

50 JAAR CITO, EEN HALVE EEUW WISKUNDE-EXAMENS?

Ivo Claus
Irene van Stiphout

DEEL 5

In september 2018 bestond Cito 50 jaar. In het vijfde artikel over de rol die Cito door de jaren heen heeft gespeeld bij de wiskunde-examens schrijven Ivo Claus en Irene van Stiphout over de normering. Zij interviewen wetenschappelijk onderzoekers Hendrik Straat en Marieke van Onna van de psychometrische afdeling van het Cito



Inleiding

Cito beschikt over een van de grootste psychometrische onderzoeksafdelingen van Europa. Deze specialisten voeren onder andere berekeningen uit voor de normering van de centrale examens. Op basis van hun technische advies en het advies van de inhoudelijk betrokken toetsdeskundigen stelt het College voor Toetsen en Examens de N-term vast.

Voor dit artikel zijn we in gesprek gegaan met twee wetenschappelijk onderzoekers van deze afdeling, Marieke van Onna en Hendrik Straat. Beiden zijn betrokken bij de normering van de centrale examens. Ieder jaar doen ongeveer 200.000 leerlingen centraal examen. Deze leerlingen verdienen een eerlijke beoordeling en een diploma met een vaste waarde. Maar eerlijk normeren is ingewikkeld. Tegen welke keuzes of dilemma's lopen Hendrik en Marieke aan?

Equivaleren

Als eerste punt noemen Hendrik en Marieke *equivaleren*. Daarbij draait het erom dat cijfers van verschillende examens van hetzelfde vak gelijkwaardig zijn. De meest nauwkeurige manier vanuit een psychometrisch oogpunt om dit te doen, is niet altijd de meest praktische. De aanpak van de normering hangt onder andere af van het vak, het aantal deelnemers en het type toets, aldus Hendrik en Marieke. Hendrik: 'Het uitgangspunt is dat leerlingen *gelijkwaardige* prestaties moeten aantonen. Dit betekent dat het voor een leerling in 2019 evenveel vaardigheid vraagt om een voldoende te halen als van een leerling die in datzelfde vak in 2018 examen heeft gedaan. Het *gelijk waarderen* van gelijke prestaties heet ook wel *equivaleren* en betekent dat de moeilijkheid van toetsen wordt vergeleken om zo te berekenen welke score op de ene toets overeenkomt met een bepaalde score op de andere toets.'



Hendrik Straat en Marieke van Onna

Marieke: 'Er zijn verschillende manieren om te *equivaleren*. De meest nauwkeurige manier om dit te doen is via *anchor in package*. Deze manier wordt in het voortgezet onderwijs alleen gebruikt bij digitale toetsen. Leerlingen maken nieuwe opgaven en zogeheten *anker-opgaven*, die eerder bij andere leerlingen in eerdere examens zijn afgenomen. Via de resultaten op deze ankeropgaven kan een inschatting worden gemaakt van de relatieve moeilijkheid van examens, en daarmee van het gemiddelde niveauverschil van leerlingen in twee verschillende examenjaren. Deze manier is nauwkeurig omdat er veel gegevens onder liggen. De ankeropgaven zijn door veel andere leerlingen in bijna dezelfde omstandigheden gemaakt. Dat maakt dat de moeilijkheid van opgaven en de vaardigheid van leerlingen nauwkeurig geschat kunnen worden.' Twee andere manieren van *equivaleren* zijn *pretest* en *posttest*. Hierbij worden enkele opgaven van een toekomstig examen (*pretest*) of net afgenomen examen (*posttest*) afgenomen in combinatie met ankeropgaven in een afzonderlijk onderzoek. Dit is minder nauwkeurig dan *anchor in package* omdat de afnamecondities vaak

anders zijn. Opgaven worden bijvoorbeeld afgenomen tijdens een schoolexamen. Dat is anders dan een centraal examen. Een ander verschil is dat de groepen leerlingen van wie gegevens zijn verzameld, kleiner zijn. Door deze verschillen is er een groter betrouwbaarheidsinterval voor de geschatte moeilijkheid van het examen. Een laatste manier is via standaardsetting. Hierbij maken groepen experts een inschatting van de moeilijkheid van opgaven van twee verschillende examens.'

Hendrik: 'Anchor in package is de meest precieze manier om de norm te handhaven. Ieder jaar berekenen we hoe nauwkeurig onze schattingen zijn. We zien dat de schattingen bij

anchor in package preciezer zijn dan die van pre- en posttest.' De kernvakken worden zo nauwkeurig mogelijk geëquivalet. Hendrik wijst erop dat de aard

van het vak bepaalt wat voor analyse er gedaan wordt. 'Voor sommige vakken zijn er heel weinig leerlingen waardoor we weinig data hebben. Denk bijvoorbeeld aan Russisch, Fries en Arabisch. De kernvakken Nederlands, Engels en wiskunde willen we zo nauwkeurig mogelijk equiveren, omdat deze vakken een groter gewicht hebben bij de beslissing over het wel of niet verkrijgen van het diploma. Deze examens zijn op havo/vwo op papier, dus anchor in package kan niet. Dus willen we voor deze vakken een pretest of posttest. Voor leerlingen hangt hier immers veel van af.'

Vanuit een psychometrisch perspectief zou het ideaal zijn als anchor in package bij alle vakken gebruikt zou kunnen worden. Dat zou de meest nauwkeurige normhandhaving zijn. Marieke en Hendrik beseffen dat dit praktisch niet haalbaar is. 'De inhoud bepaalt hoe de examens eruitzien, niet de psychometrie.'

Balans

Dit brengt ons op de balans tussen inhoud en psychometrie. Niet alle vragen die inhoudelijk interessant zijn, zijn psychometrisch gezien geschikt. En andersom geldt dat vragen die heel onderscheidend zijn, inhoudelijk wellicht niet interessant zijn. Hoe kijken Marieke en Hendrik hier tegenaan?

Hendrik: 'Voor alle examens is de syllabus leidend. Niet alles wat in de syllabi staat, kan in één toets worden getoetst, dan zouden examens veel te lang worden.

De keuze van de onderwerpen wordt gemaakt op basis van de inhoud. Als alleen gestreefd zou worden naar de hoogste betrouwbaarheid, zouden de examens er vreemd uitzien en zou er veel verloren gaan. Het gaat er juist om dat leerlingen breed worden opgeleid en veel verschillende onderwerpen aandacht krijgen. Zou de psychometrie voorop staan, dan zou je een eenzijdig examen krijgen. Het gaat om het vinden van een juiste

balans tussen inhoud en meetkwaliteit. De toetsdeskundigen en CvTE waken over deze balans.' Marieke: 'Een makkelijke instapvraag is puur psychometrisch gezien niet interessant. Zo'n vraag draagt immers niet bij aan het maken van het onderscheid tussen vaardige en minder vaardige leerlingen. Maar voor leerlingen is zo'n opgave juist wel belangrijk om in een context te kunnen komen en op zijn of haar gemak te raken.'

Psychometrie kan helpen om grip te krijgen op de kwaliteit van vragen, benadrukken Marieke en Hendrik. Bij vragen met veel punten, zoals vragen met veel rekenstappen of onderzoeksvragen, heb je voldoende obser-

vaties nodig van alle scorepunten om de moeilijkheid van het behalen van ieder extra scorepunt goed te kunnen schatten. In de praktijk blijkt dat leerlingen op

bijvoorbeeld een 8-puntsvraag veel scores in het midden hebben, en dat de scores 0, 1, 2 en 6, 7 en 8 nauwelijks voorkomen. Dat betekent dat het onderscheid tussen de moeilijkheid van 0, 1 of 2 aan de ene kant, of tussen 6, 7 of 8 scorepunten aan de andere kant, niet duidelijk is. En er ontstaat een schijnnaauwkeurigheid door deze scorepunten wel allemaal toe te staan.

Psychometrie kan volgens Hendrik ook helpen om zicht te krijgen op afhankelijkheid tussen opgaven. Dit doet zich voor als een leerling die een eerste vraag bij een context niet weet en dit tot gevolg heeft dat hij of zij de rest ook niet kan maken. Hendrik: 'Dit soort problemen is op verschillende manieren te ondervangen. Een optie is om een tussenantwoord te geven. Dit kan door in de vraag bijvoorbeeld op te nemen om $x = 2,45$ te nemen als het eerste onderdeel niet gelukt is.' Bij opgaven waarin veel gerekend moet worden en een rekenfout in het begin ertoe kan leiden dat een leerling de vraag niet helemaal kan beantwoorden, zou volgens Hendrik gevraagd kunnen worden naar een beschrijving van de aanpak. 'Overigens blijkt uit Cito-onderzoek bij digitale vmbo-examens dat correctie op tussenstappen geen invloed heeft op het cijfer dat leerlingen voor het examen behalen. Vanuit psychometrisch oogpunt is er daarom geen noodzaak om tussenstappen te scoren, al kunnen er natuurlijk vakinhoudelijke redenen zijn om dit wel te doen.'

Eerlijkheid en transparantie

Wiskundig gezien is er veel mogelijk om zo nauwkeurig mogelijk te equiveren. Maar dan is het bijna niet meer te volgen voor docenten hoe dit gaat. Hoe kun je zorgen voor een eerlijke normering die toch nog te volgen is? Marieke: 'Een eenvoudige methode om te equiveren bestaat niet. Elke methode, hoe simpel of hoe ingewikkeld ook, heeft voor- en nadelen. Het moet een transparant proces zijn.' Volgens Hendrik betekent dit dat helder moet

'DE INHOUD BEPAALT HOE DE EXAMENS
ERUITZIEN, NIET DE PSYCHOMETRIE.'

zijn wat de uitgangspunten zijn. Verschillen van mening zijn vaak terug te voeren naar verschillende uitgangspunten. Een voorbeeld hiervan is de aanname van gelijke populaties. Hendrik: 'De aanname is dat de populatie die op een vak examen doet als geheel en over de jaren heen niet zoveel verandert. Onderdeel van het normeringsproces is om te onderzoeken of in deze aanname gaten geschoten kunnen worden. Een voorbeeld van een reden dat een populatie kan veranderen is de invoering van de kernvakkenregeling. Zo'n verandering kan betekenen dat leerlingen eerder een minder moeilijk profiel kiezen om te voorkomen dat ze het niet halen. Het kan ook zijn dat leerlingen meer aandacht gaan besteden aan de kernvakken omdat deze zwaarder meetellen in de uitslagregeling en zodoende hoger op deze vakken scoren.'

Het wiskundige model dat wordt gebruikt voor de normering kan heel complex worden, maar dan is het voor een leek moeilijk te volgen. Marieke: 'Vanwege de transparantie willen we dat het model alleen maar ingewikkelder mag worden als het evident is dat die ingewikkeldheid iets oplevert. In dat geval is het uit te leggen. Het statistische model wordt zeker niet altijd beter door meer data te verzamelen of meer statistische toetsen en bellen te gebruiken. Vaak win je het een, en verlies je het ander.' Hendrik: 'Welk model ook wordt gebruikt, het belang van de leerling staat altijd voorop. Dat betekent met name dat een eerlijke waardering van de prestatie van de leerlingen belangrijk is. Als daarvoor een ingewikkelder model nodig is, dan zullen we altijd ons best doen om een transparante, duidelijke uitleg te geven.'

De rekenmodellen die worden gebruikt voor de normering worden voortdurend tegen het licht gehouden. Marieke: 'Aanpassingen zijn altijd gebaseerd op onderzoek. Stel dat we een ander psychometrisch model willen gaan hanteren, dan gaan we eerst kijken naar wat de gevolgen van een aanpassing zouden zijn voor leerlingen. Dan rekenen we bij oude data door wat dat voor leerlingen zou betekenen. Soms worden nieuwe paden verkend om bepaalde knelpunten in de normering op te lossen, ook als dit niet direct in de praktijk gebracht kan worden.'

Voor sommige vakken is het heel lastig om een 10 te halen. Hendrik: 'Bij deze vakken kun je ook nadenken over een andere methodiek. Nu is het alleen maar mogelijk om een 10 te halen als er nul fouten zijn. Een ander uitgangspunt zou kunnen zijn dat voor het behalen van een 10 een minimaal niveau nodig is. Dan zou een leerling met een paar foutjes toch een 10 kunnen halen. Dit vraagt wel om uitleg, want het kan al gauw een hoop gegoochel lijken.'

Tot slot

Tot slot hebben we Hendrik en Marieke gevraagd om tips voor docenten om hun eigen toets goed te normeren. Hendrik: 'Bij het bepalen van de cesuur, de grens tussen onvoldoende en voldoende, gaat het erom wat het meest aannemelijk is. Zijn klassen goed vergelijkbaar dan kun je koersen op een gelijk percentage voldoende. Heb je de toetsen zo gemaakt dat ze even moeilijk zouden moeten zijn dan kun je bij hetzelfde percentage correcte antwoorden de lat leggen. Een andere optie is een bepaald gemiddeld cijfer, bijvoorbeeld

gemiddelde 6,0, of het resultaat van een andere jaarlijkse toets, als benchmark nemen.'

Daarbij ga je ervan uit dat het resultaat op de andere toets de algemene vaardigheid van jouw leerlingen

weerspiegelt. Als je merkt dat een klas over het geheel beter wordt op de andere toets, dan kun je het gemiddelde cijfer op jouw toets ook wat hoger laten zijn.

Marieke: 'Daarnaast zou je kunnen kijken of het mogelijk is om meer gegevens te verzamelen. Gegevens van één toets die door één klas is gemaakt, zeggen niet veel. Als er veel hoge scores zijn behaald, concludeer je dan dat de groep vaardig is of dat de toets te makkelijk was? Dat valt niet met zekerheid te zeggen. Dit is wel mogelijk als je meer gegevens verzamelt zoals historische gegevens over de moeilijkheid van de items of historische gegevens van de leerlingprestaties op eerdere toetsen.'

Volgens Marieke en Hendrik is het vooral van belang om te sturen op wat er past bij je verwachtingen.

Als de leerlingen wat vaardiger zijn, en je komt met $N = 1$ op een 5,8 gemiddeld uit, dan kun je misschien $N = 1,5$ doen. Dan doe je de leerlingen meer recht.

Dit natte vingerwerk is bij de centrale examens natuurlijk uit den boze.

'HET BELANG VAN DE LEERLING STAAT ALTIJD
VOOROP. DAT BETEKENT DAT EEN EERLIJKE
WAARDERING VAN DE PRESTATIE VAN DE
LEERLINGEN BELANGRIJK IS.'

Over de auteurs

Irene van Stiphout en Ivo Claus zijn toetsdeskundigen bij Cito.

Emailadressen:

Irene.vanStiphout@cito.nl en Ivo.Claus@cito.nl.