



Validiteit

**Wat verstaan moet
worden onder
'validiteit' is nog altijd
geen uitgemaakte zaak**

De eindtoets basisonderwijs

Over de gevolgen van een gebrek aan consensus over validiteit

■ Karen Heij

Dr. Karen Heij is na haar studie Toegepaste Taalwetenschap werkzaam geweest bij Bureau ICE, van 2002 tot 2016 als Algemeen Directeur. Op 21 april jl. is zij als buitenpromovendus van de Nederlandse School voor Openbaar Bestuur gepromoveerd aan de TU op haar onderzoek naar de eindtoets basisonderwijs ('Van de kat en de bel, tellen en vertellen met de eindtoets basisonderwijs'), te bestellen of te downloaden via de site www.parrhesiaonderwijsadvies.nl E-mail: karenheij@gmail.com

Validiteit is het belangrijkste aspect voor het vaststellen van kwaliteit van toetsing. Een eenduidige definitie ervan ontbreekt echter. Dat maakt validiteit enerzijds cruciaal en anderzijds lastig te beoordelen. Dit gebrek aan consensus biedt toetsontwikkelaars ruimte te kiezen voor een eigen interpretatie en invulling en dat heeft weer gevolgen voor hoe anderen kwaliteit zien en waarderen zoals blijkt uit de eindtoets basisonderwijs.

Validiteit

In de Internationale 'Standards for Educational and Psychological Testing' (2003) wordt validiteit van toetsing gedefinieerd als 'the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests'. Deze definitie is te typeren als een gulden middenweg op basis van de vele discussies die in de afgelopen eeuw gevoerd zijn over hoe validiteit te definiëren en te verantwoorden. Ook dit 'compromis' heeft de discussie niet kunnen verstommen. Dus wat verstaan moet worden onder 'validiteit' is nog altijd geen uitgemaakte zaak. En zo is er sprake van een patstelling die een goede beoordeling van kwaliteit van testen in de weg staat. Aan de hand van de eindtoets basisonderwijs kijken we naar waar dat in de praktijk toe leidt.

Waar zit de angel in het debat?

De meest gebruikte en voorkomende definitie van validiteit is in de praktijk van alledag nog altijd de definitie die door Kelley in 1927 naar voren is gebracht: Meet de toets daadwerkelijk wat hij beoogt te meten? De charme van de eenvoud van de definitie neemt niet weg

dat velen na Kelley gepleit hebben voor een bredere kijk op validiteit waarbij vooral de nadruk is gelegd op de geschiktheid van een instrument gegeven een bepaalde context. Vanuit die optiek is validiteit geen vaste eigenschap of kenmerk van een meetinstrument maar ontleent een test zijn geldigheid aan de toepassing in een bepaalde context. Een test zelf kan vanuit dat perspectief niet het predicaat 'valide' of 'niet-valide' krijgen. Anderen bepleiten dat de sociale consequenties van testgebruik als een apart aspect van validiteit onderscheiden moet worden waarbij de vraag 'Who gets hurt, who gets helped, and is it worth it?' centraal staat. Zij spreken hierbij over 'consequential validity' (Nichols & Berliner, 2007).

Als bij toetsing veel op het spel staat voor zowel winnaars als verliezers wordt er gesproken van 'high stakes' toetsen. Zeker bij deze categorie toetsen is aandacht voor wat er op het spel staat en wat de mogelijke gevolgen zijn, een belangrijk onderwerp van discussie. Hierbij is de vraag naar de mogelijkheden en beperkingen die toetsing hebben voor

het nemen van dergelijke, zwaarwegende beslissingen van belang. Madaus stelt dat de introductie van een 'high stakes test' op voorhand bevestigd moet worden op 'wie er wint en wie er verliest'. Hij pleit ervoor te zorgen dat uit de evaluatie van het gebruik van tests naar voren komt welke individuen en groepen benadeeld of juist bevoordeeld worden (Madaus, 1994). Daarmee bepleit Madaus een veel bredere interpretatie van een onderzoek naar de waarde en waardering van toetsen dan gangbaar is en door testwetenschappers in het kader van validiteit wordt bepleit. Madaus vraagt naar 'who gains and who loses' is bijzonder relevant om te beantwoorden bij de inzet van een test die grote gevolgen heeft voor de toekomstkansen van kinderen zoals in het geval van de eindtoets basisonderwijs. Als tests ingezet worden om bepaalde maatschappelijke of onderwijskundige doelen dichterbij

te brengen, is het relevant een aantal vragen, die in het kader van de waardering van de effecten van een testprogramma relevant zijn, vooral op voorhand te onderzoeken. Madaus pleit voor de beantwoording van een viertal vragen alvorens high stakes toetsing op grote schaal in te voeren:

- Leidt de inzet van de test(s) tot een onderwaardering van dat wat niet in de test aan bod komt?
- Hoe zullen potentiële risico-leerlingen de gekozen inhoud van de test ervaren?
- Wat zullen de gevolgen zijn voor hen die laag scoren?
- Heeft de uitkomst gevolgen voor het volgen van vervolgonderwijs? Zo ja, wat zijn de gevolgen van het gebruik van de tests in relatie tot kanselijkheid?

Terwijl wetenschappers zich bezighouden met discussies over de definities van validiteit, de reikwijdte ervan of juist het omzeilen van het begrip door met andere vragen te kijken naar de (lange termijn) effecten van toetsing, wordt in de praktijk van alledag doorgewerkt met 'eigen' onderbouwingen van validiteit. Het gebrek aan heldere kaders biedt toetsmakers maar ook beleidsmakers ruimte om op nationaal niveau of institutioneel niveau zelf te kiezen voor definities en werkwijzen. Zo bepalen ze niet alleen zelf of en in hoeverre ze meegaan met de verschillende stromingen in het validiteitsdebat maar zetten daarmee ook de toon voor hoe in het publieke en maatschappelijke debat tegen validiteit van toetsing en daarmee tegen verantwoording van de kwaliteit ervan wordt aangekeken.

Wat gebeurt er in de praktijk in Nederland?

Onderzoek naar definiëring en praktische toepassing van het begrip validiteit in de Nederlandse context laat zien dat hier vooral de meer smalle opvatting over validiteit het denken en handelen beheerst. Hier overheerst de mening dat consequenties die samenhangen met het gebruik van tests buiten de scope van



Als bij toetsing veel op het spel staat wordt er gesproken van 'high stakes' toetsen

validiteitsonderzoek vallen. In de publicatie 'Toetsen op school' stelt Cito dat het bij validiteit gaat om de vraag of de toetsscores de informatie bieden waarin we geïnteresseerd zijn en of op basis van deze scores de gewenste beslissingen genomen kunnen worden. Daarbij wordt ervoor gekozen te werken met wat genoemd wordt 'validiteitsbewijzen'. De maat 'betrouwbaarheid' wordt daarbij gezien als één van de bewijzen van validiteit (Cito, 2013). De COTAN, de Commissie Testaangelegenheden Nederland, die onder andere is belast met de beoordeling van andere eindtoetsen en LVS-toetssystemen, hanteert bij beoordeling van voorgelegde toetsen onder andere de validiteit. Validiteit definieert COTAN als 'de mate waarin een test aan zijn doel beantwoordt: kan men uit de testcores die conclusies trekken die men op het oog heeft?' (Evers e.a., 2009). COTAN beoordeelt op drie aspecten van validiteit: inhoudsvaliditeit, begripsvaliditeit en criteriumvaliditeit.

Alle eindtoetsen basisonderwijs, behalve de Centrale Eindtoets, worden beoordeeld door de Expertgroep Toetsen PO. Zij hanteert, naast het oordeel van de COTAN, een eigen beoordeling waarbij validiteit gepositioneerd wordt als een aspect van de inhoudelijke kwaliteit van een eindtoets (Van der Vleuten e.a., 2020). De onderbouwing van de validiteit van een eindtoets vraagt een koppeling van de opgaven van de toets aan de domeinbeschrijvingen voor taal en rekenen zoals beschreven in het Referentiekader Taal en Rekenen en opgenomen in de 'Toetswijzer voor eindtoets taal en rekenen' (CvTE, 2014). Toetsopgaven moeten op inzichtelijke en verantwoorde wijze gekoppeld zijn aan de domeinen en de onderdelen zoals daar beschreven en ook moet worden aangetoond dat de items in de toets bijdragen aan de validiteit van de toets waarbij het gaat het om aspecten als relevantie, objectiviteit en efficiëntie van de items.

Wat opvalt voor de uitwerking en toepassing van validiteit hier is dat het begrip 'consequen-

tial validity' geen onderdeel lijkt uit te maken van de praktijk van toetsverantwoording.

De Centrale Eindtoets en selectie

De Centrale Eindtoets is de variant van de eindtoets die door de overheid beschikbaar wordt gesteld in het kader van de wettelijke verplichting om aan het eind van het basisonderwijs een eindtoets af te nemen en door Cito ontwikkeld wordt onder regie van het College van Toetsen en Examens. Deze toets kent geen openbaar beoordelingskader. Uit de jaarlijkse rapportages en tussentijdse verantwoordingen is alleen indirect af te leiden hoe de validiteit van deze eindtoets inhoudelijk, als construct en als criterium wordt onderbouwd. Voor de onderbouwing van de validiteit van de gekozen inhoud van de eindtoets wordt al sinds 1968 min of meer dezelfde formulering gebruikt: 'Resultaten van voorafgaand onderwijs geven meestal zeer waardevolle aanwijzingen over wat er in daarop volgend onderwijs van een leerling kan worden verwacht' (Samenwerkende Instituten, 1968). In 2015 onderbouwt CvTE de inhoudelijke validiteit op vergelijkbare wijze: 'Het lijkt gerechtvaardigd te veronderstellen dat leerprestaties uit het verleden iets zeggen over leerprestaties in de toekomst.' Daar voegt CvTE aan toe dat 'verschillen in leerprestaties op de onderdelen van de Centrale Eindtoets vermoedelijk voor een belangrijk deel kunnen worden verklaard door verschillen in eigenschappen als 'intelligentie', 'concentratie', 'motivatie' en doorzettingsvermogen'. Via de Centrale Eindtoets worden deze eigenschappen indirect en afgewogen gemeten.' (CvTE, 2015).

Het beoogde, te meten construct, wint niet aan duidelijkheid met deze formulering. Want waar gaat het nu om in de eindtoets? Gaat het nu om leerprestaties? Om schoolgeschiktheid? Om een combinatie van bovengenoemde aspecten? Of gaat het nu om het niveau van taal en rekenen? Los van de vraag wat nu precies het te meten construct is, is het relevant te bevragen hoe valide de aanname is dat



Wat maakt iemand nu specifiek geschikt voor VMBO-T?

leerprestaties uit het verleden, en dan op het gebied van taal en rekenen, bepalend zullen zijn voor toekomstig functioneren. Dit geldt ook voor het begrip 'schoolgeschiktheid'. Hoe is dat voor een bepaalde onderwijsvorm vast te stellen als er van de verschillende vormen van vervolgonderwijs geen profielen beschikbaar zijn en al zeker niet gebaseerd op degelijk onderzoek? Wat maakt iemand nu specifiek geschikt voor VMBO-T? In de praktijk blijkt een ordening op basis van de gecombineerde score op een set taal- en rekenopgaven bepalend voor het advies en blijkbaar daarmee de geschiktheid. De onderbouwing die hiervoor gegeven wordt, is gebaseerd op correlaties: correlaties tussen score en de positie na één of drie jaar in het voortgezet onderwijs. Maar daarmee is de vermeende juistheid of onjuistheid van de voorspelling niet te onderbouwen. Het is immers niet waar te nemen wat er zou zijn gebeurd als er een ander onderwijstraject

was aangeboden. Noch is duidelijk wat nu precies verklaart dat een leerling na één of drie jaar nog wel of niet langer in het geadviseerde spoor zit. Een correlatie is geen verklaring en van oorzaak en gevolg hoeft dan ook zeker geen sprake te zijn.

De Centrale Eindtoets en kwaliteit

Hoewel het geven van onafhankelijke informatie over de keuze voor het best passend vervolgonderwijs volgens de verantwoordingsdocumenten van Cito/CvTE het primaire doel van de toets is, is ook het verbeteren van de kwaliteit van onderwijs een belangrijk doel. De wet spreekt in dit verband over de eindtoets als 'hulpmiddel voor de zelfevaluatie van scholen' en van een 'instrument voor de externe verantwoording en beoordeling van de leerresultaten van scholen'. Als toelichting op dat laatste wordt gesteld dat de Centrale Eindtoets de functie heeft 'van hulpmiddel om geobjectiveerd te kunnen nagaan of de eindopbrengsten van deze vaardigheden op schoolniveau voldoen aan het bij de wet vastgestelde minimum qua onderwijsopbrengsten c.q. leerresultaten' (Ministerie OCW, 2012). Daarmee is de eindtoets wettelijk gezien selectietoets en resultaatmeting in één. Dat betekent dat er zowel sprake is van een normgerelateerde beoordeling als van een criteriumgerelateerde beoordeling waarbij de eerste functie primair is wat gevolgen heeft voor de constructie en selectie van opgaven. Want voor het selectiedoeleinde worden vooral opgaven geselecteerd die een goed 'discriminerend vermogen' hebben (p-waarde tussen de .40 en .90 met een gemiddelde van .70). De beide functies van toetsing staan niet alleen toetstechnisch op gespannen voet met elkaar, ze staan een scherpe blik op de validiteit van de eindtoets in de weg. Dit wordt extra 'gehinderd' door het feit dat het exacte doel van de eindtoets in relatie tot het algehele doel van onderwijs niet duidelijk omschreven is. De vraag, relevant in het kader van validiteit, of de eindtoets in zijn aard en uitwerking het doel van onderwijs bevordert of juist belemmert, is daardoor niet

te beantwoorden. Toetsontwikkelaars krijgen ruimte om zich te beperken tot dat wat in het kader van hun beoordeling aangetoond moet worden (de 'andere' eindtoetsen) of tot wat zij zelf relevant vinden om aan te tonen of kunnen aantonen (Centrale Eindtoets). Gezien de grote gevolgen die op het spel staan voor 'winnaars' en 'verliezers' is het van belang de toets en de uitkomsten ervan te zien in het licht van de maatschappelijke ordening die ermee tot stand gebracht wordt. Toetsen zoals de eindtoetsen zijn beleidsinstrumenten en dienen als zodanig beoordeeld te worden door degenen die deze instrumenten willen inzetten en door hen die hier op moeten toezien. Gezien de zorg over de toenemende kansengelijkheid is het uitermate relevant de inzet van een eindtoets te beschouwen op de mate waarin die inzet bijdraagt aan het verbeteren van kansengelijkheid. Het is meer dan zorgwekkend dat deze vraag geen antwoord kent.

Tot slot

Concluderend kan worden gesteld dat onderzoek naar de validiteit van de eindtoets ernstig wordt bemoeilijkt en bedreigd door het feit dat de eindtoets basisonderwijs meerdere doelen moet dienen. Daarmee is nog onduidelijker geworden wat nu het construct is dat gemeten moet worden en waar de vermeende voorspellende waarde aan ontleend wordt. In de praktijk is er vooral ruimte ontstaan voor toetsontwikkelaars (en hun beoordelaars) om eigenstandig te besluiten hoe zich te verantwoorden over wat zij zien als de validiteit van de eindtoets en daarmee ook te bepalen hoe anderen naar de toets en de uitkomsten kijken. Het feit dat de validiteit van een toets met grote gevolgen voor alle betrokkenen nooit goed is onderzocht en niet eenduidig beantwoord kan worden, is een ernstige zaak. Gezien de belangen die met de eindtoets gepaard gaan en de geconstateerde afname van kansengelijkheid wordt het tijd de validiteit van de eindtoets in het licht van deze aspecten grondig te bevragen. Madaus biedt daarvoor een nuttig startpunt. ■

Literatuur

- American Educational Research Association, American Psychological Association, Joint Committee on Standards for Educational, Psychological Testing (US), & National Council on Measurement in Education. (1985). Standards for educational and psychological testing. American Educational Research Association.
- College voor Toetsen en Examens. (2015). Verantwoording Centrale Eindtoets PO. Utrecht: auteur.
- Evers, A., Lucassen, W., Meijer, R., & Sijtsma, K. (2009). COTAN beoordelingssysteem voor de kwaliteit van tests (geheel herziene versie). Amsterdam: NIP.
- Expertgroep Doorlopende Leerlijnen Taal en Rekenen. (2008). Over de drempels met taal en rekenen: Hoofdrapport van de Expertgroep Doorlopende Leerlijnen Taal en Rekenen (p. 107). Enschede: Stichting Leerplan Ontwikkeling.
- Madaus, G. (1994). A technological and historical consideration of equity issues associated with proposals to change the nation's policy. *Harvard Educational Review*, 64(1), 76-95)
- Nichols, S. L. ., & Berliner, D. C. (2007). Collateral Damage: How High-Stakes Testing Corrupts America's Schools. Cambridge, Massachusetts. Harvard Education Press.
- Samenwerkende Instituten (1967). Amsterdamse Schooltoetsen: Verslag van het eerste onderzoek 'L.O.-Schooltoets Amsterdam, 1966'. Groningen: J.B. Wolters.
- Sanders, P. (red.) (2017). [2012] Toetsen op school: Voortgezet Onderwijs. Arnhem: Cito.
- Van der Vleuten, C. & Glas, C. (2020). Beoordelingskader Eindtoetsen (pp. 1–45).