profiel*. Den Haag: Nederlands Interdisciplinair Demografisch Instituut, 15-25.
- Uiterwijk, H.
1994 *De bruikbaarheid van de Eindtoets Basisonderwijs voor allochtone leerlingen*. Arnhem: Cito (proefschrift Katholieke Universiteit Brabant).
- Uiterwijk, H. & T. Vallen
1997 Onderzoek naar bias voor allochtone leerlingen in de Cito-Eindtoets Basisonderwijs. *Pedagogische Studiën*, 74 (1), 21-32.
2005 Linguistic sources of item bias for second generation immigrants in Dutch tests. *Language Testing*, 22 (2), 211-234.
- Veen, I. van der
2001 *Successful Turkish and Moroccan students in the Netherlands*. Leuven/Apeldoorn: Garant.
- Verhelst, N.
1992 *Het eenparameter logistisch model (OPLM), een theoretische inleiding en een handleiding bij het computerprogramma*. Arnhem: Instituut voor toetsontwikkeling.