AMN Doorstroomtoets

Hoe werkt een adaptieve toets?

Begin 2024 doet voor het eerst de doorstroomtoets zijn intrede in het primair onderwijs. De doorstroomtoets vervangt de eindtoets en legt de focus op de doorontwikkeling van de leerling. In 2023 keurde het College voor Toetsen en Examens (CvTE) de AMN Doorstroomtoets officieel goed als doorstroomtoets. De AMN Doorstroomtoets is uniek in zijn soort door zijn adaptiviteit op vraagniveau mét terugbladerfunctie.

Maar hoe werkt adaptiviteit binnen een toetsafname precies, en hoe zorg je voor een betrouwbare inschatting van het niveau van de leerling? Deze en andere vragen worden in deze whitepaper beantwoord door onze orthopedagoog en psychometrist.

Toelating

Het CvTE stelt hoge eisen voor de toelating van een doorstroomtoets. Dat is ook logisch aangezien het vervolgonderwijs van ongeveer 180.000 leerlingen mede afhangt van het resultaat van de doorstroomtoets. Het resultaat van de toets wordt immers door de leerkracht gebruikt om zijn initiële advies eventueel bij te stellen.

Adaptief

Sinds 2018 is de AMN Eindtoets - nu AMN Doorstroomtoets - adaptief. Dit betekent dat de leerling opgaven aangeboden krijgt die nauw aansluiten bij zijn/haar gemeten vaardigheidsniveau. De toets is zelfs adaptief op vraagniveau (zie p. 4). De toets maakt na elke vraag een inschatting van het vaardigheidsniveau van de leerling. Zodra deze inschatting heel betrouwbaar is, krijgt de leerling geen opgaven meer aangeboden. Hierdoor varieert de lengte van de AMN Doorstroomtoets per leerling. Bij de ene leerling kan sneller een betrouwbare inschatting gemaakt worden van het niveau dan bij de andere leerling (zie p. 3). Bij de AMN Eindtoets duurde dit gemiddeld tweeënhalf uur. Bij de AMN Doorstroomtoets zal dit naar verwachting gelijk blijven.

Adaptief toetsen heeft een aantal belangrijke voordelen.

- Korter en sneller toetsen, wat voor meer betrokkenheid en minder toetsangst kan zorgen (Ling et al., 2017).
- Niet iedere leerling krijgt dezelfde opgaven en iedere leerling krijgt slechts een klein deel van de vragen van de totale vragenbank. Hierdoor is de geheimhouding van de opgaven beter gewaarborgd (He & Reckase, 2014).
- Doordat de toets continu het niveau van de leerling inschat en de nieuwe vragen daarop aanpast, sluit de toets (steeds beter) aan bij het niveau van de leerling. Bij een te makkelijke of een te moeilijke toets kan de leerling gedemotiveerd raken. Een toets die wél goed aansluit bij het niveau van de leerling, wordt vaak positiever ervaren. Dit komt de motivatie ten goede (Martin & Lazendic, 2018).
- Doordat iedere leerling andere opgaven krijgt aangeboden, is het risico op afkijken kleiner (Kainz et al., 2015).
- Het vaardigheidsniveau kan preciezer worden geschat voor kinderen met een hogere of lagere vaardigheid dan gemiddeld (Weiss, 2011).

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.
- Bij een papieren doorstroomtoets kan de leerling een vraag overslaan om er later op terug te komen. Dat geeft mogelijk wat extra tijd en rust voor de leerling. Bij veel adaptieve toetsen is dat niet mogelijk. Bij de AMN Doorstroomtoets kan dit wel. De leerling kan terug naar een vraag om deze te verbeteren of een vraag overslaan. Het uiteindelijke resultaat blijft betrouwbaar.

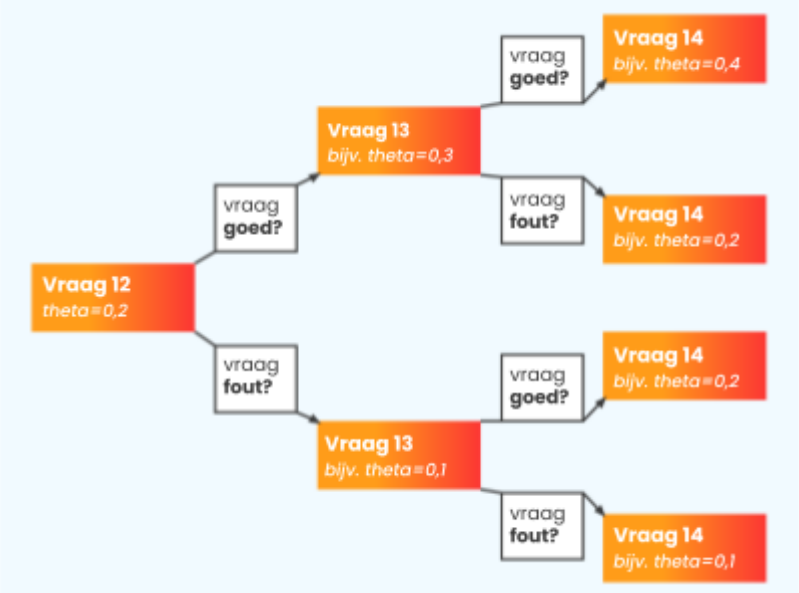
Algoritme

Een adaptieve toets maakt gebruik van een algoritme. Een algoritme beschrijft hoe de toets moet verlopen: hoe de vaardigheid van de leerling moet worden geschat, op welk punt de toets moet stoppen en hoe de combinatie van de toetsvragen moet zijn, zodat alle domeinen bevraagd worden.

De vaardigheid (theta) schatten

De vaardigheid van de leerling wordt uitgedrukt in 'theta'. Theta heeft meestal een waarde tussen -3 en 3. Leerlingen maken eerst een aantal vragen van een gemiddeld niveau. Op basis daarvan wordt de (initiële) theta geschat. Dat is het startpunt van het adaptieve algoritme. Op basis daarvan selecteert de toets een vraag die hierop volgend de meeste informatie kan geven over de vaardigheid.

Bij elk antwoord van de leerling rekent de toets opnieuw uit wat de theta van de leerling is (zie afbeelding). Bij een goed antwoord schat de toets een hogere theta in en daarmee dus een hogere vaardigheid van de leerling. Bij een hogere vaardigheid horen moeilijkere vragen. De toets selecteert dan een nieuwe vraag met een hogere moeilijkheid. Bij een fout antwoord schat de toets een lagere theta in en daarmee een lagere vaardigheid. Op basis van die informatie selecteert de toets een andere vraag met een lagere moeilijkheid.



Afbeelding 1: schematische weergave van de steeds betere inschatting van de vaardigheid van de leerling. Dit is een illustratie; de stappen die gemaakt worden zijn in werkelijkheid subtieler.

Op die manier gaat de toets 'op zoek' naar de vaardigheid van de leerling. Zo komt de toets steeds dichter bij de 'echte' waarde van theta en dus het vaardigheidsniveau van de leerling. De vaardigheid van de leerling wordt dus steeds betrouwbaarder ingeschat. Deze betrouwbaarheid wordt na elke vraag opnieuw berekend. Wanneer de schatting zeer betrouwbaar is, krijgt de leerling geen vragen meer en stopt het desbetreffende onderdeel. Soms wordt de hoge mate van betrouwbaarheid (net) niet bereikt. Om te voorkomen dat de toets vragen blijft aanbieden, is er een maximaal aantal vragen ingesteld. Op het moment dat een leerling het maximaal aantal vragen heeft bereikt, stopt het onderdeel. Ook in die gevallen is de vaardigheidschatting voldoende betrouwbaar.

Alle domeinen in de juiste verhouding

Naast dat in het algoritme is vastgelegd hoe de vaardigheid van de leerling wordt geschat en wanneer de toets moet stoppen, houdt het ook rekening met welke inhoudelijke lesstof aan bod moet komen. Daarom zijn er in het algoritme ook 'regels' opgenomen die beschrijven welke domeinen bevraagd moeten worden en in welke verhoudingen. Zo ontstaat er inhoudelijk een representatieve afspiegeling van de te toetsen lesstof. Je wilt bijvoorbeeld niet dat een leerling 21 vragen uit het domein 'meten en meetkunde' krijgt en maar 1 vraag uit het domein 'getallen'. Daarom is in het algoritme voor ieder subdomein vastgelegd hoe groot het aandeel moet zijn, bijvoorbeeld "het percentage toetsvragen uit het domein 'getallen' ligt tussen de 30% en 40% van het totaal aantal rekenvragen". Zo zijn er ook regels met betrekking tot opgaven die in een context zijn geplaatst en vragen zonder context én regels voor de onderliggende subonderdelen. Op basis van al deze regels, selecteert het algoritme de juiste vragen uit de grote vragenbank. Op die manier krijgt iedere leerling een inhoudelijk gebalanceerde toets.

Verschillende varianten van adaptieve toetsen

Een adaptieve toets betekent dat de toets zich aanpast aan het vaardigheidsniveau van de leerling. Je kunt hierbij grofweg twee soorten onderscheiden: adaptief op vraagniveau en adaptief op groepsniveau.

- De AMN Doorstroomtoets is adaptief op vraagniveau. De leerlingen beginnen met een aantal startvragen waarna de vragen op een volledig adaptieve manier worden aangeboden. Dit houdt in dat na ieder antwoord dat de leerling geeft, het vaardigheidsniveau opnieuw wordt berekend en op basis daarvan de volgende vraag wordt geselecteerd. Dit gebeurt dus na iedere vraag opnieuw.
- Een andere variant is adaptief op groepsniveau. Dit wordt ook wel een 'multistage test' (MST) genoemd (Mead, 2006). Dit betekent dat de leerling een reeks vragen maakt, waarna de vaardigheid wordt berekend. Vervolgens wordt een nieuwe reeks vragen aangeboden. Dit herhaalt zich een aantal keer tot het onderdeel stopt. Uit hoeveel vragen de reeksen bestaan en hoeveel van deze reeksen er zijn, verschilt per toets.

Beide soorten toetsen (volledig adaptief en MST) passen zich dus aan het niveau van de leerling aan, maar de mate en 'snelheid' waarin dit gebeurt verschilt. De voordelen van adaptieve toetsen zoals op de eerste pagina beschreven, zijn bij een volledige adaptieve toets groter dan bij een MST.

Statistische methode IRT

De AMN Doorstroomtoets maakt gebruik van Item Response Theory (IRT). Dit is een familie van statistische methodes. Je zou kunnen zeggen dat je met opgaven een bepaalde vaardigheid van een leerling meet, en met IRT krijg je inzicht in hoe goed elke individuele vraag die vaardigheid meet. AMN maakt gebruik van het 2-parameter logistisch model van Birnbaum (1968). Dit model geeft een indicatie van de kans dat een leerling een vraag goed beantwoordt op basis van diens vaardigheid en de kenmerken van een vraag. Die kenmerken heten ook wel itemparameters. In het 2-parameter logistisch model zijn, zoals de naam al doet vermoeden, twee parameters die van belang zijn: de discriminatieparameter en de moeilijkheidsparameter. Iedere toetsvraag heeft een waarde voor beide parameters.

- De moeilijkheidsparameter (β-parameter of bèta-parameter) geeft aan hoe moeilijk een vraag is. Hierbij is het belangrijk om in de vragenbank vragen te hebben met zowel hoge, gemiddelde en lage moeilijkheidsparameters - zodat het hele spectrum wordt gedekt. Vragen met een lage moeilijkheidsparameter (makkelijke vragen) worden aan de minder vaardige leerlingen aangeboden. Het levert immers weinig informatie op om deze leerlingen zeer moeilijke vragen aan te bieden. Voor de zeer vaardige leerlingen is het andersom: voor deze leerlingen leveren heel moeilijke vragen het meeste informatie op (dus met een hoge moeilijkheidsparameter).
- De discriminatieparameter (α-parameter of alpha-parameter) geeft aan hoe goed een vraag onderscheid kan maken tussen het vaardigheidsniveau van leerlingen. Hoe hoger de α van een vraag, hoe beter. De α is echter wel gekoppeld aan de moeilijkheidsparameter: het geeft aan hoe goed de vraag 'discrimineert' als de leerling een vaardigheid heeft die ongeveer gelijk is aan de moeilijkheid van de vraag. Of te wel: je moet de vraag aanbieden aan een leerling die ongeveer hetzelfde niveau heeft (als de vraag niet past bij het niveau van de leerling, heb je nog niets aan een hoge α). De α-parameter van een vraag geeft dus aan hoe goed deze onderscheid kan maken tussen de vaardigheidsniveaus van leerlingen. Als deze parameter laag is, dan geeft de vraag relatief weinig informatie. Dit betekent dat als een leerling de vraag goed heeft, dat niet direct hoeft te betekenen dat die leerling ook beter is in die specifieke vaardigheid dan een leerling die de vraag fout had. Er kunnen dan andere factoren een rol spelen.

Om een zeer precieze inschatting van het niveau van de leerling te kunnen maken, heb je vragen nodig die qua moeilijkheid dicht bij de (geschatte) vaardigheid van de leerling liggen én voldoende onderscheidend vermogen hebben.

Valide

In de Toetswijzer doorstroomtoetsen PO (College van Toetsen en Examens, 2022) worden eisen gesteld aan de inhoud van de toets. Hierin wordt onder andere beschreven dat de toets de onderdelen lezen, taalverzorging en rekenen moet meten uit het Referentiekader Taal en Rekenen (Meijerink et al., 2009). In dit referentiekader is beschreven welke vaardigheden leerlingen zouden moeten beheersen aan het eind van het primair onderwijs. Dit is beschreven aan de hand van referentieniveaus. Door elke vraag te koppelen aan dat referentiekader is de inhoudsvaliditeit van de doorstroomtoets gewaarborgd. Dat houdt in dat de toets daadwerkelijk meet wat hij moet meten.

Betrouwbaar

De mate van betrouwbaarheid van een toets geeft weer in hoeverre je dezelfde uitkomst zou krijgen als een leerling de toets opnieuw zou maken. De betrouwbaarheid wordt uitgedrukt met een waarde tussen 0 en 1 (of 0 en 100%), waarbij 1 (of 100%) staat voor maximale betrouwbaarheid. Vanaf een waarde van 0.7 spreekt men van een hoge mate van betrouwbaarheid (Taber, 2018). Elk jaar onderzoeken we de betrouwbaarheid van de AMN Eindtoets (nu doorstroomtoets) gebaseerd op de methode van Glas en Emons (2017) en Bechger et al. (2003).

In 2023 was de betrouwbaarheid van de gehele AMN Eindtoets 96%. De betrouwbaarheid van de onderdelen lezen, taalverzorging en rekenen waren respectievelijk 89%, 92% en 96%. Aangezien de AMN Doorstroomtoets inhoudelijk vrijwel gelijk is aan de AMN Eindtoets, zal deze naar verwachting ongeveer dezelfde resultaten laten zien.

Objectief

De opgaven in de AMN Doorstroomtoets worden automatisch gescoord. De antwoorden van de leerlingen worden allemaal gescoord op basis van dezelfde standaarden. Het is onmogelijk dat eenzelfde antwoord de ene keer goed wordt gerekend en de andere keer niet. De scoring is daarmee objectief.

Efficiënt

De AMN Doorstroomtoets meet alleen de voorgeschreven 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 zo efficiënt mogelijk 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 (2022). Toetswijzer doorstroomtoetsen PO. https://zoek.officielebekendmakingen.nl/stcrt-2022-26159.html
- Expertgroep Doorlopende Leerlijnen Taal en Rekenen (2009). Referentiekader taal en rekenen. De referentieniveaus. Enschede: SLO.
- Glas, C.A.W., & Emons, W.H.M. (2017). Blueprint voor psychometrische verantwoording normering toetsadviezen en ijking op de referentieniveaus. EPO: Utrecht
- Guttman, L. (1945). A basis for analyzing test-retest reliability. Psychometrika, 10 (4), 255-2 82.
- He, W., & Reckase, M. D. (2014). Item Pool Design for an Operational Variable-Length Computerized Adaptive Test. Educational and Psychological Measurement, 74(3), 473–494. https://doi.org/10.1177/0013164413509629
- Kainz, O., Cymbalák, D. and Jakab, F. (2015) 'Adaptive web-based system for examination with cheating prevention mechanism', Lecture Notes on Software Engineering, 3(2), pp. 90–94. https://doi.org/10.7763/LNSE.2015.V3.172
- Ling, G., Attali, Y., Finn, B., & Stone, E. A. (2017). Is a computerized adaptive test more motivating than a fixed-item test? *Applied Psychological Measurement*, *41*(7), 495–511. https://doi.org/10.1177/0146621617707556
- Martin, A. J., & Lazendic, G. (2018). Computer-Adaptive Testing: Implications for students' achievement, motivation, engagement, and subjective test experience. *Journal of Educational Psychology*, *110*(1), 27–45. https://doi.org/10.1037/edu0000205
- Mead, A. D. (2006). An Introduction to Multistage Testing. Applied Measurement in Education, 19(3), 185–187. https://doi.org/10.1207/s15324818ame1903_1
- Meijerink, H. P., Letschert, J. F., Streun, A. van, Bergh, H. H. van den, & Rijlaarsdam, G. C. W. (2009). Referentiekader taal en rekenen. www.slo.nl/publish/pages/5901/referentiekader_taal_en_rekenen_referentieniveaus.pdf
- Taber, K. S. (2018). The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education. Research in Science Education, 48(6), 1273–1296. https://doi.org/10.1007/s11165-016-9602-2
- Weiss, D. J. (2011). Better Data From Better Measurements Using Computerized Adaptive Testing. Journal of Methods and Measurement in the Social Sciences, 6(1), 1–27. https://doi.org/10.2458/azu_jmmss.v2i1.12351