De globale en lokale betrouwbaarheid van de screeningstoets van de Bareka Profieltoets Rekenen bij kinderen in groep 3

A.H. van Hoogmoed, W.E. Kupers, A.V. Sijp, J. Vloo, W.H. Hofstetter, en M.C.S. Paap

Samenvatting

Vlot en accuraat oplossen van kale sommen (oftewel technisch rekenen) ligt aan de basis van succes in rekenen en wiskunde. Om kinderen adequaat te kunnen helpen bij eventuele achterstanden in het rekenen, is het van belang om het niveau van leerlingen goed in kaart te brengen. De Bareka Profieltoetsen zijn ontwikkeld om zowel het vlot als accuraat oplossen van kale sommen in kaart te brengen middels respectievelijk de automatiseringstoets en screeningstoets. In deze studie is de betrouwbaarheid van de screeningstoets voor leerlingen in groep 3 onderzocht, voor de plus- en minsommen onder de 10, tussen de 10 en 20 zonder doorbreking van het tiental, en tot 20 met doorbreking van het tiental. De resultaten ondersteunen de opbouw van de drempels. De globale betrouwbaarheid van de screeningstoets bleek goed voor groep 3. De lokale betrouwbaarheid laat zien dat de screeningstoets vooral goed discrimineert bij kinderen die benedengemiddeld scoren. Dit is passend bij het doel van de toets, namelijk inzicht geven in welke leerlingen hulp nodig hebben op welk niveau. De toets is dus geschikt om te indiceren welke leerlingen hulp nodig hebben bij welk type sommen, om zo een optimale ontwikkeling te kunnen ondersteunen op het gebied van rekenen en wiskunde.

Kernwoorden: rekenen, Bareka profieltoets, betrouwbaarheid, item respons theorie, testkwaliteit

1 Inleiding

Rekenen is van groot belang voor individuen, omdat lage rekenvaardigheden gerelateerd zijn aan een verhoogde kans op vroegtijdig schoolverlaten, een grotere kans op werkloosheid en een grotere kans op depressieve gevoelens (Bynner & Parsons, 2001; Parsons & Bynner, 2005). Uit internationale onderzoeken blijkt echter dat het niveau van rekenvaardigheden in Nederland daalt (Mullis, Martin, Foy, & Hooper, 2016; OECD, 2016). Nationaal onderzoek laat zien dat achterstanden in veel gevallen al vroeg na de start van het formele rekenonderwijs ontstaan. Zo blijkt dat de 80% beheersingsnorm voor sommen tot 100 aan het einde van groep 4, die is vastgesteld door de onderwijsinspectie en de PO-raad (Gelderblom, 2009), door veel leerlingen niet gehaald wordt (Bandstra, Danhof, Faber, Minnaert & Ruijssenaars, 2013). Deze leerlingen hebben meer onderwijstijd nodig om de sommen zonder fouten te maken en te automatiseren. Door het cumulatieve karakter van rekenen zorgt dit voor problemen in het verdere rekenonderwijs, omdat er in dat geval wordt voortgeborduurd op een niet stevige basis (Bandstra et al., 2013; Meijerink, 2008).

Bij het rekenen op school wordt voortgebouwd op eerder besproken begrippen en rekenprocedures, waarbij ervan uit wordt gegaan dat de leerlingen die beheersen en geautomatiseerd hebben. Als dat niet het geval is, mist de leerling de aansluiting met de leerstof en zal de achterstand steeds groter worden (Noteboom, 2008). Vanwege het cumulatieve karakter van rekenen is het belangrijk dat achterstanden in het rekenen in de eerste leerjaren van de basisschool opgemerkt worden (Clarke, Doabler, Nelson & Shanley, 2015; Gelderblom, 2009). Bovendien weten we uit eerder onderzoek dat leerlingen met langdurige achterstanden en/of specifieke leerproblemen hun motivatie voor rekenen snel kunnen verliezen, waardoor er een vicieuze cirkel kan ontstaan (Fulk, Brigham, & Lohman, 1998; Lackaye, Margalit, Ziv, & Ziman, 2006; Woolf

Tabel 1.
Rekendrempels (met toestemming overgenomen uit Danhof, Bandstra & Hofstetter, 2014)

Drempel	Subdrempel	Somtype	Voorbeeld
Drempel 1: Optellen en aftrekken tot 10	Drempel 1a: Drempel 1b: Drempel 1c:	Optellen tot 10 Aftrekken tot 10 Splitsen tot en met 10	4 + 5 8 - 6 10 = 3 en …
Drempel 2: Getalbegrip tot 100	Drempel 2a: Drempel 2b: Drempel 2c: Drempel 2d:	Sprong naar het volgende tiental Sprong naar het vorige tiental Sprong van 10 vooruit Sprong van 10 terug	39 + 1 40 - 1 62 + 10 57 - 10
Drempel 3: Optellen en aftrekken tot 20	Drempel 3a: Drempel 3b:	Optellen met overschrijding van het tiental Aftrekken met overschrijding van het tiental	7 + 8 16 – 7
Drempel 4: Optellen en aftrekken tot 100	Drempel 4a: Drempel 4b: Drempel 4c: Drempel 4d:	Optellen met tientallen Aftrekken met tientallen Optellen met overschrijding van het tiental Aftrekken met overschrijding van het tiental	64 + 20 78 - 30 47 + 8 55 - 7
Drempel 5: Tafels van vermenigvuldiging	Drempel 5a: Drempel 5b:	Eenvoudige tafels (1, 2, 3, 4, 5 en 10) Moeilijke tafels (6, 7, 8 en 9)	5 x 4 8 x 6

et al., 2010). Om frustratie en verlies van motivatie te voorkomen is het cruciaal dat de instructie van de leerkracht en het niveau van de lesstof aansluiten bij het niveau van de leerling (Chaiklin, 2003; Gelderblom, 2009; Noteboom, 2008; Stone, 1998). Aandacht voor technisch rekenen, oftewel het maken van kale sommen, is hierbij van belang om eventuele stagnatie in het latere rekenleerproces te beperken (Ruijssenaars, Danhof, Hofstetter, & Minnaert, 2017).

De Bareka Profieltoetsen zijn ontwikkeld om inzicht te geven in welke vaardigheden kinderen wel en niet beheersen op het gebied van technisch rekenen. De toetsen zijn gebaseerd op het drempelmodel (Danhof, Bandstra, Mushati-Hamadani, Minnaert, & Ruijssenaars, 2008), waarbij uit wordt gegaan van een aantal basisautomatismen en procedurele kennis die de basis vormen voor hoofdrekenen. Het drempelmodel bestaat uit vijf drempels (mijlpalen) in de ontwikkeling van het hoofdrekenen (zie Tabel 1). Hierbij wordt uitgegaan van het cumulatieve karakter van rekenen, waarbij de beheersing (dat wil zeggen vlot kunnen oplossen) van de lagere drempels een voorwaarde is voor het beheersen van hogere drempels. Zo is het bijvoorbeeld van belang om het antwoord op de som 6 + 7 te weten om de som 36 + 7 vlot uit te kunnen rekenen.

De Bareka profieltoetsen bestaan uit een screeningstoets en een automatiseringstoets op basis van de bovengenoemde drempels. De screeningstoets meet de procedurele kennis van een leerling, dat wil zeggen het vermogen om een bepaalde som op te lossen, zonder dat er sprake is van tijdsdruk. Hierbij worden per somtype vier opgaven aangeboden aan de leerling. De automatiseringstoets meet de mate waarin een leerling een bepaald type som vlot kan oproepen uit het geheugen (Danhof et al., 2008). Bij de automatiseringstoets moet een leerling daarom zoveel mogelijk sommen van een bepaalde subdrempel maken binnen twee minuten.

De automatiseringstoets bevat alle subdrempels uit het drempelmodel (zie tabel 1). De screeningstoets bevat naast de subdrempels nog extra somtypen, omdat professionals uit de onderwijspraktijk graag een gedetailleerder beeld wilden van de somtypen die een leerling wel en niet beheerst. Zo omvat de automatiseringstoets van drempel 3 alleen sommen met overschrijding van het tiental met uitkomsten tussen de 10 en 20, maar bevat de screeningstoets daarnaast ook sommen tussen de 10 en 20 zonder overschrijding van het tiental. Daarnaast bevat de screeningstoets representatieve somtypen voorbij drempel 5 die later in het curriculum voorkomen, namelijk sommen met grote getallen,

breuken en procenten voor het niveau tot en met groep 7 van het basisonderwijs. Zo is de screeningstoets gestoeld op het drempelmodel, maar wijkt hier iets van af om een zo gedetailleerd mogelijke weergave te geven van de vaardigheden van leerlingen.

Bijna 12% van de reguliere Nederlandse basisscholen maakt gebruik van de Bareka, een percentage dat nog groeit. Daarnaast wordt de Bareka ook gebruikt in het speciaal (basis)onderwijs. Vaak wordt op basis van de Cito-toets rekenen-wiskunde een risicogroep in beeld gebracht. Bij deze groep wordt de screeningstoets afgenomen om te kijken welke somtypen wel en niet beheerst worden en of er hiaten in (procedurele) kennis aanwezig zijn. Bij leerlingen uit de risicogroep die op basis van de screeningstoets wel de benodigde somtypen lijken te beheersen (en in die zin geen extra instructie nodig hebben), wordt de automatiseringstoets afgenomen om na te gaan of de kennis ook reeds geautomatiseerd is, of dat hiermee nog geoefend moet worden.

Ondanks het frequente gebruik van de profieltoetsen in de onderwijspraktijk, zijn er geen gegevens bekend over de betrouwbaarheid van de toetsen. Het doel van deze studie was om de betrouwbaarheid van een deel van de screeningstoets van de Bareka in kaart te brengen. Omdat achterstanden in het technisch rekenen zo snel mogelijk gesignaleerd dienen te worden, is ervoor gekozen om onderzoek te doen bij leerlingen uit groep 3. Eind groep 3 wordt verwacht dat leerlingen de sommen tot 10 geautomatiseerd hebben en de sommen tot 20 kunnen oplossen (Noteboom, Aartsen & Lit, 2017). Daarom zijn in dit onderzoek de somtypen (binnen de Bareka aangegeven met IT) meegenomen die vallen onder drempel 1 en 3 (zie Tabel 1). Naast de globale betrouwbaarheid is ook de lokale betrouwbaarheid van de toets onderzocht, omdat het van belang is dat deze vooral betrouwbaar is voor leerlingen die beneden-gemiddeld scoren, zodat zij op tijd ondersteund kunnen worden. De betrouwbaarheid is onderzocht op basis van de item-respons theorie (IRT). Deze methode van onderzoek geeft naast de betrouwbaarheid ook inzicht in de relatieve moeilijkheid van de items. Hiermee kan worden gekeken of de verschillende onderzochte somtypen in groep 3 oplopen in moeilijkheid, zoals te verwachten volgens het drempelmodel. Daarnaast wordt gekeken of de verschillende items op basis van hun moeilijkheid binnen het somtype passen, wat het drempelmodel zou ondersteunen.

2 Methode

2.1 Participanten

De doelpopulatie bestaat uit Nederlandse leerlingen uit groep 3 van het regulier basisonderwijs. De participanten voor het onderzoek zijn geworven middels een gelegenheidssteekproef. Vier leerlingen zijn geëxcludeerd, drie omdat zij niet de gehele toets hadden gemaakt, en één vanwege hulp van de leerkracht. De uiteindelijke steekproef bestond uit 600 leerlingen ($M_{leeftijd}$ = 6 jaar en 11 maanden, SD = 5 maanden, range 6 jaar en 1 maand – 8 jaar en 11 maanden), waarvan 300 jongens en 300 meisjes. De participanten waren afkomstig van 22 reguliere basisscholen en 32 verschillende klassen. Van 24 scholen nam 1 groep 3 deel, van 7 scholen 2 groepen 3 en van 1 school 4 groepen 3.

2.2 Meetinstrument

De screeningstoets van Bareka bestaat uit vier sommen per somtype. Er zijn in het onderzoek zes somtypen afgenomen die passen bij het niveau van leerlingen in groep 3. Dit zijn IT1 (bv. 8+1), IT2 (bv. 14+3), IT3 (bv. 4+8), IT11 (bv. 8-2), IT12 (bv. 15-3) en IT13 (bv. 15-6). Hierbij staat IT in de screeningstoets voor item, maar verwijst inhoudelijk naar somtype.

Er is gebruik gemaakt van de schriftelijke (niet-digitale) versie van de screeningstoets. In het kader van een ander onderzoek, zijn naast de oorspronkelijke items een aantal extra items afgenomen. Deze laten we in dit artikel echter buiten beschouwing.

2.3 Procedure

Scholen en leerkrachten zijn door twee auteurs van dit artikel (AS en JV) telefonisch, per mail of via social media benaderd. Na instemming met deelname zijn zij voorzien van meer informatie en werd er een brief

meegegeven voor de ouders/verzorgers. Deze konden passief toestemming geven voor deelname van hun kind. Voor het onderzoek is toestemming verleend door de Ethische Commissie Pedagogische wetenschappen en Onderwijskunde te Groningen.

Vervolgens zijn scholen voorzien van de schriftelijke versie van de toetsen en een afnameprotocol. Op deze wijze konden leerkrachten de toets zelf afnemen. Op één school is de toets afgenomen door twee auteurs van dit artikel (AS en JV). De dataverzameling vond plaats in maart en april 2018. Het afnameprotocol voorzag in een korte instructie over de verschillende opgaven, waarbij de leerlingen geïnstrueerd werden om de opgaven te maken zonder gebruik van rekenmachine of andere hulpmiddelen. Leerlingen kregen voor zowel de plus- als de minsommen vijftien minuten de tijd. De tijdslimiet moest voorkomen dat leerlingen met sommen aan de slag gingen die te moeilijk waren.

2.4 Statistische analyse

Er worden twee typen betrouwbaarheid geschat: globale en lokale betrouwbaarheid, beiden met behulp van het Rasch model, ook wel het 1-parameter logistisch model (1PL) genoemd. Het 1PL behoort tot de familie van item respons theorie (IRT) modellen, welke als voordeel hebben dat de betrouwbaarheid kan worden onderzocht conditioneel op de geschatte positie op de latente vaardigheid (in ons geval rekenvaardigheid). De itemresponsfunctie, die uitdrukt hoe groot de kans is dat een bepaald item correct wordt beantwoord als functie van de vaardigheid, is voor het 1PL als volgt gedefinieerd:

$$f_i(\theta) = P(X_i = 1|\theta) = \frac{\exp(\theta - \beta_i)}{1 + \exp(\theta - \beta_i)}$$

waarbij θ de latente vaardigheidsscore aanduidt, β_i de score (0 of 1) op item i, en β_i de moeilijkheidsparameter die hoort bij item i. geeft de hoeveelheid vaardigheid aan die nodig is om een kans van 0.5 te hebben om een item juist te beantwoorden (Verhelst, 1993). θ heeft 0 als gemiddelde, de spreiding wordt geschat. Bovengenoemd model kan ook worden uitgebreid naar een multilevel Rasch model, waar rekening kan worden gehouden met de geneste structuur (leerlingen binnen klassen). Dit model zal ook worden geschat, en de uitkomsten (intraklassecorrelatie en model fit) worden vergeleken met het "gewone" Rasch model. Indien de uitkomsten dicht bij elkaar liggen, worden enkel die van het "gewone" Rasch model gepresenteerd. De modellen worden vergeleken aan de hand van de BIC fit maat (Bayesiaans Informatie Criterium). Een voordeel van de BIC is dat hij weinig gevoelig is voor steekproefgrootte. Uit verschillende onderzoeken is gebleken dat dit criterium de meest accurate resultaten oplevert bij het vergelijken van IRT modellen (Cohen & Cho, 2016). Om na te gaan of het Rasch model het meest passende model is, wordt ook het 2 parameter logistisch (2PL) model geschat, en wordt de fit vergeleken met die van het Rasch model.

De betrouwbaarheid kan worden gezien als de ratio van de betrouwbare variantie ten opzichte van de totale variantie. Binnen het kader van de IRT kan met de volgende formule worden gewerkt om een schatting te krijgen van de globale betrouwbaarheid:

$$\rho = \frac{var[E(\theta|y)]}{var(\theta)}$$

waar de teller de posterior variantie van de geschatte persoonsparameter weergeeft (y is het geobserveerde responsiepatroon), en de noemer de totale populatie variantie van de latente variabele (Bechger, Maris, Verstalen, & Béguin, 2003). De formule ligt ten grondslag aan de berekeningen gemaakt in dit artikel. Om rekening te houden met de geneste structuur, is een weging toegepast. De teller is berekend door eerst de de posterior variantie van de geschatte persoonsparameter per groep te bepalen, en vervolgens een gewogen som te nemen (weging o.b.v. groepsgrootte t.o.v. totale groep).

Binnen het kader van IRT wordt meetonnauwkeurigheid ook wel uitgedrukt in Informatie (bijvoorbeeld Fisher Informatie): een hogere Informatiewaarde duidt op een hogere meetnauwkeurigheid en dus lagere meetfout. De Informatie die een test oplevert ten aanzien van de positie van een persoon op de schaal is niet per definitie constant over die schaal: deze Informatie is direct gerelateerd

aan de psychometrische eigenschappen van de items in de test. Aangezien de gekwadrateerde standaardmeetfout voor persoon *p* gelijk is aan de reciprook van de testinformatie , $I(\theta_p)$ kan de lokale betrouwbaarheid als volgt worden geschat:

$$r(\theta_p) = 1 - \frac{\mathrm{SE}(\theta_p)^2}{VAR(\theta_p)} = 1 - \frac{VAR(\theta_p)}{\mathrm{I}(\theta_p)}$$

Naast het evalueren van de betrouwbaarheid, is de verwachting dat de parameters oplopen voor de drie typen plussommen en de drie typen minsommen onderzocht. Verder is aan de hand van een Wright Map gekeken of de moeilijkheid van de toets goed aansluit op de verdeling van de vaardigheidsscores. In een Wright Map worden de geschatte vaardigheidsscores afgezet tegen de geschatte moeilijkheidsparameters.

De globale betrouwbaarheid is geschat met behulp van het software programma MIRT (Glas, 2010). De overige analyses zijn uitgevoerd in het software programma R versie 3.4.3 (R Development Core Team, 2012), met behulp van de volgende twee pakketten: mirt versie 1.28 (Chalmers, 2012) en WrightMap versie 1.2.1 (Torres & Freund, 2014). Het R pakket mirt maakt gebruik van de *marginal maximum likelihood* (MML) schattingsmethode. Een betrouwbaarheid van .80 of hoger wordt gezien als goed voor minder belangrijke beslissingen op individueel niveau en een betrouwbaarheid tussen de .70 en .80 als voldoende (Evers, Lucassen, Meijer, & Sijtsma, 2010).

3. Resultaten

3.1 Prestaties op de screeningstoets

Allereerst is inzichtelijk gemaakt hoe de prestaties van de leerlingen zijn op de verschillende somtypes. De resultaten zijn weergegeven in Tabel 2. Hieruit blijkt dat de drempels inderdaad oplopend zijn qua moeilijkheid en dat de minsommen moeilijker zijn dan de plussommen.

3.2 Vergelijking van IRT modellen en globale betrouwbaarheid

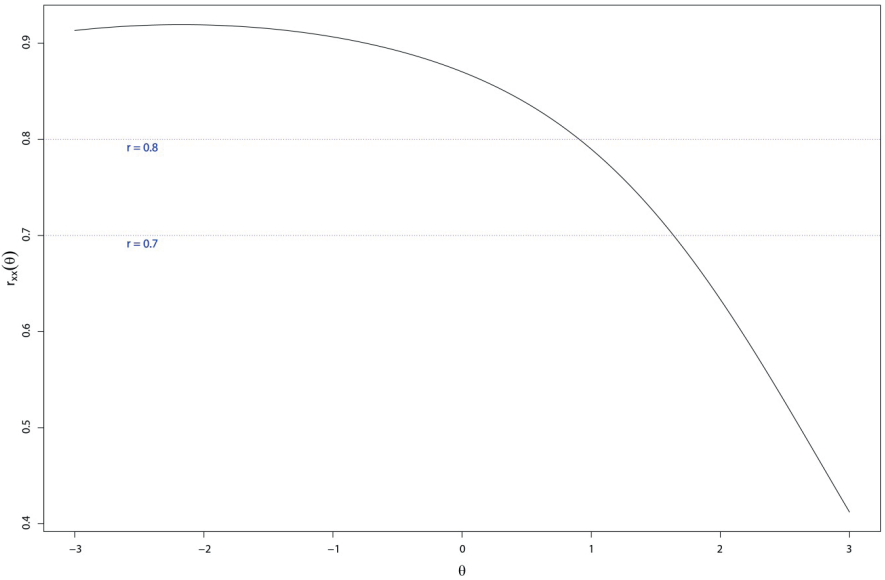
De BIC waarden bedroegen 10859 voor het Rasch model, 10864 voor het multilevel Rasch model, en 10930 voor het 2PL model. De voorkeur gaat uit naar het model met de laagste BIC waarde, in dit geval het "gewone" Rasch model. Ook uit de aanvullende analyses bleek dat de verschillen tussen de twee Rasch modellen klein waren. De intraklassecorrelatie in het multilevel Rasch model bedroeg .03; dit betekent dat het deel van de variantie in latente scores dat is toe te wijzen aan het klasniveau erg laag is. Het visueel inspecteren van de item parameters voor de twee Rasch modellen lieten verwaarloosbaar kleine verschillen zien. De globale betrouwbaarheid bedroeg .83 voor het "gewone" Rasch model en .82 voor het multilevel Rasch model; dus ook de impact op de betrouwbaarheid was klein. Beide waarden zijn conform de richtlijnen ($r \geq .80$) voor 'tests voor minder belangrijke beslissingen op individueel niveau' van de COTAN geclassificeerd als goed (Evers et al., 2010). Gezien de minieme verschillen en de superieure fit van het "gewone" Rasch model, zullen we hieronder de resultaten van het gewone Rasch model rapporteren.

3.2.1 Lokale betrouwbaarheid.

In Figuur 1 is zichtbaar dat de lokale betrouwbaarheid het hoogst is voor de theta-waarden -3.0 tot 1.0 ($r > 0.8$). Vanaf de theta-waarde 1.0 daalt de lokale betrouwbaarheid, maar tot de theta-waarde 1.5 is de lokale betrouwbaarheid voldoende ($r \geq 0.7$). Dit betekent dat de toets het beste onderscheid kan maken tussen laag scorende leerlingen en leerlingen die gemiddeld scoren. Dit komt overeen met het doel van de toets, namelijk het screenen van leerlingen die bepaalde somtypen nog niet beheersen. Uit Figuur 1 blijkt dat de toets niet betrouwbaar is voor het discrimineren tussen gemiddelde en hoger presterende leerlingen. Vanaf de latente trekwaarde 1.5 is de betrouwbaarheid namelijk kleiner dan 0.7 en volgens de COTAN richtlijnen voor 'tests voor minder belangrijke beslissingen op individueel niveau' onvoldoende (Evers et al., 2010). Dit wijst erop dat de toets meer dan voldoende betrouwbaar is voor de zwakke tot gemiddelde leerlingen en niet betrouwbaar en geschikt voor leerlingen die aanzienlijk hoger dan gemiddeld scoren.

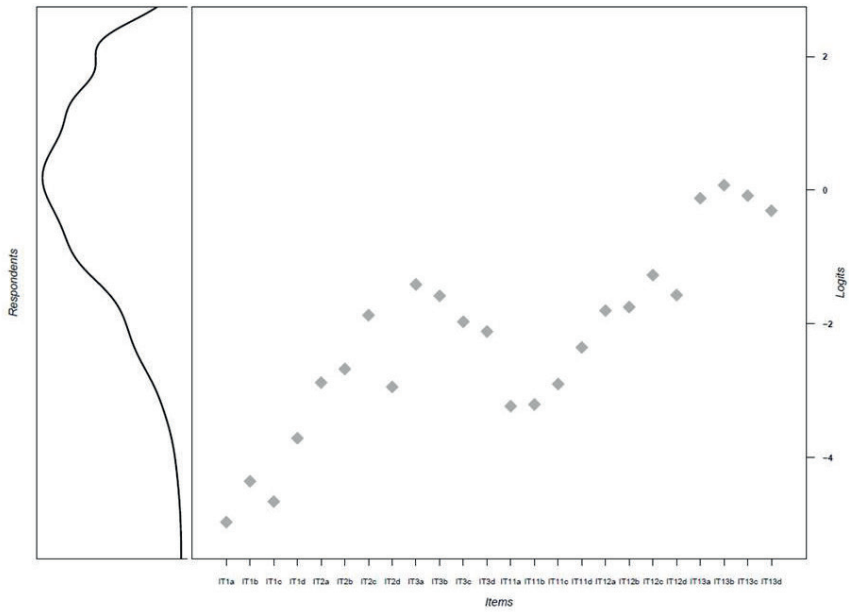
Tabel 2.
Percentage leerlingen per aantal goed op elk somtype (per drempel optellend tot 100 %)

Somtype	Aantal opgaven goed				
	0	1	2	3	4
IT 1: optellen onder de 10	0.3	0.7	1.8	7.8	89.3
IT 2: Optellen tussen 10 en 20	2.0	3.8	7.7	21.7	64.8
IT 3: Optellen over het tiental	7.0	5.2	12.5	22.5	52.8
IT 11: aftrekken onder de 10	1.3	3.5	5.3	18.5	71.3
IT 12: aftrekken tussen 10 en 20	7.8	7.8	12.5	19.7	52.2
IT 13: aftrekken over het tiental	27.0	12.3	12.7	21.8	26.2



Figuur 1. Lokale betrouwbaarheid gebaseerd op het Rasch model. De geschatte latente trek waarde (θ) op de x-as en de lokale betrouwbaarheid (r_{xx}) op de y-as.

3.2.1.1 Moeilijkheid van de zes somtypen.
In Tabel 3 is de gemiddelde b-waarde, oftewel de moeilijkheid, per somtype weergegeven. Voor de screeningstoets geldt dat de moeilijkheid van de somtypen, zowel voor de plussommen als de minsommen, oploopt. De plussommen van somtype IT1 zijn makkelijker dan de sommen van somtype IT2 en IT3 en de sommen van IT2 zijn makkelijker dan de sommen van IT3. Ook voor de minsommen geldt dat de minsommen van IT11 makkelijker zijn dan die van IT12 en IT13 en die van IT12 makkelijker dan die van IT13. Dit is ook zichtbaar in de Wright map die is weergegeven in Figuur 2, waarbij de eerste helft van de weergegeven punten de plussommen vormen en de tweede helft de minsommen. Uit Figuur 2 blijkt dat, overeenkomend met de verwachting, de vier items van somtype IT2 moeilijker zijn dan de vier items van somtype IT1. Opvallend aan somtype IT2 is dat item IT2c (de som 12 + 7) beduidend moeilijker is dan de andere drie items van somtype IT2. IT2c blijkt zelfs moeilijker te zijn dan IT3c en IT3d van de screeningstoets. Voor de overige items van IT2 geldt dat de moeilijkheid hiervan, zoals verwacht, lager ligt dan de moeilijkheid van de items van somtype IT3. Voor de minsommen geldt dat er een oplopende lijn te zien is in de moeilijkheid van de items, waarbij alle items van somtype IT11 makkelijker zijn dan de items van somtype IT12 en alle items van somtype IT12 makkelijker zijn dan de items van som-



Figuur 2. Wright map. Verdeling van scores op de latente trek links, moeilijkheid van sommen rechts (met op de x-as het somnummer en op de y-as de theta-waarde/beta-waarde).

Tabel 3. *Gemiddelde b-waarden (moeilijkheid) per somtype*

Somtype		Gemiddelde b-waarden
IT1:	Plussommen tot tien	-4.423
IT2:	Plussommen over tien tot twintig	-2.593
IT3:	Plussommen over tien tot twintig met doorbreking van het tiental	-1.701
IT11:	Minsommen tot tien	-2.923
IT12:	Minsommen over tien tot twintig	-1.598
IT13:	Minsommen over tien tot twintig met doorbreking van het tiental	-0.109

type IT13. Verder blijkt uit Figuur 2 dat de meeste leerlingen de moeilijkheid van de sommen aankunnen, omdat het vaardigheidsniveau van de leerlingen over het algemeen hoger ligt dan het vaardigheidsniveau dat nodig is om de sommen te beantwoorden.

De moeilijkheidsparameters van de items van de screeningstoets en de betrouwbaarheidsintervallen rondom de parameters zijn weergegeven in Tabel 4. Uit de betrouwbaarheidsintervallen blijkt dat de schattingsfout van de parameters acceptabel is. De b-waarden en de theta-waarden liggen op dezelfde schaal met een gemiddelde van 0. De theta-schaal van de participanten loopt van -4.542 tot 2.186 (M = 0, SD = 1.46). Een aantal items hebben een b-waarde die lager ligt dan de laagste theta-waarde, namelijk de items IT1a en IT1c. Dit betekent dat deze items weinig informatie opleveren. Daarnaast zijn er items met een b-waarde kleiner dan -2.5, die daardoor erg ver van het gemiddelde liggen. Alle items van somtype IT1 hebben een extreem lage b-waarde (zie Tabel 4 en Figuur 2). Voor somtype IT2 geldt dat de items IT2a, IT2b en IT2d een extreem lage b-waarde hebben. Ook de items IT11a, IT11b en IT11c hebben een extreem lage b-waarde (d.w.z. lager dan -2.5). Voor de somtypen IT3, IT12 en IT13 geldt dat ze geen extreme b-waarden bevatten. Verder is opvallend dat de b-waarden van de screeningstoets nauwelijks boven nul uitkomen, wat betekent dat de items over het algemeen relatief gemakkelijk zijn voor deze groep leerlingen.

Tabel 4. *Itemparameters van de plus- en minsommen*

Plussommen			Minsommen		
Item	b	95%-BI	Item	b	95%-BI
IT1			IT11		
a: 8+1	-4.967	-5.568 -4.366	a: 8-2	-3.232	-3.582 -2.883
b: 2+5	-4.356	-4.839 -3.872	b: 9-8	-3.207	-3.554 -2.860
c: 5+3	-4.659	-5.196 -4.123	c: 7-5	-2.901	-3.224 -2.578
d: 4+6	-3.711	-4.107 -3.315	d: 9-6	-2.353	-2.643 -2.063
IT2			IT12		
a: 14+3	-2.879	-3.201 -2.557	a: 15-3	-1.803	-2.069 -1.536
b: 4+15	-2.676	-2.984 -2.367	b: 16-5	-1.749	-2.014 -1.485
c: 12+7	-1.871	-2.139 -1.602	c: 19-6	-1.271	-1.521 -1.021
d: 2+16	-2.945	-3.271 -2.618	d: 18-5	-1.570	-1.828 -1.311
IT3			IT13		
a: 8+8	-1.411	-1.665 -1.158	a: 15-6	-0.121	-0.355 0.113
b: 9+6	-1.582	-1.841 -1.324	b: 16-9	0.074	-0.159 0.308
c: 5+7	-1.969	-2.241 -1.696	c: 13-8	-0.082	-0.316 0.152
d: 4+8	-2.116	-2.394 -1.837	d: 14-6	-0.307	-0.542 -0.072

Noot. b = moeilijkheidsparameter; gebaseerd op Rasch model. 95%-BI staat voor 95%-betrouwbaarheidsinterval

Discussie

Het hoofddoel van dit onderzoek was het onderzoeken van de betrouwbaarheid van de screeningstoets van de Bareka profieltoetsen bij leerlingen in groep 3. De resultaten laten zien dat de globale betrouwbaarheid van de toets goed is voor een test voor minder belangrijke beslissingen op individueel niveau (Evers et al., 2010).

Wat betreft de lokale betrouwbaarheid blijkt dat de toets voor leerlingen die laag tot gemiddeld scoren goed discrimineert tussen leerlingen. De toets geeft dus goed inzicht in verschillen in niveaus tussen deze leerlingen. Voor leerlingen die bovengemiddeld scoren discrimineert de toets echter minder goed. Dit sluit goed aan bij het doel van de screeningstoets, namelijk in kaart brengen welke leerlingen hulp nodig hebben. Voor dit doel is het belangrijk om betrouwbaar te kunnen meten bij leerlingen die benedengemiddeld scoren, maar niet zozeer om betrouwbaar te kunnen meten bij leerlingen die bovengemiddeld scoren.

De profieltoetsen van Bareka, waaronder de screeningstoets, zijn gebaseerd op het drempelmodel, waarin gesteld wordt dat voor het maken van sommen binnen een bepaalde drempel kennis nodig is uit de onderliggende drempels (Danhof, Banstra, & Hofstetter, 2015). Op basis hiervan wordt verwacht dat de drempels een oplopende moeilijkheidsgraad hebben. Hier is onderzocht of de verschillende somtypen inderdaad oplopend zijn in moeilijkheid. Zowel bij de plus- als minsommen blijkt dat sommen onder de 10 het gemakkelijkst zijn. Sommen tussen de 10 en 20 zonder overschrijding van het tiental zijn moeilijker. Sommen tot 20 met overschrijding van het tiental zijn de moeilijkste van de afgenomen somtypen. Dit laat zien dat de somtypen en drempels inderdaad oplopen in moeilijkheidsgraad, passend bij het drempelmodel.

Ook binnen de somtypen is er iets te zeggen over de relatieve moeilijkheid van de items. Zo valt bijvoorbeeld op dat de som 12 + 7 binnen somtype 2 relatief moeilijk is. Bij deze som is de addend (het getal dat erbij moet worden geteld) relatief groot. Dat dit meer fouten oplevert dan sommen waarbij een klein getal moet worden opgeteld (14 + 3, 4 + 15, 2 + 16) duidt erop dat in ieder geval

een deel van de leerlingen waarschijnlijk nog tellend optelt. Hetzelfde wordt gezien bij de plussommen over het tiental en de minsommen boven de 10 (IT12 en IT13), waarbij de sommen moeilijker zijn naarmate het getal dat moet worden afgetrokken groter is. Dit geeft aan dat het informatief is om in de toets sommen met zowel kleine als grote addends te hebben. Aangezien er weinig verschil is in moeilijkheid tussen de sommen die beginnen met het grote getal (14 + 3) en sommen die beginnen met het kleine getal (4 + 15 en 2 + 16) lijken leerlingen weinig moeite te hebben met het omkeren van de som.

De resultaten laten zien dat de screeningstoets van de Bareka profieltoetsen betrouwbaar is voor leerlingen in groep 3, vooral bij leerlingen die ondergemiddeld scoren. Dit betekent dat deze toets goed bruikbaar is om leerlingen te identificeren die moeite hebben met een bepaald type sommen, waarna instructie voor deze leerlingen kan worden aangepast en/of het oefenen geïntensiveerd kan worden. Op deze manier kan de Bareka screeningstoets bijdragen aan passend rekenonderwijs voor leerlingen die niet voldoende hebben aan de reguliere instructie in de klas.

Dit onderzoek is het eerste onderzoek dat de betrouwbaarheid van een deel van de veel gebruikte profieltoetsen Bareka aantoont. De betrouwbaarheid in dit onderzoek is gebaseerd op het 'gewone' Rasch model waar geen rekening werd gehouden met de geneste structuur. Uit de analyses bleek de fit (BIC) van het Rasch model beter te zijn dan de fit van het multilevel Rasch model. Dit geeft aan dat de itemparameters vergelijkbaar zijn over de verschillende klassen waar de leerlingen in zitten.

Een tekortkoming van dit onderzoek is dat er gebruik is gemaakt van extra items aan het eind van elk somtype. Hoewel verwacht wordt dat dit weinig invloed heeft op de betrouwbaarheid, zou vervolgonderzoek met de bestaande versie wenselijk zijn om dit uit te sluiten. Daarnaast wordt om het volledig rekenprofiel van een leerling in kaart te kunnen brengen niet alleen gebruik gemaakt van de screeningstoets, maar ook van de automatiseringstoets. Ook wordt de toets niet alleen in groep 3 gebruikt, maar ook in hogere groepen op de basisschool. Toekomstig onderzoek naar de betrouwbaarheid van de moeilijkere items in de screeningstoets voor oudere leerlingen en onderzoek naar de betrouwbaarheid van de automatiseringstoets is dan ook gewenst.

Literatuur

Bandstra, P., Danhof, W., Faber, S., Minnaert, A., & Ruijssenaars, W. (2013). *Rapport Rekenproject: Leerbaarheid van hoofdrekenen*. Geraadpleegd van http://www.steunpunttaalenrekenenvo.nl/sites/default/files/Rapport%20leerbaarheid%20van%20hoofdrekenen.pdf

Bechger, T., Maris, G., Verstralen, H., & Béguin, A. (2003). Using classical test theory in combination with item response theory. *Applied Psychological Measurement, 27*(5), 319-334.

Bynner, J., & Parsons, S. (2001). Qualifications, basic skills and accelerating social exclusion. *Journal of Education and Work*, *14*(3), 279-291.

Chaiklin, S. (2003). The zone of proximal development in Vygotsky's analysis of learning and instruction. In A. Kozulin, B. Gindis, V. S. Ageyev, & S. M. Miller (Eds.), *Vygotsky's educational theory in cultural context* (pp. 39–65). Cambridge, MA: Cambridge University Press.

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software, 48*(6), 1-29.

Clarke, B., Doabler, C. T., Nelson, N. J., & Shanley, C. (2015). Effective Instructional Strategies for Kindergarten and First-Grade Students at Risk in Mathematics. *Intervention in School and Clinic*, *50*(5), 257-265. doi:10.1177/1053451214560888

Cohen, A. S., & Cho, S.-J. (2016). Information criteria. In W. J. van der Linden (Ed.), *Handbook of item response theory, models, statistical tools, and applications* (Vol. 2, pp. 363-378). Boca Raton, FL: Chapman & Hall/CRC Press.

Danhof, W., Bandstra, P., & Hofstetter, W. (2015). Rekendrempels nemen. *Volgens Bartjens, 34*(3), 4-7.

Danhof, W., Bandstra, P., Mushati-Hamadani, E., Minnaert, A., & Ruijssenaars, W. (2008). Onderzoeksproject leerbaarheid van hoofdrekenen: naar criteria voor differentiatie en/of

planning. *Panama-post*, *27*(2), 24-28.

Evers, A., Lucassen, W., Meijer, R. R., & Sijtsma, K. (2010). *COTAN beoordelingssysteem voor de kwaliteit van tests*. Retrieved from https://www.psynip.nl/wp-content/uploads/2016/07/COTAN-Beoordelingssysteem-2010.pdf

Fulk, B. M., Brigham, F. J., & Lohman, D. A. (1998). Motivation and self-Regulation: a comparison of students with learning and behavior problems. *Remedial and Special Education*, *19*(5), 300–309.

Gelderblom, G. (2009). Iedereen kan leren rekenen. Opgevraagd van https://www.poraad.nl/nieuws-en-achtergronden/iedereen-kan-leren-rekenen

Glas, C. A. W. (2010). *Preliminary Manual of the Software Program Multidimensional Item Response Theory (MIRT).* Enschede, the Netherlands: Department of Research Methodology, Measurement and Data-Analysis, University of Twente.

Lackaye, T., Margalit, M., Ziv, O., & Ziman, T. (2006). Comparisons of Self-Efficacy, Mood, Effort, and Hope Between Students with Learning Disabilities and Their Non-LD-Matched Peers. *Learning Disabilities Research and Practice*, *21*(2), 111–121. http://doi.org/10.1111/j.1540-5826.2006.00211.x

Meijerink, H. (2008). *Over de drempels met taal en rekenen. Hoofdrapport van de Expertgroep Doorlopende Leerlijnen Taal en Rekenen.* Enschede: SLO.

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2016). *TIMSS 2015 International Results in Mathematics*. Boston, MA.

Noteboom, A. (2008). *Fundamentele doelen Rekenen-Wiskunde.* Enschede: SLO.

Noteboom, A., Aartsen, A., & Lit, S., (2017). *Tussendoelen rekenen-wiskunde voor het primair onderwijs.* Enschede: SLO

OECD. (2016). *PISA 2015 Results (Volume I). Excellence and Equity in Education*. Paris, France.

Parsons, S., & Bynner, J. (2005). *Does numeracy matter more?*. Geraadpleegd van http://eprints.ioe.ac.uk/4758/1/parsons2006does.pdf

R Development Core Team. (2012). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.

Ruijssenaars, A.J.J.M, Danhof, W., Hofstetter, W.H. & Minnaert, A.E.M.G (2017). Automatisering van basale rekenkennis en rekenproblemen: drempels in het tot stand komen van feitenkennis en procedurekennis. *Orthopedagogiek: onderzoek en praktijk, 56* (3-4), 71-85.

Stone, C. A. (1998). The metaphor of scaffolding: Its utility for the field of learning disabilities. *Journal of Learning Disabilities*, *31*(4), 344–364.

Torres, D., & Freund, R. (2014). *WrightMap: IRT Item-Person Map*. R package version 1.0.

Verhelst, N. D. (1993). Itemresponstheorie. In T. J. H. M. Eggen & P. F. Sanders (Eds.), *Psychometrie in de praktijk*. Arnhem: Centraal Instituut voor Toetsontwikkeling.

Woolf, B. P., Arroyo, I., Muldner, K., Burleson, W., Cooper, D. G., Dolan, R., & Christopherson, R. M. (2010). The effect of motivational learning companions on low achieving students and students with disabilities. In *International Conference on Intelligent Tutoring Systems* (pp. 327–337). http://doi.org/10.1007/978-3-642-13388-6_37

Auteurs

Anne van Hoogmoed is Universitair Docent Orthopedagogiek bij het Nieuwenhuisinstituut voor Onderwijsonderzoek aan de Rijksuniversiteit Groningen. **Elisa Kupers** is Universitair Hoofddocent Orthopedagogiek en Klinische Onderwijskunde bij het Nieuwenhuisinstituut voor Onderwijsonderzoek aan de Rijksuniversiteit Groningen. **Andrea Sijp** is afgestudeerd aan de master Orthopedagogiek en de master Onderwijskunde van de Rijksuniversiteit Groningen. **Joanka Vloo** is afgestudeerd aan de master Orthopedagogiek en de master Onderwijskunde van de Rijksuniversiteit Groningen. **Wilfred Hofstetter** is ontwikkelingspsycholoog en rekenspecialist bij Effectief Onderwijs. **Muirne Paap** is Universitair Docent Methodologie bij het Nieuwenhuisinstituut voor Onderwijsonderzoek aan de Rijksuniversiteit Groningen.

Correspondentieadres: A.H. van Hoogmoed, Rijksuniversiteit Groningen, Grote Rozenstraat 38, 9712 TJ Groningen; E-mail: a.h.van.hoogmoed@rug.nl

Abstract

The local and global reliability of the screening subtest of the Bareka-tests Arithmetic in children in grade 1

Quickly and accurately solving basic number operations lies at the core of success in arithmetic and mathematics. To adequately help children who lag behind in arithmetic, insight into their level of basic number facts needs to be established. The Bareka-tests are developed to measure accuracy and fluency in basic number operations using a separate fluency test and screening test. Here, we examined the reliability of the screening test in children in grade 1. Operations under 10, between 10 and 20 and over the first decade were studied. Good global reliability was shown. Local reliability estimates showed good discrimination specifically in children scoring below the mean. Therefore, the test is suitable to indicate which children need help with which type of operations, and as such can be used to foster optimal development in arithmetic.

Keywords: arithmetic, test, reliability, item response theory, test quality