

VIER DESKUNDIGEN NEMEN DE CESUUR ONDER DE LOEP

WAAR LIGT DE CESUUR? EEN MASTERCLASS

Een 5,5 is een voldoende. In het onderwijs hanteren we al sinds jaar en dag deze grens tussen onvoldoende en voldoende, in vaktermen: de cesuur. Is dat gewoon een kwestie van afspraak of zit er een logica achter dit getal? Het voelt vreemd om een rekentoets te krijgen waarbij een 4,5 al voldoende is. Komt dat omdat we gewoontedieren zijn of is er meer aan de hand? Hoe komen cijferbepalingen tot stand en wat zijn de overwegingen?

Theo Eggen is hoogleraar aan de Universiteit Twente, verbonden aan Cito en het Research Center voor Examinering en Certificering (RCEC). **Bernard Veldkamp** is hoogleraar aan de Universiteit Twente en tevens verbonden aan RCEC. **John de Jong** is hoogleraar Taaltoetsing aan de VU. **Tjeerd Haitjema** is psychometrisch expert bij Bureau ICE. Met deze experts heeft Toets! op verschillende momenten gesproken, waarbij ieder gesprek zijn eigen verloop kende. In dit artikel vatten we hun inzichten en antwoorden samen.

WAT IS EIGENLIJK EEN CESUUR?

Theo Eggen: 'Een cesuur is een standaard. Een cesuur is altijd gekoppeld aan een besluit. Boven de grens neem je een ander besluit dan onder de grens: voldoende of onvoldoende, geslaagd of niet geslaagd, toegelaten of afgewezen. De eindtoets voor het basisonderwijs, zoals Cito die maakt, heeft eigenlijk vier cesuren; vier blokjes waar vier verschillende besluiten bij horen. Er zijn overigens ook toetsen waarbij zo'n besluit niet aan de orde is. Bijvoorbeeld bij het meten van vooruitgang. Dan is er helemaal geen cesuur nodig.'

ZIT ER EEN LOGICA ACHTER DE 5,5 ALS VOLDOENDE?

'Zeker', zegt **De Jong**. 'Een 5,5 staat voor een score die hoort bij iemand die ten minste 50% van het getoetste niveau, en wat daarin beschreven staat, goed doet. Onder de 5,5 zitten scoorders bij wie de kans groter is dat ze het niveau niet beheersen. De score 5,5 (en hoger) wordt toegekend aan de groep mensen bij wie de kans groter is dat ze het wel kunnen dan dat ze het niet kunnen. En in deze groep zal je niemand meer tegenkomen waarbij de kans groter is dat hij

het niet kan dan dat hij het wel kan. Dat is de definitie van het Raschmodel. In feite doen we nogal ingewikkeld met deze psychometrische onderbouwing. Iedere docent weet dat de grens ongeveer op de helft ligt, een 5 is onvoldoende en een 6 is net voldoende. Niemand haalt het in zijn hoofd om te zeggen dat je een 1 voldoende is, net zomin als dat je minstens een 9 moet halen om te slagen. Dat bestaat niet.'

WE KRIJGEN EEN REKENTOETS, WAARBIJ EEN 4,5 VOLDOENDE IS. IS DAT EEN GOEDE ONTWIKKELING?

Daar is **De Jong** heel kort over: 'Ze zijn gek.' **Haitjema** vult aan: 'Het lijkt me een halfslachtige erkenning op de inhoudelijke kritiek uit het veld. Maar het is bizar om dan de cesuur te verlagen. Óf je weerlegt de kritiek, óf je besluit dat 2F nog te hoog gegrepen is en je verlegt je ambitie. Maar dat doe je niet door de cesuur te verlagen. Beter is om dan te besluiten, tijdelijk, op een niveau lager, 1F, te toetsen.'

Eggen ziet dat anders. Het vasthouden aan de cijferschaal is misschien onzin. Maar als je dat doet kan in het begin een 4,5 wel acceptabel zijn. Tijdelijk. Als je maar toewerkt naar een 5,5 als grens. **Veldkamp** onderschrijft de tijdelijkheid van die aanpassing: 'Bij de invoering van de rekentoets is geadviseerd om de cesuur bewust lager te leggen omdat de meeste examenkandidaten voor rekenen nog niet op de referentieniveaus van Meijerink zitten. Door de lat eerst laag te leggen en hem vervolgens in een paar jaar tijd stapsgewijs te verhogen tot de referentieniveaus, geef je het vo en het mbo de tijd om het rekenonderwijs op een verantwoorde manier in te voeren, zodat je van examenkandidaten ook kunt verwachten dat ze de rekentoets halen.'

WAT HEB JE NODIG OM TE KUNNEN VASTSTELLEN OF IEMAND IETS VOLDOENDE BEHEERST?

De Jong: 'Je hebt eerst een beschrijvingskader nodig van wat mensen moeten kunnen doen en hoe goed ze dat moeten kunnen doen. Zo'n beschrijvingskader beschrijft ook hoe iemand zich ontwikkelt van nagenoeg geen beheersing naar praktisch volledige beheersing. Vervolgens zoek je naar een punt waar de kans groter is dat iemand het wel kan dan dat hij het niet kan. Dat is het basisprincipe. Neem het examen van een bakkersopleiding. Als iemand goed deeg kan rollen is dat niet voldoende om te slagen. Maar hij hoeft ook niet alle finesses te beheersen, want dan hebben we geen bakkers meer in Nederland. Je hoeft geen 5 sterren te hebben om een restaurant te beginnen. Je vraagt dus nooit het maximum, maar ook nooit het minimum: je zit er altijd ergens tussenin. Wanneer weet je meer wel dan niet? Dat zal je met elkaar moeten bepalen. Internationale onderzoeken laten zien dat het heel belangrijk is voor een hoge onderwijsopbrengst dat alle betrokkenen heel duidelijk weten wat de bedoeling is van het onderwijs, wat de standaarden zijn en wat de inhoud van het onderwijs zou moeten zijn.'

Omdat veel leerlingen in de basisberoepsgerichte leerweg moeite hebben om rekenniveau 2F te behalen, komt er een eenvoudigere variant. Wel krijgen de vmbo-bb-leerlingen vanaf 2015-2016 vier kansen. Tegelijkertijd wordt er gestart met een pilot met een rekentoets 2A, de eenvoudigere rekentoets voor vmbo-bb. Maar er is geen inhoudelijke beschrijving van 2A.

IS DIT WENSELIJK?

Veldkamp: 'Ik ben heel blij met die eenvoudige variant. Als je naar de toetsresultaten van vmbo-bb-leerlingen kijkt en in gesprek gaat met het veld, dan moet je constateren dat 2F voor veel van hen onhaalbaar is en je kunt je afvragen of zij dat niveau echt nodig hebben in hun werk.' Maar **Haitjema** is geen voorstander: 'Niet wenselijk, zo wordt de inhoud van een nog onbekende toets leidend voor de invulling van een niveau. Er is nu nergens een inhoudelijke beschrijving van 2A, dus docenten, leerlingen en andere betrokkenen in het onderwijs weten niet waar ze zich op moeten voorbereiden. Het is bijna onethisch. Breng eerst maar in kaart waar leerlingen precies moeite mee hebben, bespreek wat wenselijk is wat ze wel moeten beheersen, pas indien noodzakelijk de didactiek daarop aan, maak een inhoudelijke beschrijving en maak die openbaar. Daarna pas komt toetsconstructie om de hoek kijken. Een goede toets moet gebaseerd zijn op een duidelijk inhoudelijk kader. Pas dan kun je met elkaar controleren of het examen goed is en een daarbij passende cesuur opstellen.'

Ook **De Jong** vindt dit een verkeerde ontwikkeling: 'Zelfs daar waar we syllabussen hebben, hebben we nog te weinig houvast. Laat staan dat we in een syllabus opschrijven wat er in het examen zit. Omgekeerde wereld. Omdat het



‘Een cesuur is een standaard. Een cesuur is altijd gekoppeld aan een besluit.’

niet bekend is wat we eigenlijk willen meten, zijn de voexamens nu al niet zomaar een goede vertaling van wat we willen dat mensen kunnen.'

SAMEN BEPAAL JE DUS DE INHOUD. DAN VAN INHOUD NAAR UITSLAG: HOE STEL JE EEN CESUUR OP?

Eggen: 'Heb je nog weinig data, dan werkt men het meest met de standaardbepalingsmethode Angoff. Je neemt meerdere experts, minstens 8 om een goede steekproef te hebben, en het gezamenlijk oordeel van de experts geeft je de cesuur. Je laat hen een student in gedachten nemen die op de grens van voldoende/onvoldoende zit. Vervolgens laat je hen per item aangeven of ze denken dat 'hun' grensstudent deze wel of niet goed zal maken. Je verzamelt alle oordelen en middelt ze, of sluit de uitersten uit. Belangrijk bij deze methode is dat de experts zeer vertrouwd zijn met de inhoud van wat er getoetst moet worden en met de toets zelf. Laat de experts ook eerst met elkaar discussiëren over de vragen voordat zij hun inschattingen met betrekking tot de 'grensstudent' geven. Bekend is dat veel beoordelaars de grensstudent overschatten. Zeker bij innovaties, nieuwe toetsen, denken beoordelaars dat leerlingen meer goed zouden moeten doen dan ze in de praktijk blijken te kunnen. Op basis van de expertoordelen kun je besluiten om de cesuur, de 5,5, op een bepaalde plek te leggen. Vervolgens is het wijs om ook even af te wegen wat de gevolgen zijn van de cesuur op deze plek: hoeveel studenten zullen er nu zakken en slagen? Hierover kun je in een nieuwe ronde met de experts opnieuw nadenken. Deze methode werkt goed en is redelijk uitvoerbaar. Je zorgt zo

tegelijkertijd voor draagvlak. Ongeveer zo is het ook gegaan bij de rekentoets. Je ziet dan dat de vakdeskundigen de cesuur hoog leggen en dat de resultaten in de praktijk anders uitwijzen. De gedachte was dat het rekenniveau omhoog zou gaan. Maar het niveau gaat niet omhoog omdat je een toets afneemt. Daar is meer voor nodig.'

SOMS STUREN WE AAN OP EEN BEPAALD SLAGINGSPERCENTAGE. HOE DOE JE DAT DAN MET EEN CESUUR?

Eggen: 'Bij een goed beschreven inhoudelijk kader werkt zo'n Angoff-procedure best redelijk. Maar je kunt ook een combinatie maken met een variant van de Hofstee-methode. Neem de kennisbasistoets bij de lerarenopleidingen van de hogescholen. Daar stellen we een cesuur vast, maar combineren dat met de mogelijkheid om te schuiven (Hofstee). Je vraagt je dan af: wat is het minimale en maximale aanvaardbare percentage personen dat kan slagen? En wat is het minimale en maximale percentage kennis dat iemand moet hebben en om te slagen? Bijvoorbeeld: je moet minimaal 40% van de voorgeschreven kennis hebben en je mag nooit meer dan 80% van de kandidaten laten zakken. Een soort compromismethode, feitelijk twee schuiven waar je mee mengt tot je tot een aanvaardbare cesuur komt.'

WAAROM ZIJN DE SLAGINGSPERCENTAGES VAN DE VERSCHILLENDE VAKKEN NIET HETZELFDE?

Eggen: 'Verschillende deskundigenpanels voor de verschillende vakken kunnen via de Angoff-procedure uitkomen op verschillende cesuren. Weet je wat dwaas is? We slagen er eindelijk in om meer leerlingen een exact

vakkenpakket te laten kiezen. Het is al jaren zo dat van de vwo'ers bijna de helft het vak wiskunde B kiest. Iedereen doet examen Nederlands. En de slagingspercentages voor wiskunde zijn bijna altijd een procent of 10 lager dan die voor Nederlands, ondanks dat wiskunde maar door een geselecteerde groep is gemaakt. Dit zal met de status van het vak wiskunde te maken hebben: het is moeilijk, en dus vinden wij lagere slagingspercentages aanvaardbaar. Zo zie je hoe vreemd we soms bezig kunnen zijn. Dan hebben we eindelijk meer leerlingen, maar dan laten we ze zakken en demotiveren hen op deze manier om verder exact te kiezen.'

DUS HOE GROOT DE INSPANNING OOK ZOU ZIJN VAN LEERLINGEN, LERAREN, LESMETHODES, UREN, ENZOVOORT, JE ZAL DAT NIET TERUGZIEN IN DE EXAMENRESULTATEN?

Eggen: 'Nee, en dat zou natuurlijk wel moeten. Ik denk wel dat deskundigen hier op de langere termijn een halt aan moeten toeroepen, maar vooralsnog is het zo.'

EERDER WERD GESPROKEN OVER EEN GOED INHOUDELIJK KADER. WELKE BASIS BIJEN DE REFERENTIE-NIVEAUS VOOR BIJVOORBEELD DE REKENTOETS?

Eggen: 'Dat vind ik heel lastig. Want men zegt wel dat je opgaven op een bepaald niveau hebt, maar wat betekent dat eigenlijk? Wat zegt je dat dan? Maakt iemand met dat niveau die opgave goed of heeft diegene een kans van 50% of meer om dat item goed te maken? Dat weet men eigenlijk niet zo goed. De items van een niveau zijn echt niet altijd even moeilijk. Er zitten opgaven bij waarvan je verwacht dat iemand die precies op dat niveau zit ze niet goed zal maken, en dat iemand die hoger dan die cesuur zit, ze wel goed zal maken. Kortom, die niveaus worden vertaald in opgaven die zelf variëren in moeilijkheidsgraad, maar die als het goed is wel beperkt zijn in spreiding. Dus het is niet zo maar klip en klaar bij welk niveau een item hoort, of hoe moeilijk een niveau is.'

EN HOE STUREND IS HET COMMON EUROPEAN FRAMEWORK VOOR HET MAKEN VAN TOETSEN VOOR DE MODERNE VREEMDE TALEN?

De Jong: 'Brian North heeft ooit de descriptor (beschrijvingen van taaltaken, red.) voor niveaus van taalvaardigheid over de hele wereld onderzocht en verzameld. In Europa is dat opgepakt in de vorm van een Europees Referentiekader. Dat kader is gebaseerd op 500 losse descriptor. Die zijn als losse statements geschaald, dat wil zeggen: ieder statement ('kan de krant lezen', 'kan uitleggen waarom hij te laat is gekomen') is geschaald van 'kan helemaal niks' tot 'kan praktisch alles'. Die schaal is vervolgens ingedeeld in intervallen omdat het als geheel onbeheersbaar was. Die zijn gedefinieerd als A1, A2 etc. En wat is de grens tussen A1 en A2? Dan kom je weer op het verhaal van de cesuur. Wanneer beheers je A2? Toen ze de schaal maakten hebben ze de descriptor ingedeeld. De descriptor in het interval A1 hoorden bij A1 en de descriptor daarboven, in het interval A2, hoorden bij A2. Vervolgens heeft men die descriptor samengevat en dat zijn de globale niveaubeschrijvingen geworden. Nu komt de vraag: waar moet je zitten als gebruiker van de taal om

'Als een toets inhoudelijk rammelt, kan ik er wel allemaal ingewikkelde analyses op uitvoeren, maar dan blijft het een gedrocht.'



BERNARD VELDKAMP

'Gestandaardiseerde toetsen passen goed bij een klas met standaardleerlingen, maar daar zijn er maar weinig van.'



