Psychometrisch Onderzoek naar de Kwaliteit van Onderzoeksmaterialen

Student: S.W. van Beilen (S4504151)
Begeleider en eerste beoordelaars: dr. A.H. Stoevenbelt
Tweede beoordelaar: prof. dr. D.D. van Bergen

Rijksuniversiteit Groningen
Faculteit der Gedrags- en Maatschappijwetenschappen
Bachelorwerkstuk Pedagogische Wetenschappen
Juni 2025

Aantal woorden: 5000

Summary

This study investigates the psychometric quality of a newly developed mathematics test used in the replication study by Stoevenbelt et al. (2025), which aimed to examine the effects of stereotype threat on women's math performance. This study is focused on evaluating the psychometric properties of this test, such as its dimensionality, item parameters, reliability and measurement precision. The data originates from 1031 students aged 18 to 25 who took the easier version of the math test with a time limit. To evaluate test quality, three IRT models (1 parameter logistic model (1PLM), 2PLM and 3PLM) were compared, and the assumptions of local independence, unidimensionality, and monotonicity were tested. The 2PL model was selected for further analysis. Unidimensionality and monotonicity were partially violated, undermining the interpretability of the test scores. Findings indicate that although global reliability was acceptable (α=.777), the test lacked sufficient precision for participants with low or high ability levels. Approximately half of the participants had high standard errors, and several items showed poor alignment with participant abilities.

These results raise concerns about the validity and reliability of the test for assessing individual math ability in the context of stereotype threat research and suggest that future versions should include improved item targeting, no or appropriate time limits, and better model fit.

Uit onderzoek blijkt dat stereotiepe bedreiging ertoe kan leiden dat vrouwen slechter presteren op wiskunde toetsen, doordat zij bang zijn dat hun prestaties de negatieve stereotypen over hun groep zouden kunnen bevestigen (Spencer et al., 1999). In het replicatieonderzoek van Stoevenbelt et al. (2025) herhaalden ze het onderzoek van Johns et al. (2005) om zo de sterkte en robuustheid van het effect van stereotiepe bedreiging op de wiskundeprestaties van vrouwen te beoordelen in een grootschalig, gepreregistreerd, multilab (gezamenlijk onderzoek dat wordt uitgevoerd tussen meerdere onderzoekscentra) replicatieproject.

Om een replicatie waardevol te maken, moeten de metingen zowel in de oorspronkelijke studie als in de replicatiestudie betrouwbaar en valide zijn (Goos et al., 2024). Betrouwbaarheid zorgt ervoor dat de resultaten consistent en herhaalbaar zijn, terwijl (construct) validiteit aangeeft in hoeverre de metingen daadwerkelijk het construct meten dat bedoeld is (Bandalos, 2018). Alleen door beide te waarborgen kan de replicatiestudie de bevindingen van de oorspronkelijke studie op een geloofwaardige manier bevestigen (Golafshani, 2003). In het onderzoek van Stoevenbelt et al. (2025) zijn de psychometrische kenmerken van het meetinstrument nog niet onderzocht.

Betrouwbaarheid is een cruciale eerste stap om tot geloofwaardige bevindingen te komen. Wanneer een meting geen consistente resultaten oplevert, is het moeilijk om vertrouwen te hebben in zowel de uitkomsten als de gemeten effecten (Mellenbergh, 2011). Betrouwbaarheid is een voorwaarde voor validiteit en speelt een grote rol in de psychometrie. Het geeft aan of scores hetzelfde blijven in verschillende situaties, zoals verschillende sets van items, verschillende toetsvormen of op verschillende tijdstippen (Bandalos, 2018).

Constructvaliditeit is een vorm van validiteit en gaat over de vraag of een meetinstrument daadwerkelijk het begrip meet dat het moet meten en of de gekozen manier van meten daarbij passend is (Barry et al., 2014). Eén van de drie fasen van construct validiteit is de structurele fase. In deze fase worden statistische methoden gebruikt om de psychometrische eigenschappen van meetinstrumenten te onderzoeken, zoals de factorstructuur of interne consistentie (Flake et al., 2017).

Item Response Theorie

Een manier om psychometrische kenmerken, betrouwbaarheid en validiteit, te meten is via item respons theorie (IRT, Bandalos, 2018). IRT-modellen beschrijven de relatie tussen de eigenschap die door het instrument wordt gemeten en de antwoorden die iemand op een toets geeft (itemrespons). De itemrespons kan dichotoom (twee categorieën), of polytoom zijn (meer dan twee categorieën). Het gemeten construct kan een vaardigheid (zoals de

wiskundeprestatie), houding of overtuiging zijn (DeMars, 2010). IRT berekent de kans op een correct antwoord, afhankelijk van het vaardigheidsniveau[1] van de persoon. Elk item wordt beschreven met een itemresponsfunctie (item karakteristieke curve of ICC). Deze curve laat binnen IRT zien hoe de kans op een correct antwoord verandert afhankelijk van het vaardigheidsniveau van een deelnemer (Bandalos, 2018).

De meest gebruikte IRT-modellen voor dichotoom gescoorde items zijn het 1-parameter logistieke model (1PL) of Rasch-model, het 2-parameter logistieke model (2PL) en het 3-parameter logistieke model (3PL). Deze modellen verschillen in het aantal parameters dat wordt gebruikt om de itemreponsfunctie te beschrijven. Er kunnen drie parameters worden onderscheiden: de moeilijkheidsgraad (b-parameter); de discriminatie (a-parameter); en de pseudogokkans (c-parameter, Rasch, 1980).

Het 1PL-model is het eenvoudigste IRT-model en wordt gekarakteriseerd door de moeilijkheidsparameter. De moeilijkheidsparameter bepaalt het punt op de latente eigenschap (zoals het vaardigheidsniveau) waar de kans om het item correct te beantwoorden 50% is. Hoe groter de waarde van *b*, hoe moeilijker het item gemiddeld wordt. Dus bij een kleine *b* waarde is het item relatief makkelijk, omdat respondenten met een lagere vaardigheid toch een grotere kans hebben om het item te beantwoorden (DeMars, 2010). In dit model is de pseudogokkansparameter niet aanwezig en is de discriminatieparameter gelijk voor elk item (Bandalos, 2018).

Het 2PL-model bevat naast de moeilijkheidsparameter ook de discriminatieparameter, die aangeeft hoe goed het item onderscheid maakt tussen deelnemers met verschillende vaardigheidsniveaus. Hoe hoger de discriminatie, hoe steiler de itemresponsfunctie en hoe beter het item onderscheid kan maken (De Mars, 2010; Rizopoulous, 2006). Door rekening te houden met zowel moeilijkheid als discriminatie bepaalt het model hoe goed elk item het vaardigheidsniveau van een deelnemer meet, ook als de items niet even moeilijk zijn. Dit verhoogt de betrouwbaarheid en maakt het mogelijk gemakkelijke en moeilijke items te combineren, wat de validiteit van de toets verbetert (DeMars, 2010).

Het 3PL-model bevat alle drie de parameters: de moeilijkheid, discriminatie en pseudogokkans. De pseudogokkansparameter geeft de kans weer dat een persoon met zeer lage vaardigheid het item correct beantwoordt, simpelweg door te raden of door een andere niet-vaardigheid gerelateerde strategie te gebruiken. Het 3PL-model is complexer dan het

[1] Vaardigheidsniveau, theta-waarden en latente trek hebben dezelfde betekenis in dit onderzoek.

2PL-model doordat het de mogelijkheid biedt om gokgedrag van respondenten in te schatten. Dit maakt het geschikter voor toetsen waar gokken mogelijk is, zoals bij meerkeuzevragen (DeMars, 2010; Bandalos, 2018).

Tekortkomingen in Psychometrische Rapportage

Hoewel wetenschappers het belang van het beoordelen en rapporteren van psychometrische eigenschappen weten, toonde het onderzoek van Goos et al. (2023) een buitengewoon hoog percentage van de gepubliceerde wetenschappelijke rapporten over gezondheids- en gedragseducatie waarin geen validiteits- of betrouwbaarheidsstatistieken werden gerapporteerd.

Ook Flake et al. (2017) ontdekten dat veel constructen die in sociaal- en persoonlijkheidsonderzoek worden bestudeerd niet goed zijn getest op validiteit. Zo ontdekten ze met name een groot vertrouwen in Cronbach's alfa, dat als enige bron van bewijs voor structurele validiteit werd gebruikt. Hoewel tests zoals Cronbach's alfa en test-hertest betrouwbaarheid algemeen bekend zijn en vaak worden gerapporteerd, worden andere testen van structurele validiteit, zoals meetinvariantie, slecht begrepen en zelden uitgevoerd (Flake et al., 2017).

Wanneer de psychometrische kwaliteiten niet worden getest of gerapporteerd, kunnen onderzoekers tot foutieve conclusies en aanbevelingen komen (Barry et al., 2014). Barry et al. (2014) en Flake et al. (2020) wijzen op een structureel probleem van onderrapportage, waarbij metingen worden uitgevoerd, maar informatie over betrouwbaarheid en validiteit ontbreekt. Hierdoor blijft de kwaliteit van gebruikte instrumenten onduidelijk. Hussey en Hughes (2020) stellen dat het probleem vaak dieper ligt. Onderzoekers zijn zich soms niet eens bewust van gebrekkige meetinstrumenten en testen de kwaliteit niet, terwijl ze wel uitgaan van geldige metingen. In beide gevallen kan het voorkomen dat onderzoekers uiteindelijk iets anders meten dan oorspronkelijk de bedoeling was (Barry et al., 2014), wat leidt tot onzekerheid over de interpretatie van onderzoeksresultaten.

Met dit uitgangspunt onderzoekt deze studie de psychometrische eigenschappen van de makkelijke versie van de wiskundetoets uit het replicatieonderzoek van Stoevenbelt et al. (2025). Dit gebeurt aan de hand van item respons theorie, met als doel inzicht te krijgen in de betrouwbaarheid en (construct) validiteit van deze toets. Om vervolgens na te gaan in hoeverre de psychometrische kenmerken mogelijk van invloed zijn geweest op de interpretatie van de onderzoeksresultaten. In dit onderzoek staat de volgende onderzoeksvraag centraal:

Wat zijn de psychometrische eigenschappen van de wiskunde toets (makkelijke versie) uit het replicatieonderzoek van Stoevenbelt et al. (2025)?

De volgende deelvragen worden gebruikt om antwoord te geven op de onderzoeksvraag:

1. Wat is de modelfit van het gekozen IRT-model?
2. Wat zijn de schattingen van de itemparameters van het gekozen IRT-model?
3. Wat is de informatiefunctie en meetnauwkeurigheid van de toets voor de verschillende vaardigheidsniveaus?
4. In hoeverre meten de items op de schaal consistent hetzelfde construct?
5. Zijn de moeilijkheidsparameters goed afgestemd op de vaardigheidsniveaus?

Methode

Databeschrijving

De data die voor dit onderzoek werden gebruikt komen van een experimentele studie waarbij deelnemers willekeurig zijn toegewezen aan verschillende condities en een wiskundetoets maakten. De gehele steekproef bestond uit 1502 studenten tussen de 18 en 25 jaar. Deelnemers maakten in een gecontroleerde setting een wiskunde toets waarvoor ze 20 minuten de tijd kregen. De helft van de deelnemers werd toegewezen aan een conditie met een stereotiepe bedreiging, de rest aan een onderwijsinterventie of probleemoplossingsconditie, waarbij de indeling ongelijk was om de stereotiepe bedreiging extra te belichten. Voorafgaand aan de toets werden instructies gegeven via een audiotape en herhaald door de experimentator. De dataverzameling vond plaats tussen herfst 2019 en herfst 2022.

Het onderzoek had twee verschillende wiskundetoetsen, een makkelijke en een gemiddelde versie. Er zijn 1031 deelnemers die de makkelijke versie van de wiskundetoets hebben gemaakt. Met 1031 deelnemers wordt er voldaan aan de benodigde steekproefomvang voor alle drie de IRT-modellen (1PL-model: 100-200 deelnemers, 2PL-model: 250-500 deelnemers (Stone, 1992) en 3PL-model: 1000 of meer deelnemers (Bandalos, 2018)).

Variabelen en Instrumenten

Voor dit onderzoek wordt gebruik gemaakt van de makkelijke versie van de wiskundetoets die wordt gebruikt in het onderzoek van Stoevenbelt et al. (2025). Daarom wordt er vooral ingegaan op de wiskundetoets die nieuw is ontwikkeld voor het onderzoek van Stoevenbelt et al. (2025).

Om geschikt toetsmateriaal te ontwikkelen voor de deelnemende onderzoeksgroep in verschillende landen, werd eerst een grote pilot uitgevoerd. Zo zijn er 109 wiskundevragen uit eerdere onderzoeken naar stereotiepe bedreiging verzameld die vergelijkbaar waren met die van Johns et al. (2005) en die voortkwamen uit de Graduate Records Examination, een Amerikaanse gestandaardiseerde wiskundetoets. Uiteindelijk zijn 40 items getest in een pilot aan de Universiteit van Amsterdam. Hieruit zijn twee wiskundetoetsen ontwikkeld: een makkelijk en een gemiddelde versie, elk met 30 vragen. Elke deelnemende onderzoeksgroep deed hierna zelf een kleine toets om te bepalen welke versie het meest geschikt was vaan de deelnemers. Binnen dit onderzoek is de data gebruikt van de deelnemers die de makkelijkere versie van de toets hebben gemaakt.

De twee toetsen delen 25 items en hebben 5 unieke items. Alle items waren meerkeuze vragen, met vijf antwoordopties en één juist antwoord. De items omvatten onderwerpen uit algebra, meetkunde en statistiek. De validiteit van de wiskunde toets is niet gerapporteerd en wordt naast de betrouwbaarheid onderzocht in dit onderzoek. In dit onderzoek wordt vooral gekeken naar de losse itemscores en wordt er geen gebruik gemaakt van een somscores.

Analyseplan

De analyses en datamanagement zijn uitgevoerd in R (R Core Team, 2025), versie 2024.12.1.

Scoring van de Wiskunde Vragen

De dataset bestaat uit dichotome itemdata (goed/fout) van de wiskundetoets met meerkeuze vragen. Elk item is gescoord als 0 (fout) of 1 (goed). De laatste tien items van de toets zijn verwijderd uit de analyse, aangezien het merendeel (89%) van de deelnemers deze niet heeft kunnen beantwoorden vanwege tijdgebrek tijdens de dataverzamenling. De overige 20 items zijn meegenomen in de analyses, waarbij ontbrekende antwoorden worden geïnterpreteerd als fout en dus gecodeerd als 0.

De analyses die zijn uitgevoerd worden gekoppeld aan de deelvragen van het onderzoek.

Model Selectie

Om de deelvraag ‘Wat is de modelfit van het gekozen IRT-model?’ te beantwoorden zijn de itemparameters voor elk model geschat. Hierdoor wordt tevens de deelvraag ‘Wat zijn de schattingen van de itemparameters van het gekozen IRT-model?’ beantwoord. Vervolgens is er bekeken welk model het beste bij de data past.

De moeilijkheidsparameter zou idealiter tussen -2 en 2 moeten liggen. Deze parameter kan als volgt worden geïnterpreteerd: “Een item met b-waarde van < -2 zou zo gemakkelijk

zijn dat zelfs deelnemers met twee standaarddeviaties onder het gemiddelde een kans van minstens 50% hebben om het juiste antwoord te geven" (DeMars, 2010, p5).

De discriminatieparameter bevindt zich idealiter tussen 1.3 en 3.4. Hogere waarden van deze parameter duiden op een beter vermogen om deelnemers met verschillende vaardigheidsniveaus van elkaar te onderscheiden. De pseudogokkansparameter ligt meestal tussen 0 en 1.0 (DeMars, 2010).

Naast de schattingen van de itemparameters is ook de itemfit per model beoordeeld aan de hand van S-X^2, ook wel de chi-kwadraattoets genoemd (Orlando & Thissen, 2020). In deze toets worden deelnemers gegroepeerd op basis van hun geschatte vaardigheid. Vervolgens wordt het verschil berekend tussen de geobserveerde en verwachte antwoorden. Bij de beoordeling van de itemfit is specifiek gekeken naar significante S-X^2 waarden ($P<.05$), omdat deze kunnen wijzen op een verschil tussen de geobserveerde en verwachte antwoorden. Dit kan betekenen dat een item niet goed binnen het gekozen model past. Het model met de minste significante waarden wordt gezien als het best passend bij de data (Bandalos, 2018; DeMars, 2010; Avcu, 2021).

Daarnaast zijn zowel de lokale als globale betrouwbaarheid meegenomen in de keuze voor het best passende model (zie Methode: "Betrouwbaarheid en Meetnauwkeurigheid").

Assumpties IRT

Lokale Onafhankelijkheid

Lokale onafhankelijkheid houdt in dat de scores op een toets alleen afhankelijk zijn van de latente trek (de onderliggende eigenschap die gemeten wordt) en niet van antwoorden op een ander item in de toets (Sijtsema & Molenaar, 2002).

Om de assumptie van lokale onafhankelijkheid te controleren is er gebruik gemaakt van Yen's Q3 om itemparen te identificeren die de assumptie mogelijk schenden. Yen's Q3 is de correlatie van de residuwaarden tussen twee items. Waarden groter dan 0.2 wijzen op een mogelijke schending van lokale onafhankelijkheid (Yen, 1993).

Unidimensionaliteit

Bij het gebruik van unidimensionale IRT-modellen (zoals de modellen in dit onderzoek), wordt ervan uitgegaan dat de data zelf ook unidimensionaal zijn (Bandalos, 2018). Om deze assumptie te controleren, is er een exploratieve mokkenschaalanalyse (AISP) uitgevoerd. Deze analyse laat het algoritme bepalen hoe de schaal is opgebouwd en of er mogelijk meerdere onderliggende constructen aanwezig zijn. Als dit het geval is, wordt de assumptie van unidimensionaliteit geschonden (Drenth & Sijtsma, 2006; Avcu, 2021).

Daarnaast is er ook een confirmatieve mokkenschaalanalyse uitgevoerd. Hierbij is de assumptie van monotoniciteit eerst gecontroleerd (zie kopjes “Monotoniciteit”), omdat deze een voorwaarde is voor het uitvoeren van de confirmatieve mokkenschaalanalyse (Drenth & Sijtsma, 2006; Avcu, 2021). De confirmatieve mokkenschaalanalyse onderzoekt of de items betrouwbaar binnen de schaal passen. De H-waarde zou niet kleiner mogen zijn dan 0.3, aangezien dit zou betekenen dat er geen sprake is van unidimensionaliteit. Ook kunnen Hi-waarden van de items worden bekeken om te bepalen of er items zijn die lager scoren dan 0.3 (Sijtsema & Molenaar, 2002).

Monotoniciteit

Voor alle IRT-modellen geldt dat de kans op een correct antwoord moet toenemen naarmate het vaardigheidsniveau hoger wordt. Om deze assumptie te controleren, wordt er ook gebruik gemaakt van een mokkenschaalanalyse om de itemfit te beoordelen. Per item is er een item karakteristieke curve (ICC) geschat (Zie Inleiding). Voor elk item is beoordeeld of de curve een stijgend verloop heeft. Zolang de grijze betrouwbaarheidsband rond de curve niet daalt (Zie Bijlage 1), wordt aan deze assumptie voldaan (Drenth & Sijtsma, 2006; Avcu, 2021).

Betrouwbaarheid en Meetprecisie

Om de deelvraag “Wat is de informatiefunctie en meetnauwkeurigheid van de toets voor de verschillende vaardigheidsniveaus?” te beantwoorden, is de testinformatiefunctie geschat. Binnen IRT geeft deze functie aan hoe nauwkeurig de toets meet op verschillende vaardigheidsniveaus. De testinformatiefunctie is opgebouwd uit de som van de informatiefuncties van afzonderlijke items, die tonen hoeveel informatie en item levert voor verschillende vaardigheidsniveaus. Items zijn het meest informatief wanneer hun moeilijkheid aansluit bij het vaardigheidsniveau van de deelnemer. Te makkelijke items dragen weinig bij, omdat bijna iedereen ze goed maakt (Bandalos, 2018; DeMars, 2010; Avcu, 2021). Een hogere informatiewaarde betekent dat de toets op dat vaardigheidsniveau betrouwbaarder meet. Een lagere informatiewaarde wijst op minder nauwkeurigheid. Zo is bepaald voor welke vaardigheidsniveaus de toets goed of juist minder goed meet.

Daarnaast is de standaardmeetfout (SE) gebruikt als maat voor de meetnauwkeurigheid. De standaardmeetfout is een inverse functie van de betrouwbaarheid: hoe lager de SE, hoe hoger de betrouwbaarheid op dat vaardigheidsniveau. De SE hangt direct samen met de informatiefunctie van de toets: hoe meer informatie een toets op een bepaald vaardigheidsniveau geeft, hoe kleiner de standaardfout en hoe preciezer de schatting van dat

vaardigheidsniveau. Binnen IRT is de standaardmeetfout afhankelijk van het vaardigheidsniveau en dus niet voor alle deelnemers gelijk. Deelnemers met hetzelfde niveau hebben dezelfde SE, wat het mogelijk maakt om de meetnauwkeurigheid per vaardigheidsniveau te beoordelen (DeMars, 2010). Lagere SE-waarden wijzen op een nauwkeurigere schatting; hogere waarden op meer onzekerheid (Cook & Wind, 2024).

Om de deelvraag 'Zijn de moeilijkheidsparameters goed afgestemd op de vaardigheidsniveaus?' te beantwoorden, is er gebruik gemaakt van targeting. Targeting is een methode waarmee onderzocht wordt of de moeilijkheidsparameters van de toetsitems aansluiten bij het vaardigheidsniveau van de deelnemers. Een goede afstemming betekent dat er voor elk vaardigheidsniveau binnen de groep deelnemers voldoende items zijn die passen bij dat niveau. Targeting geeft daarmee inzicht in de meetnauwkeurigheid van de toets voor verschillende vaardigheidsniveaus. Dit hangt nauw samen met de individuele betrouwbaarheid: als de items niet aansluiten bij iemand vaardigheidsniveau, dan kan de toets voor die persoon minder nauwkeurig meten (Cook & Wind, 2024).

Resultaten

Model selectie

Om te bepalen welk model het meest geschikt is voor verdere analyses, zijn de drie IRT-modellen met elkaar vergeleken: het 1PL-, 2PL- en 3PL-model.
Het 1PL-model heeft moeilijkheidsparameters die variëren tussen -1.692 en 2.772, waarvan 14 items een positieve waarde hebben. De meerderheid van de items is dus gericht op deelnemers met een bovengemiddelde vaardigheid, aangezien een b=0 staat voor een gemiddelde vaardigheid. Hoewel de waarden redelijk verspreid liggen binnen het bereik van -2 tot 2, neigt de verdeling meer naar de positieve kant. De itemfit werd beoordeeld met behulp van de S-X^2 toets, waarbij bleek dat 12 items significant zijn ($P<.05$), wat duidt op een matige fit voor deze items binnen het model. De globale betrouwbaarheid van het 1PL-model is 0.723, wat volgens de richtlijnen van COTAN (2010) als voldoende wordt beschouwd.

Het 2PL-model heeft moeilijkheidsparameters die variëren tussen -1.704 en 1.880, waarvan 14 positieve waarden. Hiermee is dit model vooral gericht op deelnemers met bovengemiddelde vaardigheid, met een goede verdeling van de parameter. De discriminatieparameters liggen tussen 0.495 en 2.175 (zie Bijlage 2). Waarden onder 0.6 zijn volgens DeMars (2010) onvoldoende, om personen met verschillende vaardigheidsniveaus te

discrimineren. Bij de S-X2 toets bleken 12 items significant (P<.05). De globale betrouwbaarheid van dit model is 0.777, en wordt als voldoende beschouwd (COTAN, 2010).

Het 3PL-model heeft moeilijkheidsparameters tussen de -1.348 en 1.572, met 15 positieve waarden. Dit model is daarom gericht op deelnemers met bovengemiddelde vaardigheid. De parameterschattingen zijn redelijk verspreid. De discriminatieparameters in dit model variëren sterk: van 0.639 tot 5.469. Opvallend is dat vijf items een extreem hoge waarde hebben (boven de 4.0), waarvan drie zelfs boven de 5.0 liggen (zie Bijlage 3). Idealiter liggen deze waarden tussen de 1.3 en 3.4. Dergelijke uitschieters kunnen wijzen op overschatting of modelinstabiliteit (DeMars, 2010). De pseudogokkans ligt tussen 0.00 en 0.18, wat acceptabel is. Wat betreft de itemfit blijken 3 items significant te zijn, wat duidt op een redelijke fit. De globale betrouwbaarheid is 0.768, en wordt als voldoende beschouwd (COTAN, 2010).

Op basis van deze analyses is gekozen om het 2PL-model te gebruiken voor verdere analyses. Deze keuze is gebaseerd op meerdere overwegingen. De discriminatieparameters van het 3PL model zijn problematisch, wat wijst op instabiliteit. Het 2PL-model biedt een hogere globale betrouwbaarheid dan het 1PL model, en de moeilijkheidsparameters zijn beter verdeeld bij het 2PL-model. Daarom is het 2PL-model als meest geschikt beoordeeld voor de verdere analyses in deze studie.

Assumpties

Lokale Onafhankelijkheid

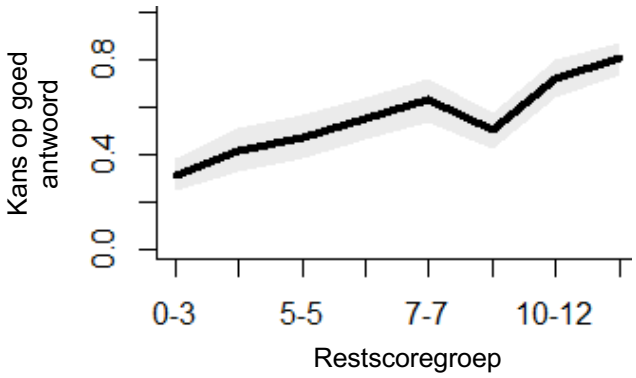
Uit de analyses blijkt één itempaar (item 19 en item 20) met elkaar te correleren (0.316). Voor een verklaring van deze correlatie moet gekeken worden of de vragen inhoudelijk op elkaar lijken. Vraag 19 is een meetkundige vraag en vraag 20 is een rekenkundige-/verhoudingsvraag. Inhoudelijk verschillen de vragen van elkaar. Daarom is er geen reden om te concluderen dat de lokale onafhankelijkheid geschonden wordt.

Monotoniciteit

Monotoniciteit vereist dat de ICC (item karakteristieke curve) overal niet-dalend is. Vraag 2 (zie Figuur 1) heeft een lokaal dalend stuk, en kan dus wijzen op een mogelijke schending. De rest van de vragen tonen een stijgende ICC. Deze assumptie is dus mogelijk geschonden. Dit kan leiden tot een verkeerde inschatting van de vaardigheid van deelnemers, wat de betrouwbaarheid van de analyses verminderd. Onderzoek laat zien dat dit item uit de test gehaald zou moeten worden (Meijer, 2014).

Figuur 1

Item Karakteristieke Curve van Vraag 2



Unidimensionaliteit

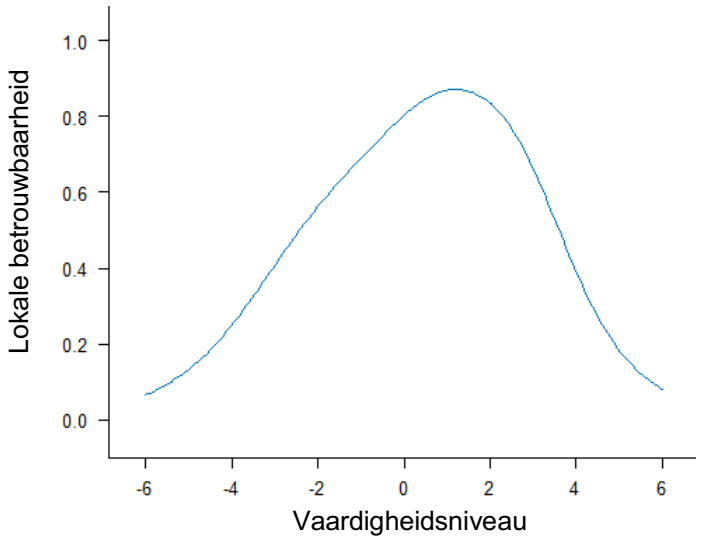
De uitgevoerde analyse (AISP, exploratieve mokkenschaalanalyse) toont aan dat er verschillende schalen zijn voor de items (zie Bijlage 4), wat duidt op een mogelijke schending van de assumptie. Om te kijken hoe groot de schending is, is de confirmatieve mokkenschaalanalyse uitgevoerd. De H-waarde is 0.270, wat lager is dan 0.3 en duidt op een matige schaal. Daarnaast zijn er 12 Hi-waarden onder de 0.30 (zie Bijlage 5), wat suggereert dat deze items weinig bijdragen aan de schaalconsistentie. Deze assumptie is dus mogelijk geschonden.

Betrouwbaarheid en Meetnauwkeurigheid

De globale betrouwbaarheid is α=0.777, wat als voldoende wordt beschouwd volgens de richtlijnen van COTAN (2010). De lokale betrouwbaarheid is het hoogst rond een vaardigheidsniveau van 2, wat betekent dat de toets vooral nauwkeurig onderscheid maakt tussen personen met een bovengemiddelde vaardigheid. Dit blijkt uit het feit dat de betrouwbaarheid op dit vaardigheidsniveau boven de 0.80 ligt (zie Figuur 2.1). De betrouwbaarheid blijft boven de 0.70 voor deelnemers met een vaardigheidsniveau van -1 tot 3. Voor vaardigheidsniveaus buiten dit bereik, zowel lager als hoger, ligt de betrouwbaarheid onder de 0.70. Volgens Buckendahl et al. (2005) wordt een betrouwbaarheid minder dan 0.70 als onvoldoende beschouwd voor het doen van uitspraken op groepsniveau.

Figuur 2.1

Lokale Betrouwbaarheid bij Verschillende Vaardigheidsniveaus

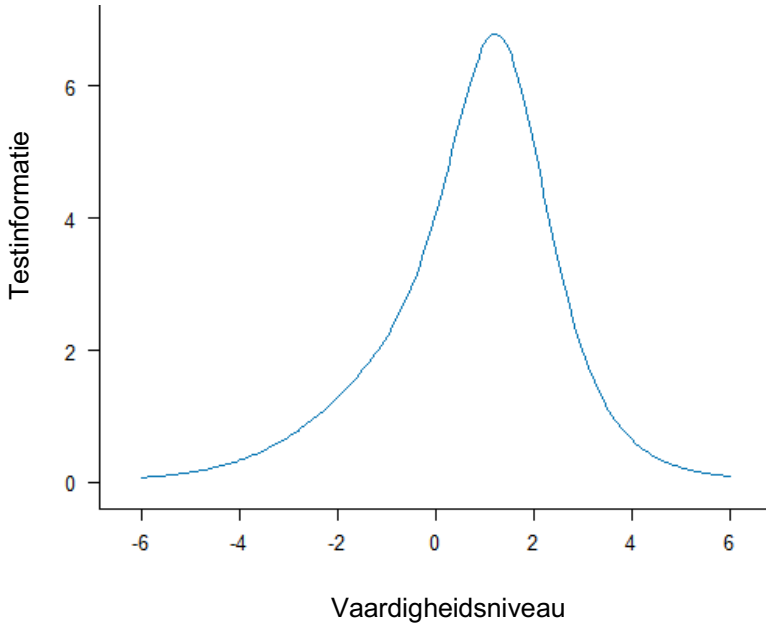


Informatiefunctie

De informatiefunctie van de toets toont een smalle, steile curve met een duidelijke piek bij een vaardigheidsniveau tussen 1 en 2 (zie Figuur 2.2). De maximale informatie ligt rond een waarde van 6.5. Dit betekent dat de toets het meest informatief is voor deelnemers met een gemiddeld tot bovengemiddeld vaardigheidsniveau. Buiten dit bereik neemt de informatiewaarde af, wat aangeeft dat de toets minder informatie biedt over vaardigheidsniveaus die lager of hoger liggen.

Figuur 2.2

Testinformatiefunctie

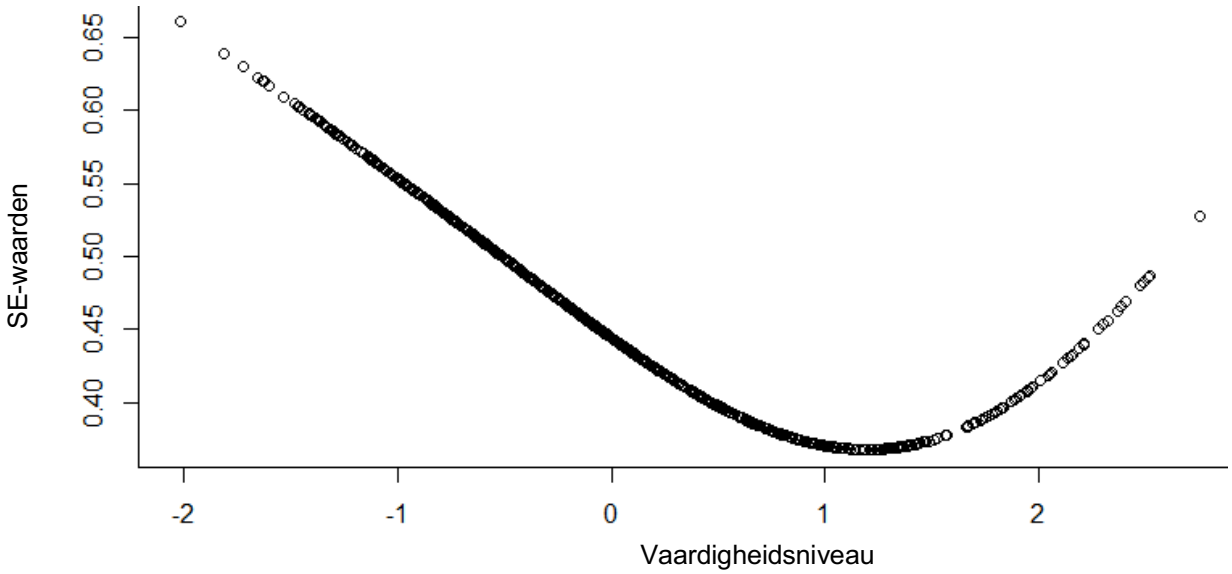


Meetnauwkeurigheid

De standaardfout mag bij een betrouwbaarheid van 0.78 ongeveer 0.469 zijn. Uit Figuur 2.3 blijkt dat de standaardfout toeneemt bij deelnemers met een vaardigheidsniveau tussen ongeveer -2 en net onder 0, evenals voor vaardigheidsniveaus boven 2.5. Dit betekent dat metingen in deze uiterste gebieden minder precies zijn, doordat de toets zelf minder informatie oplevert op die vaardigheidsniveaus. Uit Figuur 2.4 blijkt dat ongeveer 475 deelnemers een hoge standaardfout hebben, dit is net iets minder dan de helft van alle deelnemers. Voor bijna de helft van de deelnemers is de vaardigheidsschatting minder betrouwbaar, doordat hun standaardfout hoog is.

Figuur 2.3

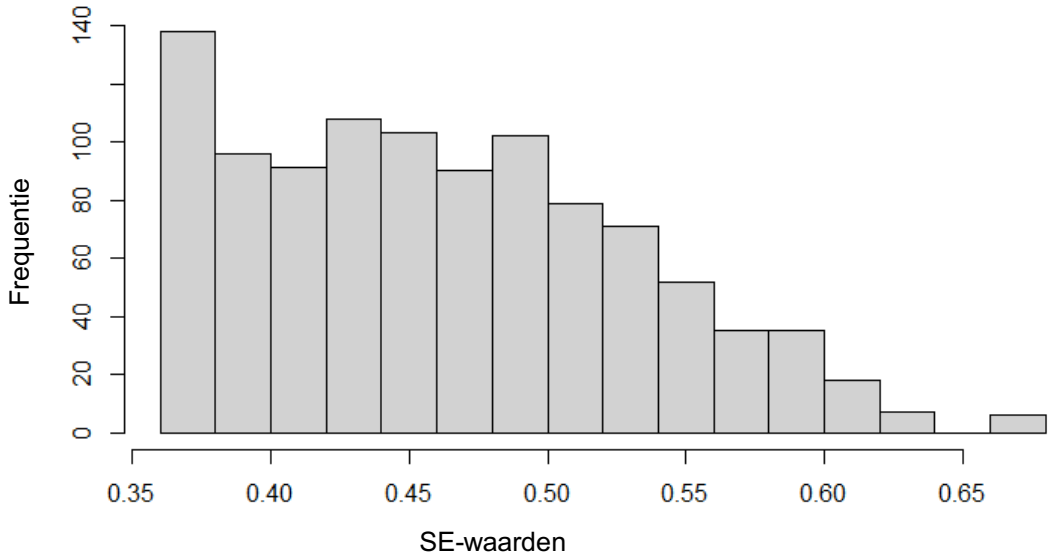
Standaardfout in Relatie tot het Vaardigheidsniveau



Noot. SE = standaardmeetfout

Figuur 2.4

Histogram Frequentie van SE-waarden

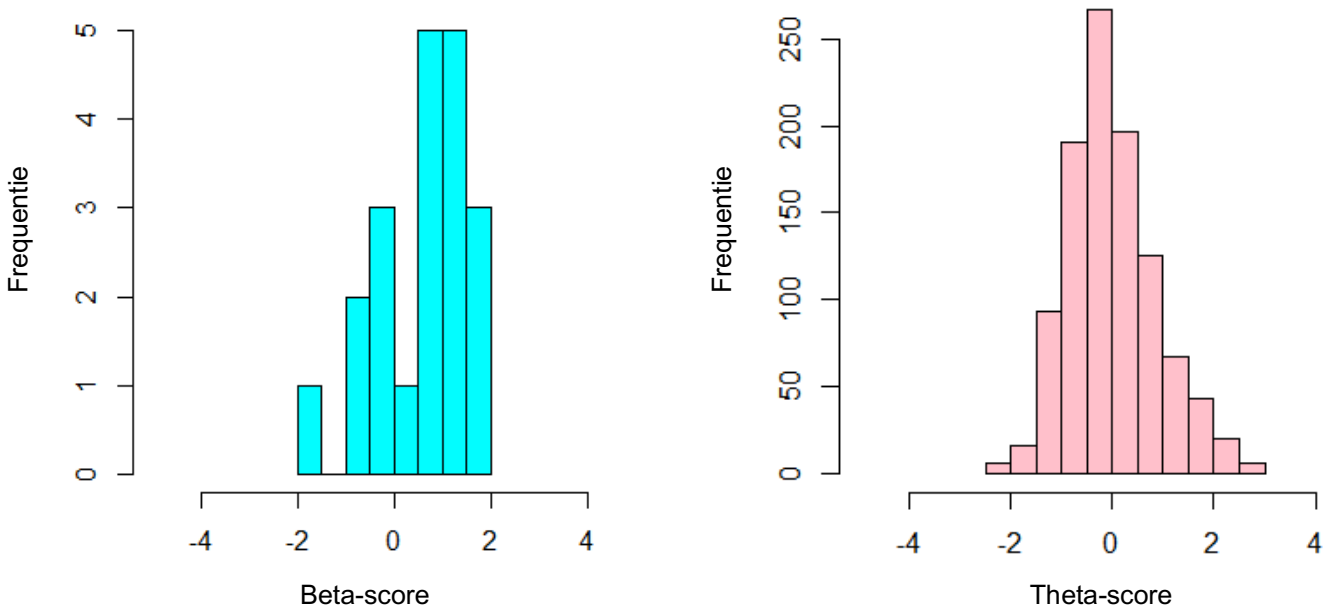


Targeting

In Figuur 2.5 (zie volgende bladzijde) zijn twee grafieken weergegeven. De blauwe grafiek toont de moeilijkheidsniveaus (beta-waarden) van de items, terwijl de roze grafiek de verdeling van de vaardigheidsniveaus (theta-waarden) van de deelnemers weergeeft. Uit de vergelijking van beide grafieken blijkt dat niet alle vaardigheidsniveaus worden gedekt door items met een overeenkomstig moeilijkheidsniveau. Dit is zichtbaar doordat de blauwe grafiek enkele gaten vertoont, oftewel gebieden waar geen items aanwezig zijn. Vooral bij vaardigheidsniveaus boven de 2 en rond -1 en -2 ontbreken passende items. Hierdoor zijn er voor ongeveer 130 deelnemers onvoldoende passende items om hun vaardigheidsniveau nauwkeurig te meten.

Figuur 2.5

Histogrammen van de Moeilijkheidsgraad en de Vaardigheidsniveaus



Samenvatting resultaten

Voor de verdere analyses is gekozen voor het 2PL-model, vanwege de hogere betrouwbaarheid ten opzichte van het 1PL-model en stabielere parameters dan het 3PL-model. De assumpties van lokale onafhankelijkheid is voldaan. De monotoniciteit wordt mogelijk geschonden bij één item (vraag 2). Ook de unidimensionaliteit is twijfelachtig. De schaal consistentie is matig (H=0.270) en meerdere items dragen zwak bij aan de schaal (12 Hi-waarden <.30).

De betrouwbaarheid van de toets is globaal voldoende (α=.777). De meetnauwkeurigheid is het hoogst bij een vaardigheidsniveau van 2 (bovengemiddeld). Voor bijna de helft van de deelnemers is de standaardfout relatief hoog. Targeting laat zien dat er onvoldoende items zijn voor deelnemers met zeer lage of hoge vaardigheidsscores. Voor ongeveer 130 deelnemers zijn er geen goed passende items.

Discussie

In dit onderzoek zijn de psychometrische eigenschappen van de makkelijke versie van de wiskundetoets uit het replicatieonderzoek van Stoevenbelt et al. (2025) geanalyseerd met het 2PL-model. De resultaten moeten voorzichtig worden geïnterpreteerd, aangezien de modelfit niet optimaal was (Bandalos, 2018; Wainer & Wang, 2000). Dat is helaas niet ongebruikelijk binnen de IRT. Zo laat onderzoek van Wainer en Thissen (1987) zien dat geen enkel model perfect aansluit bij de data. Er zullen altijd fouten in de modelfit zijn. Wainer en Thissen (1987) benadrukken dat modelkeuze niet alleen op modelfit moet worden gebaseerd, maar ook op andere factoren zoals modelstabiliteit.

Binnen het 2PL-model bleken de moeilijkheidsparameters overwegend positief, wat inhoudt dat de items vooral geschikt zijn voor deelnemers met een bovengemiddelde vaardigheid. Hoewel dit de makkelijke versie van de toets betrof, is dit verklaarbaar binnen onderzoek naar stereotiepe bedreiging, waarin effecten vaak pas optreden bij moeilijkere toetsen (Nguyen & Ryan, 2008). De makkelijkste items hadden bovendien vaak een lage discriminatie. Discriminatieparameters zijn bij makkelijke items lastig betrouwbaar te schatten, wat kan leiden tot onderschatting (DeMars, 2010). Eén item (vraag 1) had een lage discriminatie (<0.6) en onderscheidt daarmee onvoldoende tussen vaardigheidsniveaus.

De toets blijkt drie verschillende constructen te meten (de items meten niet hetzelfde construct), waardoor de assumptie van unidimensionaliteit wordt geschonden. Ook is er een lichte schending van monotoniciteit (vraag 2). Volgens Schedl, Thomas en Way (1995) kan het toepassen van unidimensionale IRT-modellen bij multidimensionaliteit leiden tot een vertekening van vaardigheidsschattingen, minder informatie en hogere standaardfouten.

De globale betrouwbaarheid van de toets is acceptabel, maar de lokale betrouwbaarheid toont dat alleen deelnemers met (net) bovengemiddelde vaardigheid betrouwbaar gemeten worden. Volgens Cook & Wind (2024) kan een hoge (globale) betrouwbaarheid verklaard worden door een brede schaal aan items en een grote steekproefgrootte, omdat grotere steekproefgroottes van personen over het algemeen meer variatie in personen impliceren, wat meer informatie biedt om onderscheid te maken tussen items. Het kan dus zijn dat de globale betrouwbaarheid van de toets overschat is.

De informatiefunctie laat hetzelfde beeld zien. De informatie is het hoogste voor deelnemers met een bovengemiddelde vaardigheid. Dit roept opnieuw de vraag op: ‘Wat is de doelgroep van de toets?’ Aangezien het de makkelijke versie betreft, zou verwacht worden dat

de meeste informatie bij gemiddelde of lagere vaardigheid ligt. De beperkte informatiewaarde is mogelijk deelt het gevolg van het verwijderen van de laatste 10 items (zie Methode). Dit was nodig om vertekening door missing data te voorkomen, maar kan de informatiewaarde hebben beïnvloed.

Concluderend, is de meetnauwkeurigheid van de toets over het algemeen laag, vooral voor deelnemers met een lagere vaardigheid. Juist bij deze groep zou je een hogere nauwkeurigheid verwachten (gezien het doel van de toets). De hoge standaardfouten in dit vaardigheidsgebied roepen dan ook vragen op over in hoeverre de toets in staat is om lagere vaardigheidsniveaus betrouwbaar te meten.

Er zijn weinig items die aansluiten bij deelnemers met lage of zeer hoge vaardigheid. De moeilijkheidsparameters zijn dus niet volledig afgestemd op alle vaardigheidsniveaus. Dit kan twee oorzaken hebben: het relatief lage vaardigheidsniveau van de voornamelijk eerstejaars psychologiestudenten, of de toets die te moeilijk was (gezien het ontbreken van eenvoudigere items), of een combinatie van beide.

Beperkingen van dit Onderzoek

Aan het begin van dit onderzoek zijn de laatste 10 items verwijderd, omdat deelnemers daar te weinig tijd voor hadden en dit mogelijk de resultaten had kunnen vertekenen. Volgens Glas en Pimentel (2008) mag dit soort missing data echter niet genegeerd worden, omdat het aantal gemaakte items samenhangt met het vaardigheidsniveau. Het negeren hiervan kan leiden tot vertekening van de itemparameters (Holman & Glas, 2005; Bradlow & Thomas, 1998).

Ook is de data die in dit onderzoek is gebruikt verzameld in een low-stake setting. Dit is een setting waarin de uitkomst van een toets weinig of geen directe gevolgen voor de deelnemer zelf heeft. Mogelijk waren deelnemers minder gemotiveerd om zichzelf in te spannen. Low-stake settingen kunnen de validiteit van de testscores van de deelnemers beïnvloede (Wise & DeMars, 2005).

Daarnaast was in dit onderzoek het wellicht passender geweest om DIF-analyses uit te voeren, hier wordt in het kopje ‘vervolgonderzoek’ meer over uitgelegd.

Vervolgonderzoek

Vervolgonderzoek zou zich kunnen richten op het afschaffen van de tijdslimiet, omdat die bij een toets die vaardigheid (en niet de snelheid) wil meten, invloed kan hebben op de

betrouwbaarheid, constructvaliditeit en de dimensionaliteit van de test (Stoevenbelt et al., 2022; Lu & Sireci, 2007). Als dit niet haalbaar is, kan overwogen worden om missing data te modelleren met het ‘sequential model’ van Tutz (1990). Zo kan de missing data systematisch op een statistische manier geanalyseerd worden.

Een tijdslimiet bij een vaardigheidstoets kan leiden tot multidimensionaliteit, doordat naast vaardigheid ook snelheid een rol gaat spelen (Lu & Sireci, 2007). Dit maakt de toets niet alleen oneerlijk en onbetrouwbaar, maar kan ook leiden tot schending van unidimensionaliteit, wat standaardfouten en parameters kan vertekenen (DeMars, 2010). Een andere mogelijke oorzaak van multidimensionaliteit is differentiële itemfunctionering (DIF), waarbij bijvoorbeeld demografische kenmerken zoals geslacht invloed hebben op antwoorden. Vervolgonderzoek kan met DIF-analyses nagaan of bepaalde items systematisch moeilijk zijn voor meisjes onder stereotiepe dreiging dan voor meisjes in de controlegroep, ondanks vergelijkbare vaardigheid (Flore, 2018).

Om de toets beter af te stemmen op de doelgroep, kunnen in de toekomst extra items worden ontwikkeld voor specifieke vaardigheidsniveaus. Hierbij moet echter gewaakt worden voor een te lange toets. Een praktische benadering is om selectief bestaande items te verwijderen om ruimte te maken voor nieuwe, doelgerichte items (Cook & Wind, 2024). Daarnaast verdient item 2 nader onderzoek, aangezien dit item in de huidige analyse tekenen van schending van monotoniciteit vertoonde, wat kan wijzen op problemen in de itemconstructie of interpretatie. Tot slot is aanvullend onderzoek nodig om de impact van de modelmisfit beter te kunnen inschatten, bijvoorbeeld door alternatieve modellen te vergelijken of itemfitanalyses uit te voeren.

Terugkoppeling maatschappelijke en wetenschappelijke relevantie

Op basis van dit onderzoek kan voorzichtig worden geconcludeerd dat de globale betrouwbaarheid van de toets acceptabel is, maar de lokale betrouwbaarheid en informatiewaarde onvoldoende voor de beoogde doelgroep. De (construct) validiteit is waarschijnlijk beperkt door schending van unidimensionaliteit en monotoniciteit, en door een slechte afstemming van de moeilijkheidsparameters op het vaardigheidsniveau van de doelgroep. De toets meet het nauwkeurigst bij een beperkte groep deelnemers met net bovengemiddeld vaardigheid en is daardoor minder geschikt om effecten van stereotiepe bedreiging vast te stellen. Dit roept vragen op over de geschiktheid van de toets in de replicatiestudie van Stoevenbelt et al. (2025), waar de beperkte meetnauwkeurigheid, vooral

bij lagere vaardigheidsniveaus, mogelijk het detecteren van effecten bemoeilijkte en daarmee de conclusies ondermijnt.

Dit onderzoek laat zien hoe essentieel psychometrisch onderzoek is. Ondanks enkele beperkingen levert het waardevolle inzichten op voor het verbeteren van de psychometrische kwaliteit en doelgerichtheid van de toets, en daarmee voor het versterken van de gebruikte onderzoeksmethoden.

Literatuurlijst

Avcu, A. (2021). Item response theory based psychometric investigation of SWLS of university students. *International Journal of Psychology and Educational Studies*, *8*(2), 27- 37. https://doi.org/10.52380/ijpes.2021.8.2.265

Bandalos, D. L., (2018). *Measurement theory and applications for the social sciences.* The Guilford Press. https://www.soc.univ.kiev.ua/sites/default/files/library/elopen/methodology_in_the_social_sciences_deborah_l._bandalos_-_measurement_theory_and_applications_for_the_social_sciences-the_guilford_press_2018.pdf

Barry, A. E., Chaney, B., Piazza-Gardner, A. K., & Chavarria, E. A. (2014). Validity and reliability reporting practices in the field of health education and behavior: A review of seven journals. *Health Education & Behavior*, *41*(1), 12-18. https://doi.org/10.1177/1090198113483139

Bradlow, E. T., & Thomas, N. (1998). Item response theory models applied to data allowing examinee choice. *Journal of Educational and Behavioral Statistics*, *23*(3), 236-243. https://doi.org/10.3102/10769986023003236

Buckendahl, C. W., Impara, J. C., & Plake, B. S. (2005). District accountability without a state assessment: A proposed model. *Educational Measurement: Issues and Practice, 21*(4), 6-16. https://doi.org/10.1111/j.1745-3992.2002.tb00102.x

Cook, R. M., & wind, S. A. (2024). Item response theory; A modern measurement approach to reliability and precision for counseling researches. *Measurement and evaluation in counseling and development*, *57*(2), 116-135. https://doi.org/10.1080/07481756.2023.2301284

DeMars, C. (2010). *Item Respons Theory.* Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195377033.001.0001

Drenth, P., & Sijtsma, K. (2005). *Testtheorie: inleiding in de theorie van de psychologische test en zijn toepassingen* (4de editie). Bohn Stafleu van Loghum.

Evers, A., Lucassen, W., Meijer, R., Sijtsma, K., & COTAN. (2009). *COTAN review system for evaluating test quality*. https://nip.nl/wp-content/uploads/2022/05/COTAN-review-system-for-evaluating-test-quality.pdf

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456-465. https://doi.org/10.1177/2515245920952393

Flake, J. K., Pek J., Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological & Personality Science*, *8*(4), 370-378. https://doi.org/10.1177/1948550617693063

Flore, P. C. (2018). *Stereotype Threat and Differential Item Functioning: A Critical Assessment.* Tilburg University. https://repository.tilburguniversity.edu/bitstreams/2a23eb08-c804-41bb-a936-59706aa0d301/download

Glas, C. A. W., & Pimentel, J. L. (2008). Modeling nonignorable missing data in speeded tests. *Educational and Psychological measurement*, *68*(6), 907-922. https://doi.org/10.1177/0013164408315262

Golafshani, N. (2003). Understanding reliability and validity in qualitative research. *The Qualitative Report*, *8*(4), 597-606. https://doi.org/10.46743/2160-3715/2003.1870

Goos, C., Bakker, M., Wicherts, J., & Nuijten, M. B. (2023). *Assessing Reliable and Valid*

Measurement as a Prerequisite for Informative Replications in Psychology. Department of Methodology and Statistic. Tilburg University. https://doi.org/10.31234/osf.io/2gau8

Holman, R., & Glas, C. A. W. (2005). Modelling non-ignorable missing data mechanisms with item response theory models. *British Journal of Mathematical and Statistical Psychology*, *58*(1), 1-17. https://doi.org/10.1111/j.2044-8317.2005.tb00312.x

Hussey, I., & Hughes, S. (2020). Hidden invalidity among 15 commonly used measures in social and personality psychology. *Advances in Methods and Practices in Psychological Science*, *3*(2), 166-184. https://doi.org/10.1177/251524591988290

Johns, M., Schmader, T., & Martens, A. (2005). Knowing is half the battle: Teaching stereotype threat as a means of improving women's math performance. *Psychological science*, *16*(3), 175-179. https://doi.org/https://doi.org/10.1111/j.0956-7976.2005.00799.x

Lu, Y., & Sireci, S. G. (2007). Validity issues in test speediness. *Educational Measurement: Issues and Practice*, *26*(4), 29-37. https://doi.org/10.1111/j.1745-3992.2007.00106.x

Meijer, R. R., Tendeiro, J.N., & Wanders, R. B. K. (2014). The use of nonparametric item response theory to explore data quality. *Handbook of Item Response Theory Modeling: Applications to typical performance assessment.* 85-110. https://pure.rug.nl/ws/portalfiles/portal/111974936/paper.pdf

Mellenbergh, G. J. (2011). *A conceptual introduction to psychometrics: Development, analysis and application of psychological and educational tests*. Eleven international publishing.

Nguyen, H. H. D., & Ryan, A. M. (2008). Does stereotype threat affect test performance of minorities and women? A meta-analysis of experimental evidence. *Journal of Applied Psychology*, *93*(6), 1314-1334. https://doi.org/10.1037/a0012702

Orlando, M. & Thissen, D. (2020). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*(1), 50-64. https://doi.org/10.1177/01466216000241003

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment test.* University of Chicago Press.

Rizopoulous, D. (2006). Ltm: An R package for latent variable modeling and item response Analyses. *Journal of Statistical Software*, *17*(5), 1-25. https://doi.org/10.18637/jss.v017.i05

Schedl, M. A., Thomas, N., & Way, W. D. (1995). *An investigation of proposed revision to section 3 of the TOEFL Test.* Educational Testing Service Princeton, New Jersey. http://dx.doi.org/10.1002/j.2333-8504.1994.tb01615.x

Sijtsema, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory.* SAGE Publications. https://doi.org/10.4135/9781412984676

Spencer, S. J., Steele, C. M., & Quin, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology*, *35*(1), 4-28. https://doi.org/10.1006/jesp.1998.1373

Stoevenbelt, A. H., Flore, P. C., Wicherts, J. M., Bergers, R., Calin-Jageman, R. J., Gomez, L.A., Hüffmeier, J., Imundo, M. N., Martin, S. D., Pan, S. C., Phillips, L. A. T., Pietschnig, J., Ripp, R., Roër, J. P., Schleu, J. E., Torka, A., Verschuere, B., Voracek, M., & Weiss, L. (2025). *Registered replication report stereotype threat*. https://osf.io/preprints/psyarxiv/qctkp_v1

Stoevenbelt, A. H., Wicherts, J. M., Flore, P. C., Phillips, L. A. T., Pietschnig, J., Verschuere, B., Voracek, M., & Schwabe, I. (2023). Are speeded tests unfair? Modeling the impact of time limits on the gender gap in mathematics. *Educational and Psychological Measurement*, *83*(4), 684-709. https://doi.org/10.1177/00131644221111076

Stone, C.A. (1992). Recovery of marginal maximum likelihood estimates in the two-parameter logistic response model: An evaluation of MULTILOG. *Applied Psychological Measurement*, *16*(1), 1-16. https://doi.org/10.1177/014662169201600101

Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology*, *43*(1), 39-55. https://doi.org/10.1111/j.2044-8317.1990.tb00925.x

Wainer, H., & Thissen, D. (1987). Estimating ability with the wrong model. *Journal of Educational Statistics*, *12*(4), 339-368. https://doi.org/10.2307/1165054

Wainer, H., & Wang, C. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement*, *37*(3), 203-220. https://doi.org/10.1111/j.1745-3984.2000.tb01083.x

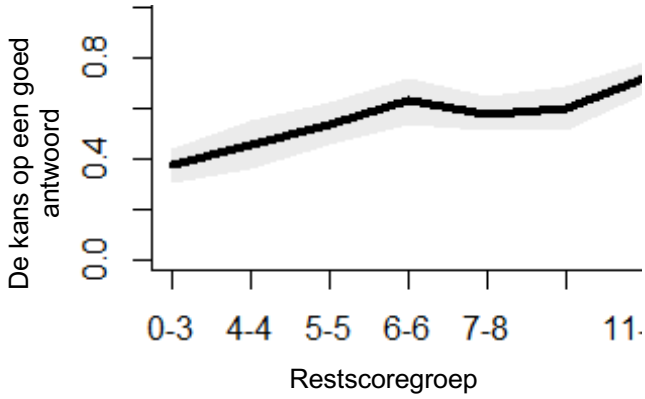
Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, *10*(1), 1-17. https://doi.org/10.1207/s15326977ea1001_1

Yen, W.M. (1993). Scaling performance assessments: Strategies for managing local item dependence, *Journal of Educational measurement*, *30*(3), 187-213. https://doi.org/10.1111/j.1745-3984.1993.tb00423.x

Bijlagen

Bijlage 1

Illustratie van een Item Karakteristieke Curve



Noot. Grijze gebied rond de curve is de betrouwbaarheidsband.

Bijlage 2

Itemparameters van het 2PL-model

Item	*a*	*b*
V1	0.495	-0.499
V2	0.719	-0.244
V3	0.603	1.198
V4	0.953	-0.899
V5	0.923	-0.736
V6	0.955	-1.704
V7	*0.851*	*0.073*
V8	0.942	-0.470
V9	0.896	0.549
V10	0.913	0.831
V11	0.910	1.156
V12	1.027	1.245
V13	1.401	0.763
V14	1.590	0.788
V15	1.910	0.921
V16	1.393	1.848
V17	1.733	1.111
V18	2.175	1.419
V19	1.754	1.880
V20	1.461	1.635

Noot. *a*=discriminatieparameter. *b*=moeilijkheidsparameter.

Bijlage 3

Itemparameters van het 3PL-model

Item	*a*	*b*	*c*
V1	0.639	-0.389	0.001
V2	1.078	0.308	0.181
V3	0.742	1.019	0.000
V4	1.269	-0.733	0.000
V5	1.150	-0.620	0.000
V6	1.355	-1.358	0.001
V7	0.967	0.086	0.000
V8	1.182	-0.386	0.000
V9	1.037	0.536	0.007
V10	0.962	0.827	0.002
V11	1.264	1.232	0.002
V12	2.159	1.344	0.126
V13	2.761	1.030	0.139
V14	3.998	1.028	0.135
V15	**5.398**	1.069	0.097
V16	3.297	1.572	0.049
V17	**5.076**	1.170	0.087
V18	**5.469**	1.329	0.037
V19	**4.372**	1.558	0.028
V20	**4.367**	1.450	0.066

Noot. a=discriminatieparameter. *b*=moeilijkheidsparameter. *c*= pseudogokkansparameter. Opvallende waarden zijn dikgedrukt.

Bijlage 4

AISP (Automated Item Selection Procedure)

Item	Schaal
V1	0
V2	0
V3	2
V4	2
V5	2
V6	2
V7	2
V8	2
V9	2
V10	3
V11	0
V12	3
V13	1
V14	1
V15	1
V16	1
V17	1
V18	1
V19	1
V20	1

Bijlage 5

Confirmatieve mokkenschaalanalysen.

Item	Hi-waarden
V1	0.180
V2	0.219
V3	0.178
V4	0.305
V5	0.289
V6	0.353
V7	0.247
V8	0.282
V9	0.256
V10	0.234
V11	0.228
V12	0.247
V13	0.278
V14	0.291
V15	0.332
V16	0.313
V17	0.314
V18	0.402
V19	0.386
V20	0.296