Bernard P. Veldkamp
Faculty of Behavioural, Management and Social Sciences, Universiteit Twente, Enschede
en Research Center voor Examinering en Certificering, Vaassen
b.p.veldkamp@utwente.nl

Het wiskundige fundament van toetsen en examens

Toetsen en examens hebben veel invloed op schoolcarrières en kansen op een baan, maar hoe worden cijfers toegekend? In dit artikel gaat Bernard Veldkamp in op het wiskundige fundament van onderwijskundig meten. De twee meestgebruikte modellen uit de psychometrie worden geïntroduceerd en hun belangrijkste begrippen, formules en toepassingen worden beschreven. Klassieke testtheorie bestaat het langst en is het meest bekend vanwege de scoringsregel, die het cijfer bepaalt door de punten van de individuele opgaves op te tellen. Item-responstheorie is minder bekend, maar wordt met name gebruikt bij grootschalige toetsen en examens zoals eindexamens en inburgeringsexamens. De voor- en nadelen van beide modellen worden benoemd. Tot slot wordt kort ingegaan op actuele uitdagingen voor het vakgebied van de psychometrie.

De centrale eindexamens zijn weer achter de rug. Tienduizenden scholieren zwoegden op examens die van grote invloed zijn op het vervolg van hun carrière. De examens vormen niet alleen de afronding van het voortgezet onderwijs, maar het diploma dat ermee wordt verkregen geeft toegang tot vervolgonderwijs, het wordt gebruikt om studenten te selecteren voor specifieke programma's en er wordt bij sollicitaties vaak gebruik van gemaakt om een kandidaat te beoordelen op geschiktheid. De behaalde cijfers worden bovendien niet alleen gebruikt om de prestaties van leerlingen te beoordelen. De gemiddelde scores die leerlingen behalen op toetsen en examens worden in de praktijk gebruikt om docenten en scholen te beoordelen. Schoolleiders gebruiken gemiddelde cijfers bij jaargesprekken en de Inspectie van het Onderwijs onderzoekt of de cijfers die behaald worden op het centraal schriftelijk examen niet te veel afwijken van de cijfers op de schoolexamens. Ten slotte gebruiken bestuurders en politici de cijfers om een indruk te krijgen van het algemene onderwijsniveau en van de prestaties van Nederlandse leerlingen ten opzichte van leerlingen uit andere landen. Vanwege dit grote civiele effect is het belangrijk dat de cijfers die leerlingen behalen op examens een betrouwbare weergave zijn van werkelijke vaardigheden van de leerlingen.

Bij het geven van cijfers voor toetsen of examens spelen verschillende aspecten een rol. Allereerst moet beoordeeld worden of en in hoeverre een leerling de individuele vragen op de juiste manier heeft beantwoord. Bij multiplechoicevragen kan deze beoordeling geautomatiseerd plaatsvinden. Bij open vragen gebeurt deze beoordeling door een of meer beoordelaars aan de hand van een beoordelingsvoorschrift. Vervolgens worden de scores op de individuele vragen, al dan niet gewogen, bij elkaar opgeteld. De resulterende score wordt vergeleken met een standaard of cesuur, die aangeeft hoeveel punten behaald moeten worden voor een voldoende beoordeling, en op basis van deze vergelijking wordt een cijfer toegekend. Zowel bij het wegen van de verschillende vragen, bij het optellen van de (gewogen) scores, als bij het vergelijken van de behaalde score met de cesuur, speelt wiskunde een belangrijke rol.

De tak van wiskunde die gebruikt wordt bij toetsen en examens staat bekend als *psychometrie*. Formeel gesproken is psychometrie de wetenschap die zich bezighoudt met de technieken van het meten van psychologische fenomenen zoals kennis, vaardigheden, attituden, eigenschappen en persoonskenmerken. Voor het meten van kennis en vaardigheden wordt gebruikgemaakt van toetsen en examens, terwijl attitude, eigenschappen en persoonskenmerken gemeten worden met psychologische testen.

In dit artikel wordt een beschrijving gegeven van de geschiedenis van de psychometrie. Vervolgens wordt ingegaan op de twee bekendste en meest toegepaste families van modellen, te weten de klassie-

ke testtheorie en de item-responstheorie. Het artikel eindigt met het beschrijven van de meest recente ontwikkelingen binnen de psychometrie.

Klassieke testtheorie

Het meestgebruikte model om cijfers toe te kennen aan toets- en examenresultaten is gebaseerd op de som-correctscore, ook wel totaalscore genoemd. Als dit model wordt toegepast, dan worden bij het nakijken punten gegeven voor elke individuele vraag. Vragen in een toets worden ook wel items genoemd. Vervolgens worden de punten van alle items opgeteld en omgezet in een cijfer. Dit model wordt toegepast bij de meeste toetsen en proefwerken in het reguliere onderwijs. Het model volgt uit de klassieke testtheorie.

Klassieke testtheorie is gebouwd op drie aannames:

$$X = T + E,$$
$$E(E) = 0,$$
$$\rho_{TE} = 0.$$

De eerste aanname geeft aan dat een geobserveerde score X opgebouwd is uit een werkelijke score T (true score) en een ruiscomponent E (error). De geobserveerde score is de som-correctscore, zoals die hierboven is geïntroduceerd. De werkelijke score is het echte niveau van de kandidaat. Dit kun je vergelijken met de gemiddelde score die een kandidaat zou halen als hij de toets een groot aantal keer zou maken, waarbij zijn hersens gespoeld zouden worden na elke poging om te voorkomen dat hij de vragen onthoudt. Hierbij wordt ervan uitgegaan dat alle items even goed bijdragen aan de te meten vaardigheid. Deze werkelijke score is een latente variabele die niet direct geobserveerd kan worden, maar die wordt afgeleid uit de antwoorden. De ruiscomponent is een meetfout die veroorzaakt kan worden door allerlei toevallige omstandigheden en die een correcte meting verstoren. Bij toetsen en examens zijn we er in geïnteresseerd om de werkelijke score T zo nauwkeurig mogelijk te meten.

De tweede aanname geeft aan dat de verwachting van de ruiscomponent gelijk is aan nul. Bij toetsen kunnen er verschillende oorzaken zijn voor ruis. Ruis kan veroorzaakt worden door eigenschappen van de persoon, van de toets, of van de situatie. Vermoeidheid kan ervoor zorgen dat de prestaties van de persoon verminderen of iemand kan een moeilijk item per ongeluk correct beantwoorden. Slechte items kunnen zorgen voor ruis ten gevolge van het meetinstrument en een surveillant die verkouden is en veel moet niesen, is een omgevingsfactor die ruis kan veroorzaken. De consequentie van de tweede aanname is dat de verwachting van de geobserveerde score gelijk is aan de verwachting van de werkelijke score. Dit geeft aan dat, op het moment dat de klassieke testtheorie geldt, we verwachten dat de geobserveerde score een goede weergave is van de werkelijke score.

De derde aanname geeft aan dat er geen correlatie is tussen de ruiscomponent en de werkelijke score van een persoon. Dat wil zeggen dat het niet uitmaakt of een kandidaat juist een hoge vaardigheid of een lage vaardigheid heeft, er is geen relatie met de gemaakte meetfout. Deze drie aannames kunnen gebruikt worden om een aantal afleidingen te maken.

Betrouwbaarheid in klassieke testtheorie

Allereerst kunnen we iets zeggen over de betrouwbaarheid ρ^2_{XT} van een examen. De betrouwbaarheid is gedefinieerd als dat deel van de variantie van de geobserveerde score dat verklaard wordt door de werkelijke score. De betrouwbaarheid kan waardes aannemen van 0 tot en met 1, mits de variantie in de geobserveerde score groter is dan 0. Een betrouwbaarheid van $\rho^2_{XT} = 0$ geeft aan dat er geen relatie is tussen de geobserveerde score en de werkelijke score waar we in geïnteresseerd zijn, oftewel, de score op de toets is volledig gebaseerd op de ruiscomponent. Een betrouwbaarheid van $\rho^2_{XT} = 1$ daarentegen, geeft aan dat de geobserveerde score volledig wordt verklaard door de werkelijke vaardigheid van de persoon, oftewel, de toets heeft geen meetfout. Deze betrouwbaarheid kan geformuleerd worden als:

$$\rho^2_{XT} = \frac{\sigma^2_T}{\sigma^2_X}.$$

Omdat $X = T + E$ en $\rho_{TE} = 0$ kunnen we afleiden dat $\sigma^2_X = \sigma^2_T + \sigma^2_E$. Daarom geldt voor de betrouwbaarheid:

$$\rho^2_{XT} = \frac{\sigma^2_T}{\sigma^2_T + \sigma^2_E}.$$

Deze laatste formule laat zien hoe de ruiscomponent van invloed is op de betrouwbaarheid. Hoe kleiner σ^2_E des te hoger de betrouwbaarheid. Bij het construeren van toetsen en examens wordt er daarom naar gestreefd om de betrouwbaarheid van toetsen en examens te maximaliseren en de ruiscomponent te minimaliseren. In de praktijk is het alleen niet mogelijk om de betrouwbaarheid ρ^2_{XT} uit te rekenen omdat we de variantie van de werkelijke score σ^2_T niet kunnen meten. Daarom zijn er verschillende manieren bedacht om de betrouwbaarheid te kunnen schatten. De drie bekendste methodes zijn test-hertestbetrouwbaarheid, de split-halfmethode en Cronbachs alfa. Bij test-hertestbetrouwbaarheid wordt een toets twee keer afgenomen bij dezelfde populatie en wordt de correlatie tussen de beide afnames genomen als schatting van de betrouwbaarheid. Een mogelijke verstoring hierbij is wel dat kandidaten vragen kunnen onthouden, waardoor deze schatting niet optimaal is. Bij de split-halfmethode wordt de toets verdeeld in twee helften en wordt de correlatie tussen de beiden helften uitgerekend als schatting voor de betrouwbaarheid. Deze methode heeft als nadeel dat de betrouwbaarheid effectief slechts berekend wordt op basis van een halve toetslengte. De derde manier is door gebruik te maken van Cronbachs alfa:

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum_i \sigma^2_i}{\sigma^2_X}\right),$$

waarbij n het aantal items is, σ^2_X de variantie van de totaalscore is en σ^2_i de variantie van de score van item i is. Cronbachs alfa is nog steeds een ondergrens van de betrouwbaarheid, maar het grote voordeel is dat Cronbachs alfa berekend kan worden op basis van bekende varianties.

De meetfout

Naast betrouwbaarheid wordt er vaak gesproken over de meetfout van een toets. Met deze meetfout wordt de wortel van de ruisvariantie bedoeld. Met behulp van de formules voor de betrouwbaarheid, kan de meetfout worden berekend als:

$$\sigma_E = \sigma_X \sqrt{(1 - \rho_{XT}{}^2)}.$$

Hoe kleiner deze meetfout, hoe nauwkeuriger de toets of het examen meet. Een opvallend kenmerk van klassieke testtheorie is dat de meetfout onafhankelijk is van het werkelijke niveau van de kandidaat. Uit de drie aannames van klassieke testtheorie volgt dat de meetfout een eigenschap is

van de toets. Deze meetfout is constant en voor de berekening maakt het niet uit hoeveel kandidaten de toets maken.

De invloed van toetslengte

Wat wel uitmaakt is het aantal items waaruit de toets bestaat. Eggen en Sanders [2] laten zien hoe Spearman en Brown afgeleid hebben dat je op basis van de formule voor de betrouwbaarheid kunt aantonen dat de betrouwbaarheid van een nieuwe toets die k keer zo lang is als de oude toets kunt berekenen met de formule

$$\rho^2_{XT_{\text{nieuw}}} = \frac{k\rho^2_{XT_{\text{oud}}}}{1+(k-1)\,\rho^2_{XT_{\text{oud}}}}.$$

De voorwaarde hierbij is dat de items in de nieuwe toets allemaal van dezelfde kwaliteit zijn als de items uit de oude toets. Om de betrouwbaarheid van de toets te verhogen en de meetfout te verkleinen wordt er in de praktijk daarom vaak voor gekozen om extra items af te nemen. De Cotan (COmmissie Test Aangelegenheden Nederland) heeft een aantal richtlijnen opgesteld voor de betrouwbaarheid. Zo geldt dat de betrouwbaarheid hoger moet zijn dan $0{,}90$ voor examens en voor andere toetsen die gebruikt worden voor belangrijke beslissingen. Voor toetsen die gebruikt worden bij minder belangrijke beslissingen, zoals voortgangstoetsen, geldt dat de betrouwbaarheid hoger moet zijn dan $0{,}80$. Als er slechts op groepsniveau wat gedaan wordt met de uitkomsten van de toetsen geldt dat de betrouwbaarheid $0{,}70$ moet zijn. Door de lengte van de toets te variëren kan geprobeerd worden om aan deze richtlijnen te voldoen.

Door de toetslengte te variëren, kan de betrouwbaarheid beïnvloed worden. In de vorige paragraaf is al opgemerkt dat een van de uitgangspunten daarbij is dat de kwaliteit van de oude en nieuwe items vergelijkbaar is. In de praktijk blijkt dit zelden het geval te zijn. De strategie die dan meestal gehanteerd wordt, is dat kwalitatief mindere items uit de toets verwijderd worden om op die manier de betrouwbaarheid te verhogen. Daarbij wordt gekeken naar de moeilijkheid van de items en naar het onderscheidend vermogen. De moeilijkheid van een item wordt berekend met het percentage correcte antwoorden op een vraag, ook wel p-waarde genoemd. Voor het onderscheidend vermogen van een item wordt als maat vaak de r_{it}-waarde gebruikt, dat is de correlatie tussen de score op het item en de toets in zijn geheel. Voor een hoge betrouwbaarheid is het van belang dat er voldoende spreiding zit in de moeilijkheid van de items en dat de toets bestaat uit items met een hoog onderscheidend vermogen ($r_{\text{it}} \geq 0{,}40$).

Validiteit

De betrouwbaarheid zegt iets over de meetkwaliteit van de toets of het examen. Bij een hoge betrouwbaarheid, meet de toets consistent. Maar meet de toets ook wat hij moet meten? Kun je testscores gebruiken om conclusies te trekken over de vaardigheid waarvoor de toets ontworpen en afgenomen is? Om die vraag te beantwoorden moet gekeken worden naar de validiteit. Bij validiteit spelen verschillende aspecten een rol. Messick [8] beschrijft dat traditioneel gezien er vooral gekeken wordt naar inhoudsvaliditeit, criteriumvaliditeit en begripsvaliditeit. Inhoudsvaliditeit zegt iets over de formulering van de items: is de inhoud duidelijk verankerd in de leerdoelen en is er een evenwichtige verdeling van de vragen of de leerdoelen? Criteriumvaliditeit zegt iets over hoe goed de toetsscore voorspellend is voor een extern criterium, zoals toekomstige prestaties of een tweede onafhankelijke meting met een ander instrument. Begripsvaliditeit, ten slotte, geeft aan hoe goed de verschillende items een representatie zijn van het onderliggende construct dat je eigenlijk wilt meten met de toets. De traditionele manier van validiteitsonderzoek vergeet alleen mee te nemen hoe de toetsscore in de praktijk gebruikt wordt, aldus Messick [8]. In recenter onderzoek naar validiteit [13], wordt validiteit dan ook veel meer gekoppeld aan geschiktheid voor een specifiek doel. Door de validiteit van een toets of examen te onderzoeken, kunnen ten slotte uitspraken gedaan worden over hoe de toetsscore gebruikt kan worden om uitspraken te doen over de kandidaten.

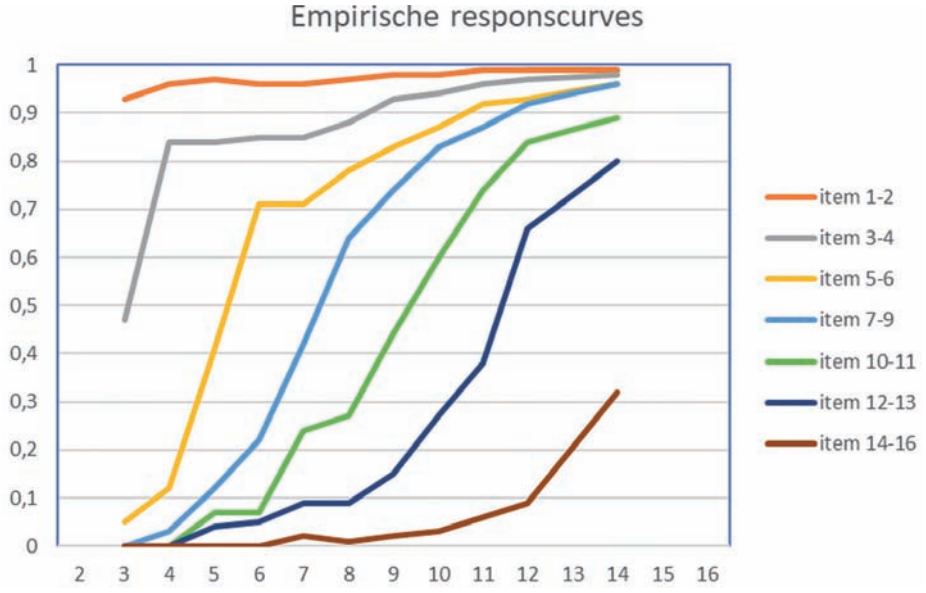
Beperkingen van de klassieke testtheorie

De klassieke testtheorie, zoals die hierboven kort is beschreven, vormt het fundament onder het gebruik van een totaalscore of som-correctscore bij het beoordelen van de resultaten van leerlingen bij toetsen en examens. Dit is alleen niet het hele verhaal. Bij de meeste belangrijke toetsen en examens, zoals de eindexamens voortgezet onderwijs, bij de inburgeringsexamens of bij de eindtoetsen basisonderwijs wordt geen gebruikgemaakt van deze methode. De reden hiervoor is dat de klassieke testtheorie een aantal grote nadelen [5] heeft. Alle indices en scores die ermee berekend worden zijn gekoppeld aan de toets en de populatie waarmee ze zijn berekend. De betrouwbaarheid van een toets wordt bijvoorbeeld berekend voor een specifieke populatie. Op het moment dat de toets bij een andere populatie, zoals een andere klas, een andere school, of leerlingen uit een ander leerjaar wordt afgenomen, moet de betrouwbaarheid opnieuw worden berekend. Een tweede voorbeeld betreft de score op de toets. Als twee leerlingen verschillende toetsen maken, dan kunnen hun scores niet onderling worden vergeleken. Om met dit soort bezwaren om te kunnen gaan is er veel onderzoek gedaan naar mogelijkheden om de indices en score te kunnen generaliseren. Door te werken met parallelle toetsen [2], kunnen de resultaten van examens en herexamens, bijvoorbeeld, worden vergeleken. Ondanks al het onderzoek naar deze mogelijkheden, bleef er veel kritiek op de klassieke testtheorie. Daarom is item-responstheorie ontwikkeld.

Item-responstheorie

Item-responstheorie ontstond gelijktijdig in de Verenigde Staten en in Europa in een poging om tot betere modellen te komen om de antwoorden van de kandidaten aan eigenschappen van toetsen en van individuele items te koppelen. Daarvoor zijn een aantal modellen ontwikkeld, die de kans dat een kandidaat een item correct beantwoordt modelleren als functie van een latente vaardigheid en een of meer eigenschappen van de items. Die vaardigheid wordt latent verondersteld omdat hij niet direct te observeren is, maar geschat moet worden uit de antwoorden die de kandidaat geeft.

Lord [6] omschreef, voor het eerst, het concept van een item-karakteristieke curve, een grafiek die de relatie tussen de vaardigheid van een kandidaat en de kans op een correct antwoord weergeeft. In Europa werkte de Deense wiskundige Georg Rasch onafhankelijk aan hetzelfde idee. Rasch (1960) bestudeerde grafieken waarin hij de kandidaten opdeelde in categorieën op basis van hun totaalscore. Hij ordende de vragen op basis van hun moeilijkheid en keek wat voor elk van de categorieën de kans was op een correct antwoord. Een



Figuur 1 Empirische item-responscurves Rasch.

voorbeeld van een dergelijke grafiek wordt gegeven in Figuur 1.

In Figuur 1 zijn items gerangschikt van makkelijk naar moeilijk en ingedeeld in zeven moeilijkheidscategorieën. Voor deze toets van zestien items hebben de kandidaten totaalscores gehaald die variëren van 2 tot 14. Het is duidelijk te zien dat Rasch bij het ontwikkelen van zijn modellen in eerste instantie nog werkte binnen het raamwerk van de klassieke testtheorie, omdat hij nog werkte met totaalscores. De grafiek laat zien wat de kans was dat leerlingen met een bepaalde totaalscore items binnen een bepaalde groep correct beantwoorden. Zo kan uit de grafiek afgelezen worden dat leerlingen met een totaalscore van 8 een kans hadden van 0.09 om bijvoorbeeld item 12 correct te beantwoorden.

Verschillende item-responstheoriemodellen

Rasch-model. Bij het bestuderen van deze grafieken, maakte Rasch onderscheid tussen de vaardigheden van kandidaten en de moeilijkheden van items, als twee onafhankelijke grootheden die van invloed waren op de kans dat een kandidaat een item correct beantwoordde. Met betrekking tot de curves die ontstonden, constateerde Rasch drie dingen:

1. De curves zijn niet-lineair.
2. De curves snijden elkaar niet.
3. De curves voor de moeilijke items stijgen minder snel dan de curves voor de makkelijke items.

Op basis van deze constateringen ging Rasch op zoek naar een model dat uitgaande van het niveau van de kandidaat en de moeilijkheid van de items een accurate voorspelling doet van de kans dat de betreffende kandidaat de vraag correct beantwoordt. Hij deed daarbij twee aannames. De aanname van unidimensionaliteit houdt in dat de kans om het item correct te beantwoorden slechts beïnvloed wordt door de vaardigheid die de toets beoogt te meten. In de praktijk zullen meestal ook andere vaardigheden, persoonlijkheid en omstandigheden waaronder de toets wordt afgenomen van invloed zijn op deze kans, maar die worden in het model niet meegenomen. De tweede aanname van lokale onafhankelijkheid houdt in dat de kans dat een kandidaat een item correct beantwoordt niet afhankelijk is van het responsgedrag op andere items in de toets. Oftewel, twee items zijn alleen aan elkaar gerelateerd via de invloed van de onderliggende vaardigheid die ze meten. Als het ene item bijvoorbeeld aanwijzingen bevat voor het correct beantwoorden van andere items, dan wordt de aanname van lokale onafhankelijkheid geschonden. Een ander voorbeeld van een schending is dat meerdere items gekoppeld kunnen zijn aan een tekst of grafiek. In die gevallen wordt de afhankelijkheid vaak expliciet mee gemodelleerd om aan de tweede aanname te blijven voldoen.

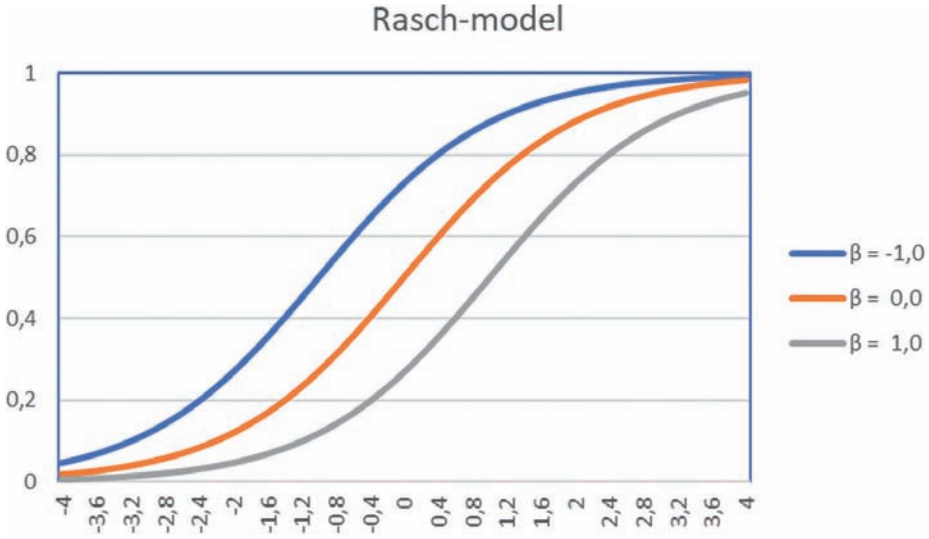
Om het niet-lineaire verband te modelleren gebruikte Rasch een logistische linkfunctie, waarbij de relatie tussen twee variabelen gemodelleerd wordt als:

$$Y = \frac{e^X}{1+e^X}$$

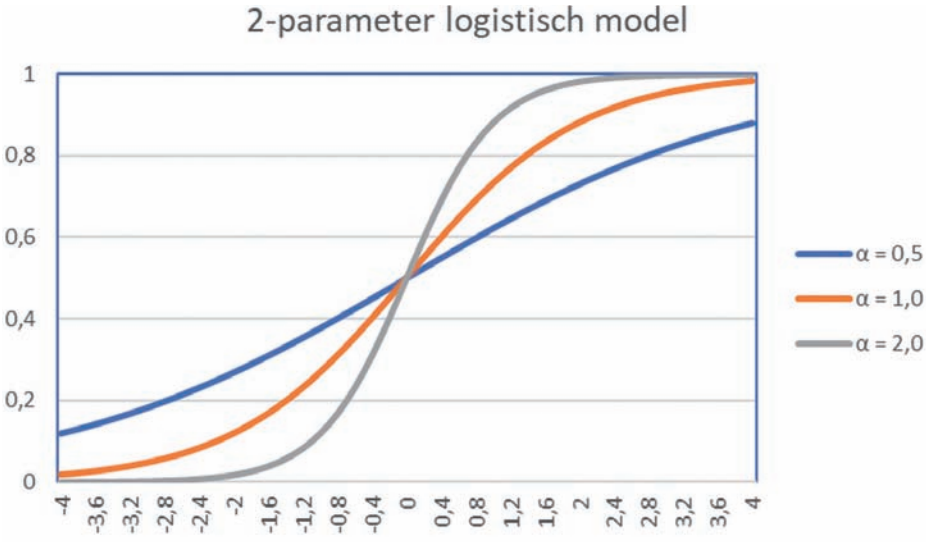
Rasch modelleerde de kans dat een kandidaat een correct antwoord gaf op een vraag ($X = 1$), afhankelijk van de moeilijkheid (β) van item i en de vaardigheid (θ) van kandidaat j, als:

$$P((X=1) \mid \theta, \beta) = \frac{e^{(\theta-\beta)}}{1+e^{(\theta-\beta)}}.$$

De vaardigheid is daarbij gedefinieerd als een latente variabele die het beheersingsniveau van de kandidaat weergeeft, waarbij $-\infty < \theta < \infty$. De moeilijkheid is gedefinieerd als een locatieparameter die aangeeft bij welke vaardigheid de kandidaat een kans heeft van $p = 0{,}50$ om de vraag correct te beantwoorden ($-\infty < \beta < \infty$).



Figuur 2 Item-karakteristieke curves Rasch-model.



Figuur 3 Item-karakteristieke curves voor het 2-parameter logistisch model.

Is de vaardigheid van de kandidaat hoger dan de moeilijkheid, dan is de kans op een correct antwoord groter dan $0{,}50$, als de vaardigheid lager is dan de moeilijkheid, dan is de kans kleiner dan $0{,}50$. Dit Rasch-model wordt ook wel het 1-parameter logistisch model genoemd, omdat de eigenschappen van het item gemodelleerd worden met één parameter, namelijk de moeilijkheid. Figuur 2 geeft de item-karakteristieke curves weer voor items met moeilijkheden $\beta_1 = -1$, $\beta_2 = 0$, $\beta_3 = 1$. Voor een kandidaat met een vaardigheid gelijk aan $\theta = 1{,}5$ is de kans op een correct antwoord op het item gelijk aan respectievelijk $0{,}92$; $0{,}82$; $0{,}62$.

Kenmerkend voor het Rasch-model is dat de item-karakteristieke curves allemaal dezelfde vorm hebben en dat ze alleen verschillen in locatie. Dit is een vrij strakke eis waaraan in de praktijk lang niet altijd wordt voldaan.

Het 2-parameter logistisch model. Om wat meer flexibiliteit aan te brengen in item-responstheoriemodellen, werd het 2-parameter logistisch model ontwikkeld [7]. Naast de moeilijkheidparameter (β) kent dit model ook een discriminatieparameter (α). Deze parameter geeft aan hoe goed een item onderscheid kan maken tussen kandidaten die een vaardigheid hebben lager dan de moeilijkheid en de kandidaten die een vaardigheid hebben hoger dan de moeilijkheid. Het 2-parameter logistisch model kan geformuleerd worden als:

$$P((X=1) \mid \theta,\alpha,\beta) = \frac{e^{\alpha(\theta-\beta)}}{1+e^{\alpha(\theta-\beta)}}.$$

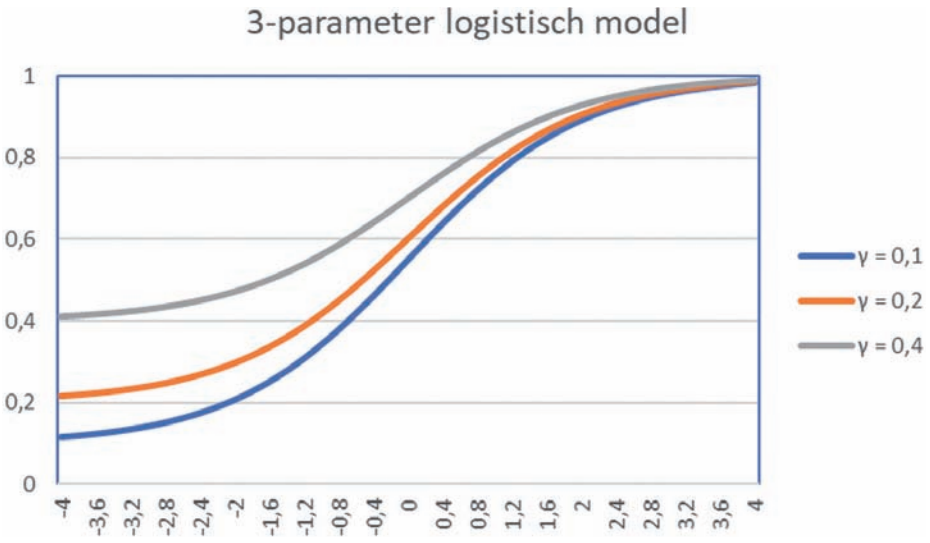
In Figuur 3 worden de karakteristieke curves weergegeven van drie items met discriminatieparameters $0{,}5$; $1{,}0$; $2{,}0$ waarbij de moeilijkheid van alle drie de items gelijk is aan $\beta = 0{,}0$.

In Figuur 3 is te zien dat het verschil in kans op een goed antwoord tussen kandidaten met een vaardigheid $\theta = -0{,}4$ en kandidaten met een vaardigheid $\theta = 0{,}4$ voor items met een laag onderscheidend vermogen ($\alpha = 0{,}5$) slechts gelijk is aan $0{,}10$, terwijl het voor items met een hoog onderscheidend vermogen ($\alpha = 2{,}0$) gelijk is aan $0{,}39$. Items met een hoog onderscheidend vermogen kunnen dus veel beter onderscheid maken tussen kandidaten met een vaardigheid lager dan de moeilijkheid en kandidaten met een vaardigheid hoger dan de moeilijkheid. Deze items leveren daarom meer informatie over de kandidaten.

Het 3-parameter logistisch model. Een tweede uitbreiding heeft te maken met de mogelijkheden voor een kandidaat om het correcte antwoord te raden. Als een toets bestaat uit multiplechoicevragen, dan heeft een kandidaat een kans groter dan nul om goed te gokken, op het moment dat hij of zij een van de antwoorden aanvinkt. Om hiervoor te corrigeren is het 2-parameter logistisch model uitgebreid met een pseudo-gokparameter (γ). Deze parameter geeft aan dat iedere kandidaat een kans heeft gelijk aan γ om het item correct te beantwoorden, ongeacht zijn of haar vaardigheid. Het 3-parameter logistisch model wordt geformuleerd als:

$$P((X=1) \mid \theta,\alpha,\beta,\gamma) = \gamma + (1-\gamma)\frac{e^{\alpha(\theta-\beta)}}{1+e^{\alpha(\theta-\beta)}}.$$

Op het moment dat een multiplechoice-item vier antwoordcategorieën heeft, zou je verwachten dat de pseudo-gokparameter de waarde aanneemt $\gamma = 0{,}25$. Deze schatting houdt er alleen geen rekening mee dat de verschillende antwoordcategorieën niet allemaal even waarschijnlijk zijn. Het gevolg hiervan is dat de pseudo-gokparameter verschillende waardes aan kan nemen, zelfs al is het aantal antwoordcategorieën vergelijkbaar. Figuur 4 laat de item-karakteristieke curves zien van items van een pseudo-gokparameter gelijk aan $\gamma_1 = 0{,}1$; $\gamma_2 = 0{,}2$; $\gamma_3 = 0{,}4$, terwijl de moeilijkheidsparameters en discriminatieparameters gelijk zijn aan $\beta = 0{,}0$ en $\alpha = 1{,}0$.



Figuur 4 Item-karakteristieke curves voor het 3-parameter logistisch model.

In Figuur 4 is goed te zien hoe de pseudogokparameter fungeert als een onderasymptoot voor de kans op een correct antwoord. Het Rasch-model, het 2-parameter logistisch model en het 3-parameter logistisch model zijn alle drie gemodelleerd met een logit-linkfunctie. Al vanaf de beginjaren van item-responstheorie is er ook gebruikgemaakt van equivalente modellen die gebaseerd waren op een probit-linkfunctie.

Polytome item-responstheoriemodellen. Beide soorten modellen kunnen gebruikt worden om de kans op een correct antwoord te berekenen voor een gegeven vaardigheidsniveau en bekende itemparameters. Hierbij moet wel opgemerkt worden dat deze modellen ervan uitgaan dat een item correct of incorrect wordt beantwoord. Dit worden ook wel dichotome items genoemd. Als een item ook gedeeltelijk correct beantwoord kan worden, of als een kandidaat meerdere punten kan behalen voor een item, dan is er sprake van een polytoom item en kan er gebruikgemaakt worden van polytome item-responstheoriemodellen [9]. Een voorbeeld hierbij is de vraag:

$$\sqrt{\frac{7{,}5}{0{,}3}-16}=\ldots?$$

Om tot het correcte antwoord te komen, moet een kandidaat zowel de deling, het aftrekken, als het worteltrekken correct uitvoeren. Als kandidaten ook punten krijgen voor het correct oplossen van deelstappen, is er sprake van een polytoom item. Daarnaast is het mogelijk dat er meerdere vaardigheden nodig zijn om een vraag correct te beantwoorden. Een voorbeeld hierbij zijn wiskundeopgaven waarbij kandidaten de benodigde informatie uit een tekst moeten halen. Naast de vaardigheid wiskunde is ook begrijpend lezen nodig om tot een goed antwoord te komen. Een overzicht van multi-dimensionale item-responstheoriemodellen en hun toepassingen wordt gegeven in Reckase [11].

Informatiefuncties en meetfout

Binnen de item-responstheorie wordt betrouwbaarheid vervangen door het concept informatie. Hoe meer informatie een item geeft over de vaardigheid van een kandidaat, des te nauwkeuriger de vaardigheid geschat kan worden. Om uit te rekenen hoeveel elk item bijdraagt, wordt gebruikgemaakt van statistische informatietheorie [1]. Fishers informatiefunctie is een van de manieren om te berekenen hoeveel informatie een geobserveerde variabele (het antwoord van de kandidaat) geeft over een latente variabele (de vaardigheid), als de kans op deze geobserveerde variabele afhangt van deze latente variabele. Voor de verschillende modellen kan de informatiefunctie $I_i(\theta)$, die laat zien hoe de informatie van item i afhankelijk is van de vaardigheid θ, uitgerekend worden als:

$$I_i(\theta)=\frac{\left[\frac{\partial P_i(\theta)}{\partial\theta}\right]^2}{P_i(\theta)\left(1-P_i(\theta)\right)},$$

waarbij $P_i(\theta)$ staat voor de item-responsefunctie $P((X=1)\mid\theta,\beta)$ van item i. Iteminformatiefuncties voor de verschillende item-responstheoriemodellen kunnen gevonden worden in bijvoorbeeld Embretson en Reise [3]. Deze informatiefuncties hebben twee nuttige eigenschappen. Allereerst kan de informatie voor een complete toets, de toetsinformatiefunctie (TIF), uitgerekend worden door de informatie van de verschillende items op te tellen:

$$TIF(\theta)=\sum_{i=1}^{N}I_i(\theta).$$

Daarnaast kan de meetfout uitgerekend worden met:

$$SE(\theta)=\frac{1}{\sqrt{TIF(\theta)}}.$$

Schatten vaardigheid en itemparameters

Er zijn verschillende schattingsmethoden ontwikkeld om de vaardigheid (θ) en de itemparameters (α,β,γ) te schatten. Al deze methoden gaan uit van de likelihood, oftewel de waarschijnlijkheid van de geobserveerde antwoordpatronen. Voor een toets van N items die beantwoord is door J personen, waarbij de antwoorden van persoon j op item i genoteerd wordt als u_{ij} ($u_{ij}=1$ voor een correct antwoord en $u_{ij}=0$ voor een incorrect antwoord), en de kans dat persoon j een correct antwoord geeft op item i als P_{ij} kun je afleiden dat:

$$P_{ij}^{u_{ij}}(1-P_{ij})^{1-u_{ij}}=\begin{cases}P_{ij} & \text{als } (u_{ij}=1),\\ 1-P_{ij} & \text{als } (u_{ij}=0).\end{cases}$$

Voor $u_{ij}=1$ geldt immers dat $P_{ij}^{u_{ij}}=P_{ij}^{1}=P_{ij}$ en voor $u_{ij}=0$ geldt dat $P_{ij}^{u_{ij}}=P_{ij}^{0}=1$. Daarmee kan de likelihood geformuleerd worden als:

$$L(u_{11},u_{12},\ldots,u_{NJ}\mid\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma})=\prod_{i=1}^{N}\prod_{j=1}^{J}P_{ij}^{u_{ij}}(1-P_{ij})^{1-u_{ij}},$$

waarbij ($\boldsymbol{\theta}$) de vector met vaardigheidsparameters weergeeft en ($\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}$) de matrix met itemparameters. Deze likelihood is de vermenigvuldiging van de kansen op correcte (P_{ij}) of incorrecte ($1-P_{ij}$) antwoorden voor alle items in de toets en alle kandidaten die meegedaan hebben. Als bijvoorbeeld de vragen 1 tot en met 5 op de volgende manier beantwoord worden door kandidaat 3: (correct, correct, incorrect, incorrect, correct) en P_{ij} op dezelfde manier gedefinieerd is als hiervoor, ziet de likelihood, hier genoteerd als $L(\cdot)$ er uit als:

$$L(\cdot)=P_{13}P_{23}(1-P_{33})(1-P_{43})P_{53}.$$

Bij het schatten van de vaardigheids- en de itemparameters wordt gezocht naar parameterwaardes die deze likelihood optimaliseren. Joint Maximum Likelihood-schatters, die proberen om tegelijkertijd zowel de vaardigheden als de itemparameters te schatten, leveren helaas inconsistente schatters op. Alternatieve methodes hiervoor zijn Marginal Maximum Likelihood-schatters en Conditional Maximum Likelihood-schatters. De Marginal Maximum Likelihood-schatters schatten eerst de itemparameters door de vaardigheidsparameters uit de likelihoodvergelijking te integreren, waarbij ervan uitgegaan wordt dat de vaardigheden van de hele populatie standaardnormaal verdeeld zijn. Na het schatten van itemparameters kunnen de vaardigheidsparameters geschat worden met een Joint Maximum Likelihood-schatting. De Conditional Maximum Likelihood-schatters gebruiken een vergelijkbare aanpak waarbij eerst de itemparameters geschat worden, conditioneel op de vaardigheid. Vervolgens worden de vaardigheidsparameters van de individuele kandidaten apart geschat. Voor een uitgebreide beschrijving van deze schattingsmethodes, zie Eggen en Sanders (1993). Naast deze beide frequentistische methodes, kunnen de vaardigheden- en de itemparameters ook geschat worden met een Bayesiaans algoritme, waarbij een priorverdeling voor zowel de vaardigheden als de itemparameters wordt aangenomen. Voor een overzicht van Bayesiaanse IRT-methodes wordt verwezen naar Fox [4]. Zowel voor de maximum likelihood-methodes als voor de Bayesiaanse methode kan

gebruikgemaakt worden van standaard opensource-softwarepakketten.

Omdat de vaardigheid en de itemparameters los van elkaar geschat worden, werken de bovenstaande schattingsmethodes ook bij missing data, dat wil zeggen dat ze ook werken als de kandidaten maar een deel van de items hebben beantwoord. Voorwaarde is wel dat deze missing data niet afhangt van de vaardigheid van de kandidaten. Het mag dus niet komen omdat de kandidaat vragen die hij of zij te moeilijk vindt over heeft geslagen. Deze eigenschap wordt gebruikt om verschillende toetsen of examens onderling vergelijkbaar te maken. Beide toetsen kunnen op dezelfde schaal gebracht worden door ze te equivaleren. Hiervoor is nodig dat een deel van de kandidaten, zowel de vragen van de eerste als van de tweede toets maken. Op die manier krijg je een gelinkt design. Twee manieren om dit te organiseren worden weergegeven in Figuur 5, waarbij de rijen verschillende groepen kandidaten weergeven en de kolommen de verschillende items. De blauwe gedeeltes geven de items aan die de verschillende groepen beantwoorden, de grijze gedeeltes staan voor items die niet aan de betreffende groep worden aangeboden.

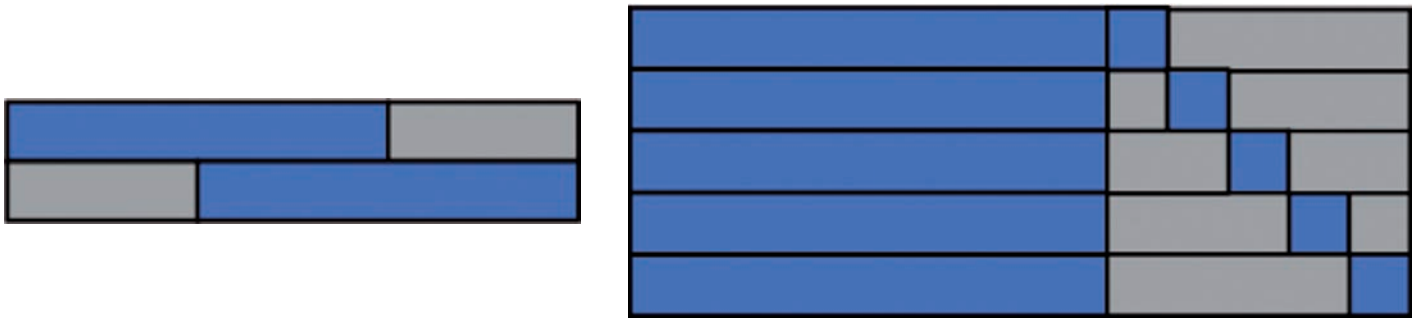
In het design van Figuur 5(a) worden twee toetsen aan elkaar gelinkt doordat twee groepen kandidaten een deel van de items gezamenlijk hebben. Allereerst worden de itemparameters van de eerste toets geschat op basis van de eerste groep kandidaten. Daarmee wordt de schaal vastgelegd. De itemparameters van de items die beide groepen gemeenschappelijk hebben hoeven niet opnieuw geschat te worden voor de tweede groep. De ontbrekende itemparameters van de tweede toets worden vervolgens op dezelfde schaal geschat als de items waarvan de parameters al bekend zijn. Dit gebeurt door deze itemparameters te fixeren. De parameters van de eerste toets en de tweede toets liggen nu op dezelfde schaal waardoor de toetsscores vergelijkbaar zijn. Bij het design in Figuur 5(b) maken vijf verschillende groepen de hele eerste toets. Daarnaast maken ze allemaal een deel van de tweede toets. De itemparameters van de eerste toets zijn identiek voor alle vijf de groepen. De vaardigheidsverdeling van de vijf groepen liggen daarmee op dezelfde schaal. Die schaal wordt vervolgens gebruikt om de itemparameters van de tweede toets op dezelfde schaal te schatten. Hiermee worden de cijfers op de tweede toets vergelijkbaar met die op de eerste toets.

Op deze manier kan er bijvoorbeeld voor gezorgd worden dat de eindexamens van verschillende jaren onderling vergelijkbaar zijn. Voorafgaand aan de examenperiode worden daarvoor items van het nieuwe examen samen met een aantal items van het oude examen uitgeprobeerd bij een kleine groep leerlingen. De informatie die dit oplevert, wordt gebruikt om de nieuwe items te kalibreren op dezelfde schaal. Voor het nieuwe eindexamen worden items gekozen die voldoen aan de randvoorwaarden voor de inhoud, maar die gezamenlijk ook een testinformatiefunctie hebben, die vergelijkbaar is met de examens van de jaren ervoor. Op deze manier worden beide examens geëquivaleerd.

Voor- en nadelen van item-responstheorie

Item-responstheorie heeft een groot aantal voordelen. Allereerst wordt de vaardigheid van de kandidaten nauwkeuriger geschat, waarbij rekening gehouden wordt met verschil in moeilijkheid en onderscheidend vermogen van de items. De schatting van de vaardigheid is onafhankelijk van de moeilijkheid van de toets en van de gebruikte items. Bovendien kan de informatiefunctie gebruikt worden om een individuele schatting van de meetfout te berekenen. Ook met betrekking tot de itemparameters kent item-responstheorie grote voordelen. Door deze parameters separaat te schatten krijgen we veel meer inzicht in de kwaliteit van de individuele items. De itemparameters kunnen bovendien onafhankelijk van de steekproef van kandidaten geschat worden. Daarnaast liggen de itemmoeilijkheid en de vaardigheid van de kandidaat op dezelfde schaal, wat veel inzicht geeft. Het gebruik van item-responstheorie kan daarom leiden tot eerlijkere en accuratere toetsing. Ten slotte maakt item-responstheorie het mogelijk om toetsen en examens adaptief af te nemen. Dat houdt in dat tijdens de afname van een toets of examen al een inschatting van de vaardigheid van de kandidaat gemaakt wordt en dat de moeilijkheid van de items afgestemd wordt op het niveau van de kandidaat. Dit leidt tot kortere toetsen en voorkomt frustratie vanwege het moeten beantwoorden van veel te makkelijke of veel te moeilijke items.

Naast deze voordelen heeft item-responstheorie ook een aantal nadelen. De vaardigheid wordt geschat op een standaardnormaal verdeelde schaal. De schattingen lopen daarmee van $-4{,}0$ tot $4{,}0$ met een gemiddelde van $0{,}0$. In tegenstelling tot wat we in Nederland gewend zijn met de klassieke som-correctscore, is deze schaal bovendien niet lineair. Voor de meeste leerkrachten en leerlingen is daarom een vertaling nodig naar een schaalscore. Dit gebeurt door middel van een standaard-setting procedure. In deze procedure werken inhoudsdeskundigen en psychometrici samen om te bepalen bij welke vaardigheidsscore een kandidaat voldoende score voor de test. Deze vaardigheidsscore krijgt de waarde $5{,}5$ op de scoreschaal. Vervolgens wordt een tabel ontwikkeld die de te behalen vaardigheidsscores ver-



Figuur 5 Voorbeelden van gelinkte designs.

taalt naar een schaalscore die loopt van 0 tot 10. Een positieve uitzondering hierbij is het Rasch-model, waarvan aangetoond is dat je daarbij wel de som-correctscore kunt gebruiken [2]. Een tweede nadeel is dat de steekproef voldoende groot (minstens een paar honderd kandidaten) moet zijn om de itemparameters nauwkeurig genoeg te kunnen schatten. Daardoor is item-responstheorie alleen toepasbaar bij grotere toetsen en examens. Een derde nadeel, ten slotte, is dat de item-responsmodellen wel een goede fit moeten laten zien bij de responsdata. Als de geobserveerde scores erg afwijken van de item-karakteristieke curves, dan is er sprake van misfit en zijn item-responsmodellen niet toepasbaar.

Conclusie

Toetsen en examens nemen een belangrijke plaats in binnen het Nederlandse onderwijs. Geschat wordt dat leerkrachten en leerlingen gemiddeld 30% van hun tijd hieraan besteden. De cijfers die de leerlingen krijgen, hebben daarnaast veel invloed op hun schoolcarrière en hun kansen op een baan. Het is daarom van groot belang dat de cijfers een betrouwbaar beeld geven van de vaardigheden van de kandidaten. In dit artikel is een korte inleiding gegeven in de psychometrie. Deze relatief onbekende tak van wiskunde is ontwikkeld om een fundament te leggen onder het meten in het onderwijs. Psychometrie heeft zich in de afgelopen honderd jaar ontwikkeld tot een rijk vakgebied en omdat er steeds meer opensource-software beschikbaar komt, is het ook toepasbaar voor een groot publiek.

Psychometrie is overigens niet exclusief ontwikkeld voor het onderwijs, al heeft de grootschalige toepassing binnen de examinering er wel een enorme boost aan gegeven. Een tweede belangrijke toepassing ligt in het analyseren van vragenlijstdata. Binnen de psychologie en de gezondheidszorg wordt bijvoorbeeld op grote schaal gebruikgemaakt van Likert-items, waarbij respondenten op een schaal van één tot drie, één tot vijf of één tot zeven alternatieven, aan moeten geven in hoeverre ze het met een uitspraak eens zijn. Concrete voorbeelden van toepassingen in de psychometrie zijn vragenlijsten die depressie meten of intelligentietesten. Met name de klassieke testtheorie wordt hierbij veel gebruikt, terwijl sinds de eeuwwisseling item-responstheorie ook binnen de psychologie en gezondheidszorg steeds vaker wordt toegepast.

Een van de uitdagingen waar de psychometrie op dit moment voor staat is dat er steeds meer data en anderssoortige data beschikbaar komt over kandidaten. Terwijl de modellen in de psychometrie met name gericht zijn op het analyseren van antwoorden, komt er steeds meer informatie beschikbaar over het leerproces. Online leersystemen leggen in logfiles exact vast wat de verschillende stappen zijn die de kandidaat doorlopen heeft om tot een antwoord te komen. Ook kan de tijd die de kandidaat nodig gehad heeft bruikbare informatie geven over zijn of haar vaardigheid. Van der Linden [12] stelde al voor hoe item-responstheorie modellen uitgebreid kunnen worden om responstijden mee te kunnen modelleren. Maar de vraag hoe logfilegegevens gemodelleerd moeten worden en de vraag hoe je (wiskundig) kunt verantwoorden dat deze data gebruikt wordt vaardigheden of groei te schatten, liggen nog open.

Referenties

1 T.M. Cover en J.A. Thomas, *Elements of Information Theory,* John Wiley & Sons, 2012.

2 T.J.H.M. Eggen en P.F. Sanders, *Psychometrie in de praktijk*, Cito Instituut voor Toetsontwikkeling, 1993.

3 S.E. Embretson en S.P. Reise, *Item Response Theory for Psychologists*, Lawrence Earlbaum Associates, 2000.

4 J.P. Fox, *Bayesian Item Response Modeling: Theory and Applications,* Springer Science & Business Media, 2010.

5 R.K. Hambleton, H. Swaminathan en H.J. Rogers, *Fundamentals of Item Response Theory*, Vol. 2, Sage, 1991.

6 F.M. Lord, A theory of test scores, *Psychometric Monographs* (1952).

7 F.M. Lord en M.R. Novick, *Statistical Theories of Mental Test Scores*, Addison-Wesley, 1968.

8 S. Messick, (Meaning and values in test validation: The science and ethics of assessment, *Educational Researcher* 18(2) (1989), 5–11.

9 R. Ostini en M.L. Nering, *Polytomous Item Response Theory Models*, No. 144, Sage, 2006.

10 G. Rasch, *Studies in Mathematical Psychology: I. Probabilistic Models for Some Intelligence and Attainment Tests*, 1960.

11 M.D. Reckase, *Multidimensional Item Response Theory*, Springer, 2009.

12 W.J. van der Linden, A hierarchical framework for modeling speed and accuracy on test items, *Psychometrika* 72(3) (2007), 287.

13 S. Wools, T.J. Eggen en A.A. Béguin, Constructing validity arguments for test combinations, *Studies in Educational Evaluation* 48 (2016), 10–18.