



Bij de ontwikkeling van een toets dienen alle invalshoeken uit het ECD-model steeds onderling op elkaar afgestemd te worden



Ontwerp van toetsen volgens de Evidence Centered Design methode

■ Erik Roelofs

Erik Roelofs is wetenschappelijk onderzoeker bij Cito, Arnhem. Email: erik.roelofs@cito.nl

In het onderwijsveld is een groeiende behoefte aan instrumenten die directe bijsturing van het leerproces mogelijk maken. Gestandaardiseerde toetsen bieden hiertoe volgens critici te weinig houvast. Neem als voorbeeld schrijfvaardigheid. Nieuwe vakdidactische en leertheoretische inzichten op dit terrein hebben geleid tot een veel omvattender definitie van schrijfvaardigheid dan tot nu toe in toetsen gebruikelijk was.

Zo is bij schrijven niet alleen taalvaardigheid (grammatica, woordgebruik, zinsbouw), maar ook het redeneren met ideeën en gedachten van belang. Daarnaast wordt de inzet van specifieke strategieën van belang geacht, zoals het inleven in de behoeften van het lezerspubliek en het maken van schema's. Ook regulatievaardigheden worden genoemd, zoals het tussentijds evalueren en reviseren van eigen schrijfproducten. Tot op heden is slechts een fractie van deze vaardigheden in beeld gebracht met behulp van bestaande gestandaardiseerde toetsen. Ook kunnen met de huidige generatie toetsen geen uitspraken worden gedaan over blokkades in de ontwikkeling van de leerlingen, de punten waarop het leerlin-

gen niet lukt een leersprong of leerprogressie te maken.

Behoefte aan toetsen die het leren ondersteunen

Mede ingegeven door deze bezwaren is er binnen verschillende vakgebieden een ontwikkeling gaande waarbij toetsen worden ontwikkeld, die valide uitspraken over leerlingen toestaan maar tegelijkertijd docenten en leerlingen handvatten bieden voor de bijsturing van het leerproces. Diagnostische toetsen vormen hierbij een belangrijke categorie. Een ideale diagnostische toets geeft antwoord op de volgende vragen:

- Heeft de leerling bepaalde (tussen)doelen van een vak of domein bereikt?
- Beheerst de leerling voorwaardelijke kenniselementen, vaardigheden en strategieën? Wat zijn hierin zijn/haar sterke en zwakke punten?
- Welke typen systematische fouten, inefficiënte oplossingsstrategieën of hulpbehoeften verklaren de resultaten? Welke blokkades of ontwikkelingsdrempels zijn er?
- Welke doelen kunnen bij een vervolginventie het best worden nagestreefd? Welke prioritering in doelen moet worden gesteld? Welke vorm en inhoud hebben vervolginventies en wat is de duur en frequentie ervan?

De methode van de Evidence Centered Design, ontwikkeld door Mislevy et al. (2006), is bij uitstek geschikt als basis voor het ontwerpen van diagnostische toetsen. Met name twee kenmerken zijn hierbij van belang.

Systematisch ontwerp vanuit een leerlingmodel: In het ECD-model worden onderdelen van toetsontwerp (deelmodellen) uitgewerkt die alle gebaseerd dienen te zijn op een theorie over hoe leerlingen leren en vorderen op een domein, een zogeheten leerlingmodel. Toetsontwerp is hierbij niet strikt volgtijdelijk. Ontwerpkeuzes kunnen en moeten onderweg worden bijgesteld op grond van verbeterde inzichten. Ook kunnen beperkingen in het verloop van het ontwerp terugvertaald worden in bijvoorbeeld een meer bescheiden leerlingmodel.

Ontwerpen vanuit het perspectief van validiteit. Bij toetsontwerp worden keuzes steeds expliciet

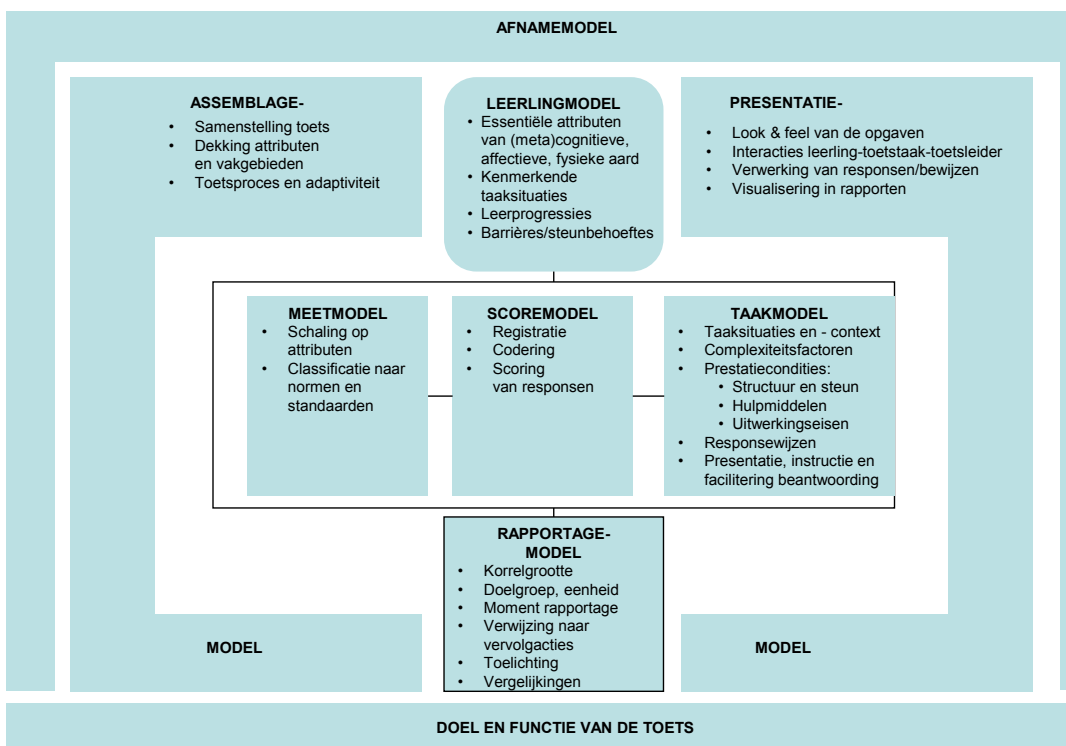
verbonden met de geldigheid van uitspraken die we over leerlingen kunnen doen. Dit betekent dat wordt gestreefd naar de best mogelijke dekking van de te meten vaardigheden en beoogde taaksituaties (inhoudsvaliditeit), het dekken van het gewenste criteriumgedrag (constructrelevantie), het voorkómen van storende factoren in de opzet, afname en antwoordscore van de opgaven (constructirrelevantie).

In het vervolg van dit artikel beschrijven we het ECD-raamwerk, dat aan de basis gelegen heeft van een ontwerpmethodiek van opgaven die recentelijk bij Cito is ontwikkeld.

Het Evidence Centered ontwerpmodel

Het Evidence Centered Design model (ECD) is een methode van systematisch ontwerp, waarbij vanuit verschillende invalshoeken wordt gewerkt aan een toets en een toetsrapportage.

In figuur 1 wordt de belangrijkste laag van het



Figuur 1. Het evidence-centered design model (naar Mislevy et al; 2006; Zieky, 2014)

model weergegeven, die van het assessmentontwerp. We merken erbij op dat het afgebeelde model is aangepast ten opzichte van het oorspronkelijke model.

In het aangepaste ECD-model onderscheiden we acht deelmodellen of invalshoeken van toetsontwerp: 1) het leerlingmodel, 2) het meetmodel, 3) het scoremodel, 4) het taakmodel, 5) het rapportagemodel, 6) het assemblagemodel, 7) het presentatiemodel 8) het afnamemodel. De kern van het ECD-model wordt gevormd door de eerste vijf deelmodellen die in figuur 1 samen een kruis vormen. Deze deelmodellen vormen samen het basisstramien van het toetsontwerp. Bij de ontwikkeling van een toets dienen alle modellen steeds onderling op elkaar afgestemd te worden. De meetpretenties, zoals verwoord in het leerlingmodel, dienen waargemaakt te worden in passende taken (opgaven), een daarbij passend scoremodel en meetmodel. Bovendien moeten resultaten betekenisvol en doelmatig worden gerapporteerd. Beperkingen op een van de onderdelen betekenen dat conclusies over leerlingen (de meetpretenties) moeten worden ingeperkt. Hieronder lichten we de deelmodellen nader toe.

Het **leerlingmodel** beschrijft de aard van het te meten aspect. Dat kan een kenniselement, een vaardigheid, een attitude, een strategie of een combinatie daarvan zijn, zoals een competentie. In het model dient naar voren te komen welke (mentale) processen zich afspelen wanneer leerlingen taken op een domein of vakgebied uitvoeren, zoals bij het schrijven van een tekst. Het leerlingmodel levert een geordend overzicht van de essentiële benodigde kenniselementen, vaardigheden, strategieën en attitude-elementen waarover de toets moet gaan rapporteren. Al deze aspecten vatten we samen onder de meer algemene term 'attribuut'. Behalve de inhoud van het attribuut beschrijft een leerlingmodel in welke taaksituaties leerlingen het attribuut inzetten, welke leerprogressies leerlingen doormaken bij de verwerving ervan en welke barrières of blokkades daarbij moeten worden overwonnen. In de

praktijk is het startpunt van een toets vaak een verzameling van (soms nog vage) onderwijsdoelen, samengevat onder rubrieken of domeinen. Vaak moeten de doelen uit verschillende rubrieken eerst worden herschikt en ondergebracht in een logisch bouwwerk van attributen. Attributen vatten de essentie van wat leerlingen moeten beheersen samen in een afzonderlijk begrip, zoals 'woordgebruik' en 'cohesie aanbrengen tussen zinnen' bij schrijven. De attributen kunnen vervolgens weer verder worden onderverdeeld naar deelattributen en subattributen, naargelang het gewenste detailniveau. Uiteindelijk rapporteert de toets over de mate van beheersing van attributen, op grond waarvan docenten en leerlingen vervolgvactiteiten kunnen starten. Het uitwerken van een leerlingmodel is het belangrijkste voorwerk voor iedere vorm van opgaveconstructie.

Het **taakmodel** beschrijft in het algemeen de sleutelkenmerken van de voor te leggen toetstaken, die nodig zijn om observaties te kunnen doen naar wat de leerling denkt en doet, zoals dat is vastgelegd in het leerlingmodel. Het taakmodel biedt een blauwdruk voor het soort opgaven dat wordt voorgelegd, samengevat in een aantal kenmerken:

(a) Kenmerken van de taaksituaties en taakcontext van de voor te leggen toetstaak. Bijvoorbeeld voor schrijven kan de context variëren van levensechte hier-en-nu taken met een realistisch publiek tot contextvrije taken. De taaksituaties dienen te worden gekarakteriseerd naar bijvoorbeeld onderwerpen, tekstdoel, tekstgenre, en presentatievorm van de tekst.

(b) Factoren die de complexiteit van een taak bepalen. Hierbij moet gedacht worden aan alle factoren die het verloop van een mentaal (of ander) proces gemakkelijker of moeilijker kunnen maken. Bij schrijven kan dit gaan over het publiek en de nabijheid tot de leefwereld. Zo is schrijven voor leeftijdsgenoten over alledaagse zaken doorgaans eenvoudiger dan schrijven over afstandelijke zaken voor een onbekende lezer (bijvoorbeeld naar een campingbeheerder). Andere voorbeelden uit de context van luistervaardigheid in een tweede taal: het gebruik van

verkortingen of accenten, spreektempo, het bezigen van spreek- of schrijftaal.

(c) Prestatiecondities waaronder een leerling de toetstaak uitvoert. Hierbij gaat het om de gegeven steun en structuur, de beschikbare hulpmiddelen en gereedschappen, en de eisen aan de uitwerking van het antwoord of het product. Zo kan een schrijftaak ondersteund worden door gebruik van vooraf afgedrukte sleutelwoorden en een (online) dictionary. Door te variëren in lengte-eisen kan de taak ook moeilijker of gemakkelijker worden gemaakt, afhankelijk van wat wordt beoogd.

(d) De manier waarop leerlingen moeten responderen. Hierbij zijn een aantal basisvormen mogelijk: verbaal, per formaal, visueel, symbolisch. Daarnaast is het onderscheid open-gesloten aan de orde.

(e) De wijze waarop de taken worden gepresenteerd, de gegeven instructies, en de facilitering van de beantwoording. Dit kenmerk heeft betrekking op hoe de interacties tussen de leerling en de toetstaak eruit gaan zien. Kortom, hoe zorgen we ervoor dat de leerling kan laten zien wat we van hem willen zien op het niveau dat wordt beoogd? Bijvoorbeeld, hoe presenteren we een luistertaak? Met of zonder beeld? Hoe stellen we een leerling in staat een fragment terug te luisteren? Hoe wordt hulp vormgegeven? Veel van deze technische vertalingen zijn aan de orde in het presentatiemodel (zie hieronder)

Het **scoremodel** bevat regels over hoe uitingen, handelingen en (tussen)producten worden geregistreerd en ontleed in scorebare eenheden, over de codering van die eenheden en over het toekennen van scores aan die gecodeerde eenheden. Hoe gecodeerd en gescoord wordt is afhankelijk van wat in het leerlingmodel is aangeduid als (hoofd-, deel-, of sub-) attributen. Verder kan het gekozen detailniveau uiteenlopen: kleinere eenheden zijn bijvoorbeeld toetsaanslagen, gekozen antwoordopties, en gezette stappen in een wiskundeopgave. Grotere eenheden zijn bijvoorbeeld geraadpleegde webpagina's, alinea's of teksten. Uiteindelijk worden de verkregen responsen op opgaven gekwantificeerd tot scores op

hoofdattributen, deelattributen en subattributen. Bij een adaptieve toets wordt bij voortschrijding door de toets een steeds betere schatting verkregen van de beheersing van een attribuut totdat op een bepaald moment overgegaan kan worden tot de rapportage van een totaalbeeld.

Dit brengt ons op het meetmodel. In het meetmodel, of psychometrisch model, worden waargenomen scores verbonden aan de veronderstelde attributen in het leerlingmodel en de relaties daartussen. Via het meetmodel krijgen we scores op attribuutniveau, aan de hand waarvan iets kan worden gezegd over de beheersing van attributen door de leerling (de vaardigheden en eventueel de onderliggende cognitieve strategieën). Verder wordt de classificatie van toetsscores (en daarmee van leerlingen) in prestatieniveaus ingesteld in het meetmodel. Ook kan de positionering van een leerling op een sequentie van leerprogressies worden bepaald met behulp van het meetmodel, al dan niet binnen bepaalde onzekerheidsmarges. Verder biedt het meetmodel de mogelijkheid de kwaliteit van de afzonderlijke opgaven, of beter nog de gescoorde responsen, te kwantificeren. Belangrijk hierbij is het onderscheidend vermogen van een opgave. Bij classificatie van leerlingen gaat het om de mate waarin de opgave leerlingen met een verschillend beheersingsniveau kan onderscheiden, gebruikmakend van bijvoorbeeld een psychometrisch classificatiemodel.

Het **rapportagemodel**. In het oorspronkelijke ECD-model (Mislevy et al., 2006) is geen sprake van een rapportagemodel. Wij nemen het rapportagemodel op omdat de gewenste rapportages van de toets sturend zijn voor de keuzes die gemaakt moeten worden bij de toetsconstructie. Bovendien, bij een toets die informatief moet zijn voor vervolgstappen in een leerproces, speelt rapportage een cruciale rol.

Een eerste inhoudelijke keuze hierbij is de korrelgrootte van de terug te rapporteren informatie. Krijgen docent en leerling een algemene rapportage over bijvoorbeeld het niveau van schrijfvaardigheid? Of is het rapport verder uitgesplitst naar deelattributen? Bijvoorbeeld bij schrijven: organisatie van ideeën, structuur, en taalgebruik.

Een verfijnde rapportage betekent dat het leerlingmodel en het taakmodel ook meer verfijnd moeten zijn.

Een volgend kenmerk van het rapportagemodel, dat deels samenhangt met het vorige, is het moment van (terug)rapportage: dat kan zijn na iedere opgave of na afloop van een toetsonderdeel of de gehele toets. In het eerste geval ligt toetsing dichtbij oefening met feedback. In het tweede geval zijn oefenen en toetsen meer gescheiden.

Een volgend aspect is het specificeren van vervolgacties. Wat zijn de wenselijke leeractiviteiten, zoals extra oefening en begeleiding, bij een bepaald profiel? Bij dit alles is van belang welke doelgroep (ouders, leerlingen, docent, school, stelsel) en gebruiksfunctie voor ogen staat bij de rapportage en welke eenheid van rapportage daarbij past. De aard van de toelichtende teksten bij de rapportage van scores hangt samen met de bovengenoemde keuzes. Deze moet worden afgestemd op het gewenste detailniveau en de doelgroep gebruikers.

Tot slot wordt in het rapportagemodel besloten over de aard van vergelijkingen: bijvoorbeeld hoe presteert de leerling ten opzichte van een standaard of ten opzichte van een referentiegroep? Hoe scoort de school ten opzichte van een referentiegroep? Verder zijn er beslissingen te nemen over de aard van de informatie: getallen, symbolen, grafieken in relatie tot wat de gebruiker ermee kan en moet.

Behalve de modellen die het basisstramien van het toetsontwerp bepalen zijn in figuur 1 twee omringende modellen afgebeeld, die bepalen hoe de hele toets en het toetsverloop eruit zien, het assemblagemodel en het presentatiemodel.

Het **assemblagemodel** beschrijft de toetssamenstelling en het toetsverloop. Opgaven dienen de beoogde attributen goed te dekken. Beschreven moet worden hoe de 'flow' van de toets eruit ziet, hoe de opgaveselectie verloopt en op welke manier eventueel adaptiviteit is voorzien. Een eerste vorm van adaptiviteit is vertakking. Hierbij krijgt de leerling bij een relatief zwak of inconsistent resultaat een vervolgoets voorgelegd, totdat dui-



Het ontwerpen van een toets verloopt niet strikt serieel

delijk is wat de meer specifieke sterke en zwakke punten van de leerling zijn. Een andere vorm van adaptiviteit is dat de moeilijkheidsgraad van iedere volgende opgave steeds afhangt van de prestatie op voorgaande opgaven en het totaalbeeld wat tot dan toe is verkregen van de leerling.

Het **presentatiemodel** beschrijft hoe de toetsonderdelen technisch werken tijdens de afname. In dit model wordt uitgewerkt (1) hoe opgaven worden gepland en ingezet in de toets (2) in welke 'look and feel' opgaven worden voorgeschied; (3) hoe de interacties tussen leerling, taak en toetsleider (indien van toepassing) bij beantwoording worden geregeld (4) hoe bewijzen worden verzameld, geregistreerd, gescoord, opgeslagen en verwerkt en (5) hoe de rapportages worden gemaakt en gevisualiseerd. Om valide gegevens te verkrijgen is het cruciaal dat de technische vertaling het inhoudelijk toetsontwerp nauwgezet volgt, om ruis in de opgaven te voorkomen.

In Tabel 1 is een voorbeelduitwerking gemaakt van de onderdelen 1 tot en met 4 voor een wiskundeopgave. Leerlingen moeten hierbij een cirkel tekenen met een straal van 4 centimeter. Een presentatievraag hierbij is of het uitmaakt wanneer leerlingen dit doen met behulp van passer en papier of met behulp van een digitale tekentool. Maakt de beantwoordingswijze uit voor de interpretatie en scoring? Betekent een cirkel met een te grote straal van 4,2 cm in de schriftelijke responsevorm hetzelfde als in de digitale vorm?

Een laatste deelmodel van het oorspronkelijke ECD-model betreft het afnamemodel. Hoewel het ook van belang is, ligt dit model wat verder af van de feitelijke ontwerpwerkzaamheden. Het afnamemodel beschrijft processen die dwars door alle andere modellen heen spelen, zoals de keuze van het platform waarop een toets wordt afgenomen, administratieve keuzes en vereisten (zoals veiligheid en dataopslag) en logistieke keuzes en vereisten (zoals toetsmomenten in een schooljaar of eisen aan toetsleiders).

Slotopmerkingen

Het ECD-ontwerpmodel veronderstelt dat verschillende professionals gelijktijdig betrokken zijn bij toetsontwerp, zoals toetsdeskundigen, onderwijskundigen, (vak)didactici, leraren, materie-

deskundigen, psychometrici en ICT-deskundigen. Deze deskundigen zorgen ervoor dat de verschillende onderdelen op elkaar afgestemd blijven. Om valide uitspraken over de leerling te verkrijgen is het zaak bij alle deelmodellen gebruik te maken van de best beschikbare inzichten. Het ontwerpen van een toets verloopt niet strikt serieel. Een leerlingmodel kan na een pilot worden aangescherpt. Ook kunnen technische randvoorwaarden maken dat het ideale taakmodel voorlopig niet kan worden gevolgd. Het idee is dat in verschillende rondes gewerkt wordt aan een toets die het optimum bereikt van wat wenselijk en haalbaar is. ■

Literatuur

- Bennett, R. E. (2010): Cognitively Based Assessment of, for, and as Learning (CBAL): A Preliminary Theory of Action for Summative and Formative Assessment, *Measurement: Interdisciplinary Research & Perspective*, 8(2-3), 70-91.
- Mislevy, R. J., & Haertel, G. (2006). Implications for evidence-centered design for educational assessment. *Educational Measurement: Issues and Practice*, 25, 6–20.
- Zieky, M.J. (2014). An introduction to the use of evidence-centered design in test development. *Psicología Educativa*, 20, 79-87.

Tabel 1. Technische vertaling van een opgave, “teken een cirkel straal van 4 cm”

Onderdeel toetsafname ►	Selectie opgave ▼	Presentatie opgave ▼	Beantwoording opgave ▼	Scoring en schaling responsen ▼
Afnameomgeving ▼	Attribuut, taak-kenmerken, complexiteit	Instructies; taak-kenmerken	Structuur en steun, hulpmiddelen, eisen, responsewijzen	Registratie, codering, scoring,
	Technische Vertaling ▼	Technische Vertaling ▼	Technische Vertaling ▼	Technische Vertaling ▼
Schriftelijk	Gedrukt opgave-boekje	Afgedrukte opdrachtstelling	Responsegereedschap: Passer en ruitjespapier	Handmatige scoring of via optisch leesbaar formulier
Digitaal	Opgaveselectie-module	Digitale tekst opdrachtstelling	Responsegereedschap: Tekentool, digitale liniaal	Automatische scoring via een scoremodule

Lamark toezichthouders geëxamineerd door Exapol

