Onderbouwing

AMN Eindtoets: adaptief met terugbladerfunctie. Hoe zit dat?

In 2017 liet het ministerie van Onderwijs, Cultuur en Wetenschap de AMN Eindtoets officieel toe als eindtoets in het basisonderwijs. De AMN Eindtoets is uniek in zijn soort door zijn adaptiviteit mét terugbladerfunctie. Dat vind je bij geen enkele andere toets.

Maar hoe werkt die adaptiviteit? Hoe betrouwbaar is de inschatting van het niveau van de leerling? En hoe is dat van invloed op het schooladvies?

Dat zijn interessante vragen die wij graag beantwoorden. In dit kennisartikel vertellen de psychometristen en psychologen van AMN hoe de eindtoets in elkaar steekt.

Toelating

Het toelatingsproces is zeer streng. De Expertgroep Toetsen PO en de Commissie Testaangelegenheden Nederland (COTAN) stellen hoge eisen voor de toelating van een eindtoets. Dat is ook meer dan logisch aangezien het vervolgonderwijs van ca. 180.000 kinderen mede afhangt van het resultaat van de eindtoets. Dit resultaat wordt immers als tweede gegeven gebruikt naast het advies van de leerkracht.

De AMN Eindtoets is officieel toegelaten. Dit houdt in dat het een valide en betrouwbare toets is die voldoet aan de hoogste kwaliteitseisen die in Nederland aan testen en toetsen gesteld worden.

Adaptief

Vanaf 2018 is de AMN Eindtoets adaptief. Dit betekent dat de leerling opgaven aangeboden krijgt die nauw aansluiten bij zijn vaardigheidsniveau. De toets maakt een zeer betrouwbare inschatting van het vaardigheidsniveau van de leerling. Om deze inschatting extra kracht bij te zetten krijgt de leerling een aantal extra controlevragen. Wanneer de inschatting zeer betrouwbaar is krijgt de leerling geen opgaven meer aangeboden. Hierdoor varieert de lengte van de eindtoets per leerling. Bij de ene leerling kan sneller een betrouwbare inschatting gemaakt worden van het niveau dan bij de andere leerling.

Een adaptieve toets heeft volgens Linacre (2000) een aantal belangrijke voordelen, waaronder:

- Korter en sneller toetsen.
- Niet iedere leerling krijgt dezelfde opgaven. Hierdoor is de geheimhouding van de opgaven veel beter geregeld.

- Doordat er gewerkt wordt met een grote itembank (‘vragenbank’) is de frequentie waarmee een opgave wordt aangeboden veel lager.
- De inhoud van de toets sluit veel beter aan bij het niveau van de leerling. Bij een te makkelijke of een te moeilijke toets raakt de leerling snel gedemotiveerd. Als de toets wél goed aansluit, blijft de leerling gemotiveerd om de toets te maken. Daardoor wordt de ervaring van de leerling ook positiever.

Er kunnen ook enkele nadelen aan adaptieve toetsen zitten. Zo is de adaptieve toets digitaal van aard. Dit houdt in dat scholen bepaalde voorzieningen moeten treffen (laptops of tablets) om te kunnen toetsen.

Daarnaast hebben andere adaptieve toetsen als nadeel dat de leerling niet terug of vooruit kan bladeren in de toets. Bij een papieren eindtoets kan de leerling een vraag overslaan om er later op terug te komen. Dat geeft wat extra tijd en rust voor de leerling. Bij veel digitale en/of adaptieve toetsen is dat niet mogelijk. Dat kan in de AMN Eindtoets wel. De leerling kan terug naar een vorige vraag om deze te verbeteren of vragen overslaan. Het uiteindelijke resultaat blijft betrouwbaar.

IRT

De AMN Eindtoets maakt gebruik van de Item Response Theory (IRT). Dit is een statistische methode waarbij wiskundige modellen het gedrag van mensen kunnen verklaren en voorspellen. AMN maakt gebruik van het 2-parameter logistisch model van Birnbaum (1968). Dit model voorspelt de kans dat een leerling een vraag goed beantwoord op basis van zijn vaardigheid en de kenmerken van een vraag. Die kenmerken heten ook wel itemparameters.

De vaardigheid van de leerling is niet direct zichtbaar. Deze wordt geschat door het afnemen van een set vragen (of: items). In het 2-parameter logistisch model zijn er, zoals de naam als doet vermoeden, twee parameters die van belang zijn:

- De discriminatieparameter (α-parameter). Deze geeft aan hoe goed een vraag onderscheid kan maken tussen leerlingen.
- De moeilijkheidsparameter (b-parameter). Deze geeft aan hoe moeilijk een vraag is.

Algoritme

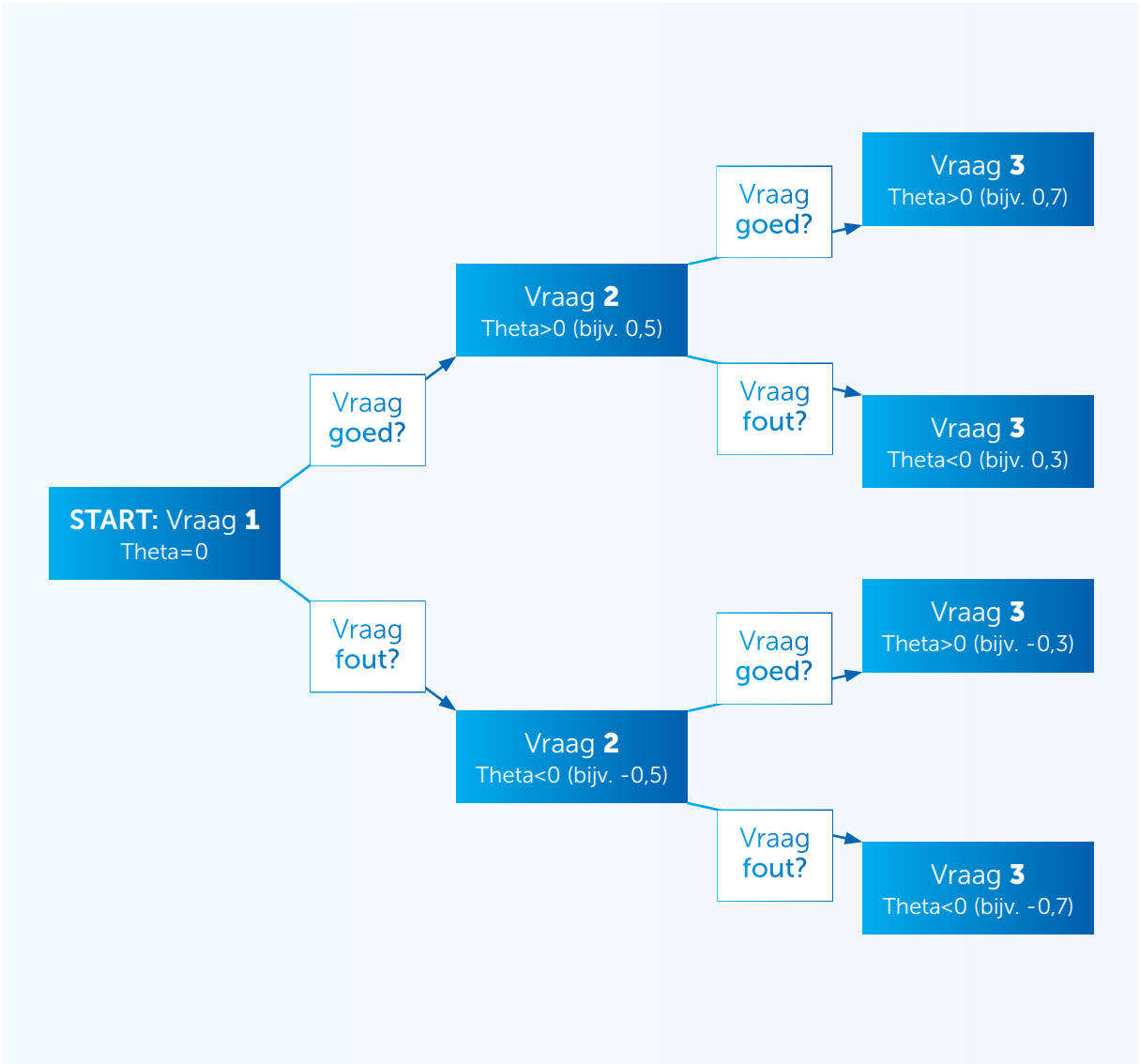
Een adaptieve toets maakt gebruik van een algoritme. Een algoritme beschrijft hoe de toets verloopt en hoe om wordt gegaan met gegeven antwoorden. Bij de AMN Eindtoets zijn twee factoren van belang: de samenstelling van de vragen en de vaardigheid van de leerling.

De vaardigheid van de leerling wordt uitgedrukt in 'theta'. Theta kan tussen de -4 en de 4 liggen. Een waarde van 0 is dus een 'gemiddelde' vaardigheid van de leerling. Dit is tevens het startpunt van een leerling als hij met de eindtoets begint: een theta van 0. Op basis van deze theta selecteert de toets de vraag waar de meeste informatie uitgehaald kan worden over de vaardigheid.

Bij elk antwoord van de leerling rekent de toets opnieuw uit wat de theta van de leerling is (zie ook het stroomschema hieronder). Bij een goed antwoord schat de toets een hogere theta in en 'ziet' dus een hogere vaardigheid van de leerling. Theta is dan groter dan 0. Bij een hogere vaardigheid horen moeilijkere vragen. De toets selecteert dan een nieuwe vraag met een grotere moeilijkheid.

Bij een fout antwoord schat de toets een lagere theta in en 'ziet' dat de leerling een lagere vaardigheid heeft. Op basis van die informatie selecteert de toets een andere vraag met een lagere moeilijkheid.

STROOMSCHEMA



Afbeelding 1: Illustratie over de schatting van de vaardigheid van de leerling

Op die manier komt de toets steeds dichter bij de 'echte' waarde van theta. De vaardigheid van de leerling wordt dus steeds betrouwbaarder ingeschat (de stappen die gemaakt worden zijn kleiner en subtieler dan zoals in de afbeelding staat). Deze betrouwbaarheid wordt na elke vraag opnieuw gecontroleerd. Wanneer de schatting zeer betrouwbaar is krijgt de leerling geen vragen meer. De toets heeft dan een goed en betrouwbaar beeld van de vaardigheid van de leerling.

Het is belangrijk dat de toets genoeg vragen genereert over de te testen stof. In het algoritme zijn regels opgenomen die ervoor zorgen dat dit ook gebeurt. Zo ontstaat er inhoudelijk een representatieve afspiegeling van de te testen stof. Zo komen bijvoorbeeld binnen het onderdeel rekenen alle domeinen en subdomeinen aan bod. Het algoritme zorgt er ook voor dat dit in de juiste verhouding gebeurt.

Daarnaast is er bij het onderdeel rekenen een goede verhouding tussen contextrijke en contextloze vragen. Contextrijke vragen zijn concreet omschreven en bevatten vaak een situatieomschrijving. Denk bijvoorbeeld aan een rekensom in verhaalvorm (redactiesom). Contextloze opgaven zijn niet-talige vragen, maar (bijvoorbeeld) juist kale rekensommen.

Het is belangrijk deze twee soorten vragen in de juiste verhouding aan te bieden aan de leerling. In het dagelijks leven krijgt de leerling namelijk met beide situaties te maken. En daar houdt de lesstof binnen het primair onderwijs ook rekening mee.

Valide

In het algemene deel van de Toetswijzer eindtoets PO (2014) worden eisen gesteld aan de inhoud van de toets. Hierin wordt onder andere beschreven dat de toets minimaal de onderdelen Lezen, Taalverzorging en Rekenen moet meten uit het Referentiekader Taal en Rekenen (2009). Door elke vraag te koppelen aan dat referentiekader is de inhoudsvaliditeit gewaarborgd. Dat houdt in dat de toets daadwerkelijk meet wat hij moet meten. De vragen in de toets meten dus geen andere aspecten dan dat aspect waar ze voor zijn ontwikkeld.

In de toets gaan we uit van drie onderdelen of dimensies: lezen, taalverzorging en rekenen. Uit onze analyses komt naar voren dat deze onderdelen unidimensioneel zijn. Er wordt dus slechts één dimensie gemeten per keer. Zo meet het onderdeel Taalverzorging ook alléén taalverzorging (en dus niet begrijpend lezen of rekenen). Het theoretische model dat we hier hanteren is dus statistisch goed onderbouwd.

Tot slot is er een goede overeenstemming tussen het schooladvies van de leerkracht en de AMN Eindtoets.

Betrouwbaar

In hoeverre zijn de resultaten van een toets een goede afspiegeling van de werkelijkheid? Dat is de betrouwbaarheid van een toets. Deze wordt uitgedrukt met een waarde tussen 0 en 1. We hebben de betrouwbaarheid van de AMN Eindtoets op verschillende vlakken onderzocht:

- De totaalscore van de AMN Eindtoets heeft een betrouwbaarheid van 0,97;
- De betrouwbaarheid van de onderdelen is 0,87 en hoger;
- De onderlinge samenhang (interne consistentie) van de vragen in de vragenbank ligt op 0,96 of hoger (Guttmans lambda 2; Guttman, 1945).

Deze hoge scores laten zien dat de gemeten vaardigheid in de toets sterk overeenkomt met de werkelijke vaardigheid van de leerling.

AMN Eindtoets: objectief en efficiënt

De AMN Eindtoets is digitaal en opgaven worden automatisch gescoord. De antwoorden van de leerlingen worden allemaal vergeleken volgens dezelfde standaarden. Het is onmogelijk dat een opgave de ene keer goed wordt gerekend en de andere keer niet. De toets is daarmee zeer objectief.

De eindtoets meet alleen de wettelijk vereiste domeinen en bevat geen extra onderdelen. De leerling wordt dus niet onnodig belast met extra vragen. Dat geldt ook voor het inschatten van het niveau van de leerling. De toets biedt namelijk niet meer vragen aan dan nodig voor een betrouwbare inschatting. Deze factoren zorgen ervoor dat de toets ook zeer efficiënt is.

LITERATUUR

Birnbaum, A. (1968). Some latent trait models. In: F.M. Lord, & M.R. Novick. Statistical theories of mental test scores (pp. 397-424). Reading: Addison-Wesley

College van Toetsen en Examens (2014). Algemeen deel toetswijzer voor eindtoets po. Inhoudelijke kwaliteitseisen aan eindtoetsen po. Utrecht: College voor Toetsen en Examens.

Expertgroep Doorlopende Leerlijnen Taal en Rekenen (2009). Referentiekader taal en rekenen. De referentieniveaus. Enschede: SLO.

Guttman, L. (1945). A basis for analyzing test-retest reliability. Psychometrika, 10 (4), 255-282.

Linacre, J.M. (2000). Computer-Adaptive Testing: A Methodology Whose Time Has Come. MESA Memorandum number 69, MESA Psychometric Laboratory, University of Chicago.