REKENEN-BASISBEWERKINGEN VOOR GROEP 3 TOT EN MET 8

Uitgangspunten van de Testconstructie

'Rekenen-Basisbewerkingen voor groep 3 tot en met 8' is een reeks toetsen die ontwikkeld is om te toetsen in hoeverre de leerling de basisbewerkingen van rekenen (d.w.z. optellen, aftrekken, vermenigvuldigen en delen) goed en vlot beheerst. Het gaat om 'kale', ook wel contextvrije, rekenopgaven die een leerling moet kunnen oplossen, zoals '4 × 60' en '56 : 7'. Daarmee wijken deze toetsen af van de vaardigheidstoetsen 'Rekenen-Wiskunde' van het Cito Volgsysteem primair en speciaal onderwijs, die opgaven uit alle rekendomeinen bevatten. Er wordt bij de toetsen 'Rekenen-Basisbewerkingen voor groep 3 tot en met 8' niet alleen naar de accuratesse gekeken, maar ook naar de snelheid van het oplossen. Deze twee constructen worden dan ook apart gerapporteerd, naast een totaalscore. De toetsen dienen om de mate te onderzoeken waarin de basisbewerkingen geautomatiseerd zijn. Dit wordt in de literatuur gezien als voorwaarde om complexere rekenopgaven goed te kunnen uitvoeren en behoort dan ook tot de kerndoelen die door het Ministerie van Onderwijs, Cultuur en Wetenschappen gedefinieerd zijn. Wel zouden de precieze betekenis en het belang van het construct 'verwerkingssnelheid' theoretisch uitgebreider kunnen worden onderbouwd aan de hand van meer literatuur over rekenen en rekenonderwijs.

De toetsinhoud is niet alleen op de kerndoelen gebaseerd, maar ook op gegevens uit de Periodieke Peilingen van het Onderwijsniveau (PPON), uit Cito Leerlingvolgsysteemtoetsen (LVS-toetsen) van de eerste en de tweede generatie, uit de hulpboeken rekenen van het Leerlingvolgsysteem, uit veel gebruikte reken-wiskundemethoden en uit extra lesstof. Met name in de methoden is nagegaan wanneer de diverse typen opgaven aangeboden worden en op welke afnamemomenten zij beheerst zouden moeten worden. De toetsen zijn overigens methode-onafhankelijk.

De reeks bestaat uit elf toetsen die corresponderen met het beoogde afnamemoment (d.w.z. E3, M4, E4, M5, E5, M6, E6, M7, E7, M8 en E8) en hebben daarmee als doelgroep leerlingen van het primair onderwijs vanaf eind groep 3 tot en met eind groep 8. De auteurs stellen dat de toetsen ook gebruikt kunnen worden in het speciaal basisonderwijs en het speciaal onderwijs clusters 2 (dove en slechthorende kinderen), 3 (motorisch gehandicapte, verstandelijk gehandicapte en langdurig zieke kinderen) en 4 (kinderen met stoornissen en gedragsproblemen). De toetsen zijn echter alleen voor het reguliere basisonderwijs genormeerd. Ze worden niet geschikt geacht voor leerlingen met een visuele handicap.

Bij de toetsen voor eind groep 3, medio groep 4 en eind groep 4 worden alleen de onderdelen 'optellen' en 'aftrekken' getoetst. Bij de toetsen vanaf groep 5 komt daar het onderdeel 'delen' bij en bij de toetsen vanaf medio groep 6 komt ook het onderdeel 'vermenigvuldigen' aan bod. Per toets overlappen 5 items met de vorige en volgende toets in de reeks.

De auteurs stellen nadrukkelijk dat de toetsen geen voorspellende pretentie hebben: het doel is na te gaan in hoeverre de leerling op een bepaald moment de basisbewerkingen beheerst. De tabel op pagina 15 van de Wetenschappelijke Verantwoording laat dan ook zien welk onderwijsaanbod zou kunnen aansluiten op diverse combinaties van het niveau van accuratesse en snelheid. Bij deze COTAN beoordeling wordt er dan ook van uitgegaan dat criteriumvaliditeit niet van toepassing is. De auteurs stellen dat het evenmin de pretentie is dat de toetsen gezamenlijk een volgsysteem vormen, maar dat is een minder duidelijke uitspraak. Als men immers op een bepaald afnamemoment een zwakke leerling identificeert, zou een half jaar later kunnen worden nagegaan of deze leerling baat heeft gehad bij een interventie en nu wél beheerst wat bij diens leerjaar hoort. Een verheldering over het begrip 'volgsysteem' en over de relatie tussen deze reeks toetsen en de toetsen Rekenen-Wiskunde zou daarom goed zijn.

Kwaliteit van het Testmateriaal

<u>Papier-en-potloodversie</u>

N.v.t.

<u>Computerversie</u>

De opgaven zijn volledig gestandaardiseerd en bevatten geen inhoud die voor welk individu of groep dan ook kwetsend zou kunnen zijn, aangezien het om contextvrije sommen gaat. De opgaven zijn correct en duidelijk geformuleerd. De toetsen worden geautomatiseerd op een schoolnetwerk afgenomen. Ook het scoren van de meerkeuzevragen verloopt geautomatiseerd en is daarmee objectief.

De instructie voor de leerling is wat summier, maar gezien de eenvoudige aard van de opgaven en de niet meer dan basale computervaardigheden die voor de beantwoording nodig zijn is de instructie adequaat. Er wordt voldoende uitgelegd dat het niet alleen gaat om het geven van goede antwoorden, maar dat er ook vlot moet worden doorgewerkt. Het ontwerp van de software maakt de kans klein dat er per ongeluk verkeerd gebruik van zal worden gemaakt of dat opzettelijk buiten het toetsprogramma om andere software wordt opgestart.
De gebruikersinterface ziet er standaard en sober uit, maar dat is, mede gelet op het doel om de leerling zich te laten concentreren op goed en vlot rekenwerk, in orde.
De toegang tot de test is beveiligd met een inlognaam en wachtwoord. Over de toegang tot de items en de testresultaten wordt in de gebruikershandleiding en de wetenschappelijke verantwoording niets gezegd.

Kwaliteit van de Handleiding
Bij de toetsen hoort zowel een gebruikershandleiding als een wetenschappelijke verantwoording. De informatie over gebruiksmogelijkheden is helder. Beperkingen ten aanzien van de doelgroep zijn hierboven reeds besproken, dat wil zeggen er wordt gesteld dat de toets niet gebruikt mag worden bij leerlingen met een visuele handicap. Duidelijk wordt gesteld dat er bij dyscalculische leerlingen niet mag worden afgeweken van de standaard afnameprocedure.
De aanwijzingen voor de testleider zijn volledig en duidelijk, inclusief precieze bewoordingen die voorafgaand aan de toetsafname gebruikt mogen worden en aansporingen die gegeven kunnen worden als een leerling erg aarzelend of langzaam werkt.
De interpretatie is tamelijk uitvoerig beschreven. Mede gezien het feit dat deze toetsen niet heel ingewikkelde instrumenten genoemd kunnen worden, lijkt dit afdoende. Waardevol zijn de aanwijzingen over hoe de uitslag aanleiding kan geven tot een specifieke aanpak van het rekenonderwijs, bijvoorbeeld bij 'accuratesse hoog – snelheid laag' of 'accuratesse gemiddeld – snelheid hoog'. Op leerlingniveau levert de rapportage automatisch een categorieënanalyse en een antwoordenoverzicht, en op groepsniveau een categorieënanalyse. In de gebruikershandleiding is tevens een tabel opgenomen waarin de relatie gelegd wordt tussen het toetsresultaat op Rekenen-Basisbewerkingen en het resultaat op de LVS-toets Rekenen-Wiskunde. Wat hierbij gemist wordt, is levensechte casuïstiek over een of enkele leerlingen; dat zou het beeld van wat men met de toetsen kan doen concreter en levendiger maken. Er wordt niet specifiek gewezen op soorten informatie die bij de interpretatie van belang kunnen zijn, zoals de invloed die bepaalde achtergrondvariabelen zouden kunnen hebben.
Slechts impliciet wordt ingegaan op de mate van deskundigheid die voor afname en interpretatie van de toetsen vereist is. Een enkele maal staat er "gebruikers en andere professionals" en vaak noemt de handleiding leerkrachten. Of leerkrachten een cursus of training gevolgd moeten hebben, wordt nergens vermeld. Dat lijkt echter niet zo'n probleem te zijn. Inhoud en gebruikswijze van de toets staan goed beschreven in de handleiding en men kan argumenteren dat gebruik van dergelijke toetsen anno 2018 tot de ‘reguliere’ vakbekwaamheid van een basisschoolleerkracht behoort.
De vereisten voor de installatie van de software en de stappen die moeten worden doorlopen worden in een aparte installatiehandleiding duidelijk uiteengezet aan de hand van illustratieve screenshots. Een ICT-medewerker zal geen problemen hebben met het maken van mappen en hernoemen van bestanden, maar of de voorgeschreven werkwijze voor de ‘gewone’ leerkracht of schooldirecteur *foolproof* is, kan niet worden beoordeeld. De informatie over de mogelijkheden van de software en de aanwijzingen voor de bediening zijn adequaat.
De gebruikershandleiding bevat een lijst van veel gestelde vragen met de bijbehorende antwoorden. Tijdens kantooruren is een helpdesk telefonisch en per e-mail bereikbaar.

Normen
De eerste twee pilotstudies met de toetsen vonden plaats in 2012, waarna in januari en juni 2013 nog twee pilotstudies zijn uitgevoerd, die de basis waren voor de voorlopige normering. De gegevens van de werkelijke afnames in 2013-2014 en 2014-2015 (op basis van dataretour) zijn gebruikt voor de definitieve normeringsgegevens van de 11 normgroepen die gerelateerd zijn aan het afnamemoment (d.w.z. E3, M4, E4, M5, E5, M6, E6, M7, E7, M8 en E8). Nadere informatie over hoe de scholen ‘geworven’ zijn en/of hoe scholen geselecteerd en/of gevraagd zijn om dit nieuwe instrument te gebruiken ontbreekt. Verder bestaat onduidelijkheid over het totaal aantal deelnemende scholen aan datare-

tour (en dus hoe het zit met de respons rate), gezien het verschil in aantal leerlingen/afnames zoals vermeld in de tekst op pagina 20 (d.w.z. n = 67758) en het totaal aantal leerlingen van de elf normgroepen in Tabel 4.2 (d.w.z. n = 54397). Hoewel het niet voor alle scholen bekend is of hele klassen hebben deelgenomen, lijkt het op basis van navraag bij 20 scholen aannemelijk dat dit het geval is en dat de toets niet alleen bij een specifiek deel van de klas wordt afgenomen. Al met al is de beschrijving van de wijze van gegevensverzameling summier.
Hoewel aan de normering voor afnamemoment E8 relatief weinig scholen (d.w.z. 37 scholen) en leerlingen (n = 984) hebben deelgenomen, zijn de gebruikte aantallen per normgroep dusdanig groot dat de grootte van de normgroepen als 'goed' wordt beoordeeld. De range tussen het aantal scholen per normgroep is 164-238 (exclusief E8) en de leerlingaantallen per normgroep liggen tussen n = 4362 en n = 6331 (exclusief E8).
De representativiteit van de diverse normgroepen voor de Nederlandse basisschoolpopulatie is allereerst op schoolniveau onderzocht met betrekking tot de variabelen 'regio' (vier categorieën), 'verstedelijking' (vijf categorieën), 'schoolgrootte' (twee categorieën) en 'schooltype' (vier categorieën). De keuze voor deze vier variabelen is bij CITO gebruikelijk en wordt verder niet onderbouwd. De indeling van schoolgrootte op basis van meer of minder dan 200 leerlingen wordt niet beargumenteerd. Schooltype wordt bepaald op basis van het percentage leerlingen met een 'gewicht', dat gebaseerd is op het opleidingsniveau van de ouders. Hoewel de steekproefproporties in het algemeen redelijk overeenkomen met de populatiewaarden, zijn er echter ook afwijkingen, van circa 7 tot 12 procent; de regio's Noord en Oost, en kleine scholen zijn ondervertegenwoordigd. Met de effectmaat $\varphi = \sqrt{(\chi^2 / n)}$ beschrijft men de mate waarin steekproefproporties overeenkomen met populatieparameters. Met name voor de variabele regio valt deze effectmaat groot uit. Er is daarom onderzocht of weging van de steekproef voor deze variabele zinvol is door de ongewogen en ongewogen resultaten met elkaar te vergelijken. Vanwege de kleine verschillen in resultaten is besloten geen weging toe te passen. Op leerlingniveau is gekeken naar de variabele 'geslacht' en is de verdeling in de steekproef vergeleken met die in de populatie. Er bleken significante verschillen te zijn voor de normgroepen M6, E6 en M7, waarbij de effectmaat φ liet zien dat het om zeer kleine effecten gaat. Dit wordt tevens bevestigd door de kleine verschillen tussen de ongewogen en gewogen resultaten. Verder staat in de Wetenschappelijke Verantwoording vermeld dat de variabele 'etniciteit' ontbreekt omdat men stelt hier geen betrouwbare gegevens over te hebben. Daarbij wordt tevens onderzoek geciteerd waaruit zou blijken dat etniciteit sterk samenhangt met de variabelen 'verstedelijking' en 'schooltype'.

Voor de totaalscore, de scores voor accuratesse en snelheid én voor de onderdelen voor de afzonderlijke rekenbewerkingen (optellen, aftrekken, vermenigvuldigen en delen) worden percentielen als normschaal gebruikt. Het COTAN Beoordelingssysteem wijst er op dat percentielen alleen zinvol zijn als ook de ruwe scores voldoende klassen kennen en dat kan men zich bij de itemaantallen voor de onderdelen afvragen, aangezien het daar gaat om 20 of 25 items. Voor de totaalscore en de scores voor accuratesse en snelheid zijn percentielscores wel een verantwoorde keuze, ook gezien het feit dat scholen daar bekend mee zijn. Het nut van percentielen hangt uiteraard ook af van de vraag of de toets voldoende betrouwbaar is, hetgeen hier het geval is, zoals uit de bespreking daarvan bij *Betrouwbaarheid* zal blijken. De keuze voor het fijnmazige percentielsysteem – dat gemakkelijk tot over-interpretatie leidt − wordt in zekere zin weer afgezwakt doordat in de rapportage alleen leerlingen met een percentielscore van kleiner dan 20 of groter dan 80 in de klassen laag respectievelijk hoog vallen. In de gebruikershandleiding zou nadrukkelijker gewaarschuwd kunnen worden dat het geen zin heeft leerlingen met kleine percentielverschillen te vergelijken of aan kleine percentielverschillen in rekenonderdelen te veel betekenis toe te kennen. Dat op basis van de percentielscore (< 20, tussen 20 en 80, > 80) wordt gesproken van een lage, gemiddelde en hoge score zou overigens de indruk kunnen wekken dat er sprake is van een criteriumgerichte normering, maar dat is niet het geval. Het zijn arbitrair getrokken grenswaarden, die niet gebaseerd zijn op empirisch onderzoek waaruit blijkt dat een leerling onder percentiel 20 niet zonder extra hulp of specialistische interventie kan meekomen in het rekenonderwijs.
In Tabel 3.3 worden van alle normgroepen voor beide schalen gemiddelde, standaardafwijking, scheefheid en kurtosis vermeld. Voor de schaal 'accuratesse' wordt ook de gemiddelde p-waarde van de toets gegeven. Deze laat zien dat het uitgangspunt dat de toetsen qua moeilijkheid zouden moeten aansluiten bij wat een leerling geacht wordt op het meetmoment te beheersen, is waargemaakt.

De standaardmeetfout wordt vermeld in Tabel 5.2 van de Wetenschappelijke Verantwoording. Dit wordt gedaan voor de accuratesse-totaalscore van alle toetsen en tevens voor de onderdelen optellen, aftrekken, vermenigvuldigen en delen. Er wordt echter geen uitleg gegeven over de standaardmeetfout en de berekening ervan. Zodoende is niet duidelijk welke betrouwbaarheidscoëfficiënt (d.w.z. alfa of glb) gebruikt is en of het inderdaad gaat om de standaardmeetfout of misschien om de standaardschattingsfout. De genoemde waarden voor de standaardmeetfout zijn namelijk niet te repliceren met behulp van de getallen in de tabel. Op basis van een standaarddeviatie van 9.51 voor de totale toets E3 en een glb van .96, komt men uit op een standaardmeetfout van $9.51*\sqrt{(1-.96)} = 1.90$ en niet 1.98 zoals vermeld staat. Daarnaast wordt geen uitleg gegeven over het gebruik van betrouwbaarheidsintervallen voor de testgebruiker.
In Tabel 6.16 wordt gerapporteerd over scoreverschillen tussen jongens en meisjes. In alle gevallen scoren de jongens hoger, maar kijkend naar de effectgrootte heeft het verschil in de groepen 3 en 4 nauwelijks betekenis. In leerjaar 5 is er soms sprake van een klein effect en in de groepen 7 en 8 is er in het algemeen sprake van een klein effect; de totaalscores laten daar een Cohens *d* van circa 0.38 zien. Ook wat betreft de snelheid zijn de scoreverschillen tussen jongens en meisjes onderzocht, waarbij een ruime meerderheid van de effecten kleiner dan 0.20 is en soms veel kleiner.

Betrouwbaarheid
De betrouwbaarheid van de toetsen is, waar het de accuratesse betreft, op klassieke wijze bepaald aan de hand van Cronbachs alfa en *Greatest Lower Bound* (glb). De toetsen zijn beoordeeld als toetsen voor minder belangrijke beslissingen op individueel niveau. Alle toetsen behalen wat hun totaalscore betreft een coëfficiënt alfa van minstens .94 en een glb van minimaal .96, dit is ‘goed’ te noemen. In Tabel 5.2 worden tevens de betrouwbaarheidscoëfficiënten voor de onderdelen vermeld. Deze zijn hoger dan .80 (d.w.z. ‘goed’) met uitzondering van 'optellen' bij de toetsen M7 en E7, dat als ‘voldoende’ beoordeeld wordt. Het is overigens niet duidelijk waarom voor sommige toetsen en onderdelen alleen alfa wordt weergegeven en niet de glb.
In Tabel 5.3 staan de waarden voor de split-half betrouwbaarheid, die gebruikt is voor het berekenen van de betrouwbaarheid van de snelheid op de afzonderlijke onderdelen voor alle normeringsmomenten. De betrouwbaarheid is zeer hoog: de laagste coëfficiënt is .85 en de meerderheid is hoger dan .90.

Begripsvaliditeit
Hoofdstuk 6 van de Wetenschappelijke Verantwoording gaat over de begripsvaliditeit van de toetsen.
De dimensionaliteit van de toetsen is onderzocht door de correlaties te berekenen tussen de onderdelen en de totaalscore. Deze variëren van .81 tot .94 en ondanks het feit dat ze vanwege de overlap geflatteerd zijn (wat vooral bij de lagere leerjaren een rol speelt aangezien daar nog geen vermenigvuldigen en delen wordt gemeten), kunnen ze als een aanwijzing worden opgevat dat er van één onderliggende vaardigheid sprake is. Verder is gekeken naar de (voor attenuatie gecorrigeerde) correlaties tussen de onderdelen. Hierover werden verwachtingen geformuleerd op inhoudelijke gronden, bijvoorbeeld dat optellen het omgekeerde is van aftrekken en dat vermenigvuldigen een vorm van herhaald optellen is, zodat hypotheses konden worden opgesteld als "De correlatie tussen optellen en aftrekken ligt hoger dan de correlatie tussen optellen en vermenigvuldigen en die tussen optellen en delen". Het merendeel van de verwachtingen kon worden bevestigd en bovendien bleek het patroon over de verschillende afnamemomenten constant. Er waren echter ook een aantal onverwachte relaties of zelfs voor alle normgroepen omgekeerd aan de verwachting, zoals de verwachting dat de correlatie tussen ‘aftrekken’ en ‘vermenigvuldigen’ lager zou zijn dan de correlatie tussen ‘aftrekken’ en ‘delen’ terwijl de correlatie tussen ‘aftrekken’ en ‘vermenigvuldigen’ voor alle normgroepen hoger was. Het is jammer dat in de Verantwoording niet getracht wordt dergelijke onverwachte resultaten bijvoorbeeld vanuit de theorie te verklaren. Voor de snelheidsscores van de diverse onderdelen werden overeenkomstige verwachtingen geformuleerd. Zo werd verwacht dat de correlatie tussen optellen en aftrekken het hoogst is en die tussen optellen en delen het laagst. Bij de bewerkingen optellen en delen werden de verwachtingen consistent bevestigd, maar bij aftrekken werd ten opzichte van vermenigvuldigen en delen een ander patroon gevonden en bij vermenigvuldigen ten aanzien van optellen en aftrekken. Ook hier ontbreekt een verklaring voor deze onverwachte resultaten.
Wat de psychometrische kwaliteit van de items betreft blijkt dat de gemiddelde waarde van de itemrestcorrelaties het laagst is bij toets M7 (r_{ir} = .41) en het hoogst bij E4 (r_{ir} = .50). Daarbij lettend op de

laagste en hoogste r_{ir} –waarden, gaat het om voldoende tot goede waarden gelet op de in het COTAN Beoordelingssysteem genoemde richtlijnen, die bovendien de item-testcorrelatie betreffen, die gezien zijn aard hoger uitvalt (en niet lager zoals in de Verantwoording vermeld staat op pagina 40). De laagste individuele item-restcorrelatie die wordt gevonden, bedraagt .19 (bij toets E7), waarmee het strikt genomen net in de categorie 'onvoldoende' valt. Gezien het grote aantal toetsen en items doet dit echter geen afbreuk aan het totaalbeeld, temeer daar het een correlatie met de rest en niet met de test betreft. Over de *p*-waarden werd reeds opgemerkt dat hierbij niet werd gestreefd naar maximale spreiding, maar juist naar wat gemakkelijker items die echt kunnen toetsen of er sprake is van een geautomatiseerde vaardigheid. Tabel 6.9 geeft hierover nadere informatie. Bovendien wordt in grote lijnen de verwachting bevestigd dat de *p*-waarden van items die tussen twee toetsen overlappen het hoogst zijn in de toets voor het laatste meetmoment.
Of de factorstructuur invariant is voor sekse, is niet door middel van factoranalyses onderzocht. Wel werd *differential item functioning* (DIF) onderzocht met de Mantel-Haenszel statistiek. Het aantal items dat DIF vertoont (d.w.z. $z > 2.58$) varieert per toets van 0 (toetsen E3 en E4, beide bestaande uit 50 items) tot 13 (toets M7, bestaande uit 80 items). Bij ongeveer de helft van de items met DIF zijn de jongens in het voordeel en bij de andere helft de meisjes. De auteurs trekken hieruit de conclusie dat de invloed van DIF zeer beperkt is, hoewel dit niet is onderzocht door de verschillen tussen jongens en meisjes te berekenen mét en zónder weglating van de items in kwestie en door het effect op testniveau in kaart te brengen. Opvallend is dat bij optel-items vrijwel geen DIF voorkomt. Een nadere inhoudelijke analyse van de items is niet gedaan.
De convergente validiteit is onderzocht met de Cito LVS-toetsen 'Rekenen-Wiskunde 2.0' en 'Rekenen-Wiskunde 3.0'. Voor het onderzoeken van de discriminante validiteit zijn de Cito LVS-toetsen 'Spelling' en 'Begrijpend Lezen' gebruikt. Bij alle onderdelen van de toetsen 'Rekenen-Basisbewerkingen voor groep 3 tot en met 8' is in alle normgroepen te zien dat de correlaties met 'Rekenen-Wiskunde' hoger zijn dan die met 'Spelling' en 'Begrijpend Lezen'. Met correlaties van rond .55 met 'Rekenen-Wiskunde 3.0' zijn de resultaten, gezien hetzelfde inhoudsdomein, niet opvallend sterk. De auteurs geven hiervoor als oorzaak dat deze toets niet automatiseren als specifieke meetpretentie heeft, maar toch zijn de correlaties matig verklaarbaar en niet erg bevredigend. Het is bij het onderzoek naar de convergente validiteit jammer dat niet een echte soortgenoottest is gebruikt. Er zijn er in Nederland voldoende om dergelijk onderzoek mee te doen, bijvoorbeeld de TTA (Tempo Toets Automatiseren), die een nauwere definitie van automatiseren geeft. Wat betreft het onderzoek naar de discriminante validiteit zijn de correlaties met 'Spelling' en 'Begrijpend Lezen' weliswaar lager dan die met 'Rekenen-Wiskunde' (circa .35 voor de toetsen van de derde generatie), wat op zich aantoont dat de rekentoetsen een niet al te talig karakter hebben, maar ze zijn ook niet uitgesproken laag.
Al met al wijzen de gevonden resultaten in het algemeen in dezelfde richting en kunnen derhalve gezien worden als voldoende ondersteuning van de begripsvaliditeit.

Criteriumvaliditeit
Volgens de auteurs zijn de toetsen niet bedoeld voor voorspellend gebruik. Criteriumvaliditeit is daarom niet van toepassing en ook niet onderzocht.

FAIRNESS MATRIJS

Fairness aspecten	Fairness onderzoek uitgevoerd?	Aangegeven reden geen onderzoek	Testmateriaal		Handleiding	Normen	Betrouw-baarheid	Begripsvaliditeit		Criterium-validiteit
Kenmerk			Instructie	Items	Info over verschillen tussen groepen en interpretatie testscore	Representa-tiviteit voor subgroepen		Item bias onderzoek	Factor-structuur	
Geslacht	ja	n.v.t.			1			2		
Herkomst (etnisch-cultureel)	nee	ja								
Leeftijd	nee	nee								

Toelichting:

1	Gemiddelde scores voor jongens en meisjes zijn onderzocht met *t*-tests.
2	Items zijn onderzocht op DIF voor sekse op basis van Mantel-Haenszel methode.

Algemene opmerkingen:
De variabele etniciteit is nergens in onderzoek meegenomen “omdat er geen eenduidige referentiegegevens voor de populatie bekend zijn” (p. 21 van de Weten-schappelijke Verantwoording).