Onderzoek

De cesuur bij de landelijke kennistoetsen van lerarenopleidingen

Monika Vaheoja, 10voordeleraar en Universiteit Twente, Theo van den Bogaart, Hogeschool Utrecht

De cesuur van een toets geeft aan vanaf welke score een student een voldoende heeft. De uitdaging is om een methode voor het vaststellen van de cesuur te hanteren die recht doet aan de vaardigheid van de student en onafhankelijk is van de specifieke toetsversie. Bij de landelijke kennistoetsen van de Nederlandse lerarenopleidingen zijn voor cesuurstelling twee methodes gebruikt: Cohen, gebaseerd op 95ste-percentielscores, en Angoff, gebaseerd op expertpanels. Aan de hand van een toetsafname bij natuurkunde en wiskunde onder 202 studenten is geanalyseerd in hoeverre de cesuren uit de Angoff- en de Cohenmethode (met verschillende parameters) op verschillende toetsversies vergelijkbaar zijn, wanneer ze worden omgezet in een latente vaardigheid door middel van itemresponstheorie. Ook zijn de interne en externe consistentie van het Angoffpanel onderzocht. Hoewel alle cesuren verschillen, kan geconcludeerd worden dat de Angoffmethode het eerlijkst is en ook, maar alleen bij wiskunde, de Cohenmethode met specifieke parameters. De inschattingen per toetsversie van het Angoffpanel zijn consistent, maar de samenhang tussen de cesuur en de moeilijkheid van de toets is laag. Het onderzoek van consistentie van Angoffpanels aan de hand van ankeropgaven laat een genuanceerd beeld zien: natuurkundepanels geven een stabiele cesuur, maar met een grote spreiding; die spreiding is bij wiskunde kleiner, maar daar is het panel uit het onderzoek milder dan eerdere panels. De aanbeveling is om een derde methode te gaan gebruiken, gebaseerd op itemresponstheorie, gecombineerd met een methode die lerarenopleiders zicht geeft op de cesuurstelling ten behoeve van de ecologische validiteit.

Inleiding

De bekostigde Nederlandse hbo-lerarenopleidingen hebben gezamenlijke kennisbases (10voordeleraar, 2020a) ontwikkeld, waarin is vastgelegd welke vakinhoudelijke kennis een student moet beheersen om een onderwijsbevoegdheid te krijgen. Aan deze kennisbases is een landelijke kennistoets (LKT) verbonden (10voordeleraar, 2020b). Deze manier van kennisborging heeft internationale aandacht gekregen en is door de OECD, de Organisatie voor Economische Samenwerking en Ontwikkeling, aangewezen als *'promising practice'* (OECD, 2018).

De landelijke kennistoetsen worden meerdere keren per studiejaar afgenomen. Iedere toetsversie wordt samengesteld op basis van een toetsmatrijs en bevat deels nieuwe toetsvragen en deels (minimaal 35%) toetsvragen die ook eerder zijn afgenomen, zogenoemde ankeritems. De toetsvragen zijn drie- of vierkeuzevragen of vragen waarop een numeriek antwoord kan worden gegeven. De cesuur geeft aan wanneer een kandidaat voor de toets is geslaagd. Omdat de verzameling opgaven bij iedere afname verschilt, roept dit de vraag op hoe je de cesuur bij verschillende toetsversies zo stelt dat die steeds dezelfde bekwaamheid van de student vereist. De ervaring bij de LKT is dat verschillen in cesuur in individuele gevallen grote impact kunnen hebben op het studieverloop van een student. De vraag naar een eerlijke cesuur kan echter ook voor andere toetsen bij andere opleidingen gesteld worden en is daarmee dus niet alleen relevant voor de LKT. De kwestie is momenteel misschien wel relevanter dan ooit, omdat opleidingen zich door de coronamaatregelen genoodzaakt zien om meerdere toetsversies af te nemen waar ze voorheen zouden volstaan met één centrale afname.

De mate waarin een student de leerdoelen die we willen toetsen beheerst, noemen we de vaardigheid (Engels: *proficiency*). Dit is een latente variabele: hij is niet rechtstreeks zichtbaar. Een toets beoogt de

vaardigheid van een student te meten. De mate waarin dit lukt, bepaalt per definitie de validiteit van de toets. In het geval van een kennistoets wordt geprobeerd de vaardigheid te meten door de student een aantal vragen voor te leggen. De antwoorden op de vragen worden van punten voorzien en de som daarvan is de toetsscore. In het geval van de LKT, waar geen deelscores worden toegekend, is de toetsscore niets anders dan het totaalaantal correct beantwoorde vragen. De cesuur is de minimale toetsscore waarbij een student voldoende vaardig wordt bevonden. Omdat de verzameling vragen, en de moeilijkheid daarvan, per toetsversie verschilt, kan ook de cesuur per toetsversie verschillen. Er bestaan verschillende methodes om de cesuur te bepalen en deze resulteren voor dezelfde toets vaak in verschillende toetsscores (De Gruijter, 2008). Er zijn zowel theoretische als praktische argumenten voor de keuze van een methode (Cizek & Bunch, 2007), maar belangrijk is dat de cesuur consistent is. Dat betekent dat de slaagkans alleen afhangt van de vaardigheid van de deelnemer en niet van de moeilijkheid van de toetsversie (Van der Linden, 1995).
Bij de invoering van de LKT werd voor de Angoffmethode gekozen. Deze methode leek het best in staat om de betrokkenheid van de lerarenopleiders bij de LKT te vergroten en hun de gelegenheid te bieden om zelf een landelijke norm te bepalen. Later heeft de Cohenmethode de Angoffmethode vervangen, aangezien de twee methodes een vergelijkbare cesuur leken te leveren, maar de Cohenmethode efficiënter en goedkoper is. In dit artikel wordt de consistentie van de twee cesuurmethodes vergeleken. Uniek van dit onderzoek is het gebruik van itemresponstheorie om toetsmoeilijkheid en vaardigheid te onderscheiden. Itemresponstheorie biedt namelijk een model om een toetsscore om te zetten in een vaardigheidsscore.

Cohenmethode

Bij de Cohenmethode (Cohen-Schotanus & Van der Vleuten, 2010) wordt een relatief referentiepunt gebruikt om de cesuur te bepalen. Dat referentiepunt is de toetsscore op het 95ste percentiel van de studenten bij wie de toets is afgenomen. De gedachte is dat de vaardigheid van dermate hoog scorende studenten stabiel is over verschillende populaties en dus een graadmeter is voor de moeilijkheid van de toets. De cesuur is in essentie 60% van deze percentielscore. Dit percentage is gekozen omdat het volgens Cohen-Schotanus en Van der Vleuten (2010) in de praktijk van cesuurstelling de meest gangbare is. Om precies te zijn wordt er een compensatie voor de gokscore g op meerkeuzevragen uitgevoerd: $c = 0{,}6(m - g) + g$, met c de cesuur en m de 95ste-percentielscore. De getallen 95 en 60 zijn de parameters van de Cohenmethode en kunnen anders gekozen worden. Er is geen onderzoek bekend waarbij de Cohenmethode wordt onderzocht op consistentie. Er is bijvoorbeeld weinig bekend over de gevoeligheid van de 95ste-percentielscore van de steekproefgrootte en de afhankelijkheid van de deelnemersvaardigheid.

Angoffmethode

Een kenmerk van de Angoffmethode is dat inhoudsdeskundigen de moeilijkheid van de toets inschatten op basis van de vastgestelde leerdoelen (Angoff, 1971). Aan de hand van een strak protocol wordt iedere toetsvraag afzonderlijk beoordeeld op moeilijkheid voor de zogenoemde grensstudent. Een grensstudent is een fictieve student die het onderwijs over de kennisbasis doorlopen heeft, enigszins voorbereid voor de toets is en de toets voor het eerst gaat maken. Deze grensstudent heeft nét de minimale kennis waardoor hij voor de toets zal behoren te slagen. De inhoudsdeskundigen schatten per toetsvraag welke fractie grensstudenten deze goed heeft. Daarbij is het mogelijk dat de inhoudsdeskundigen ook een inhoudelijke discussie voeren om zo meer overeenstemming te bereiken over de moeilijkheid. De som van de fracties over alle toetsvragen geeft de cesuur die de inhoudsdeskundige aan de toets zou geven. De uiteindelijke cesuur voor de toets is het gemiddelde van de cesuur van de verschillende deskundigen.
Een nadeel van de Angoffmethode is dat verschillende panels andere vereisten voor moeilijkheid stellen, waardoor de cesuur per panel kan verschillen (Brandon, 2004). Daarnaast blijkt het voor de panelleden lastig om de moeilijkheid van de toetsvragen in te schatten, waarbij de grootste inschattingsfouten bij de 'extreme' items worden gemaakt. Brandon stelt dat de inschatting over hoe moeilijk een toetsvraag voor een grensstudent is, niet bij alle panelleden op een consistente manier gedaan wordt. Het toevoegen van verschillende rondes waarin bijvoorbeeld empirische gegevens uit de toetsafname worden gedeeld, lijkt slechts beperkt

invloed te hebben op de overeenstemming tussen de panelleden en op de uiteindelijke cesuur. Zelfs kalibratievragen in een introductieronde leiden niet noodzakelijk tot meer consistentie. Op basis van deze kennis kan verwacht worden dat de Angoffmethode geen consistente moeilijkheid als cesuur voor verschillende toetsversies zal opleveren, als de panels bij verschillende toetsversies verschillen in samenstelling. Er is echter geen onderzoek bekend waarbij hetzelfde panel op dezelfde dag twee toetsversies normeert, wat mogelijk wel een consistente cesuur zou kunnen opleveren. In dit onderzoek analyseren we de consistentie van de twee cesuurmethodes die in de context van de LKT zijn gebruikt. De volgende onderzoeksvragen zullen worden beantwoord.

1. In hoeverre zijn de cesuren die bepaald zijn door Cohenmethodes met diverse parameters (90ste- en 95ste-percentielscore, en vermenigvuldigingsfactor 60% en 65%) en cesuren van de Angoffpanels over verschillende toetsversies vergelijkbaar, wanneer ze zijn omgezet in vaardigheid door middel van itemresponstheorie?
2. In hoeverre verschillen de cesuren van verschillende Angoffpanels op dezelfde toets?
3. In hoeverre zijn de schattingen van moeilijkheid van toetsvragen van Angoffpanels intern consistent?

De parameters die in de eerste deelvraag zijn gekozen, zijn in de literatuur (Cohen-Schotanus & Van der Vleuten, 2010; Taylor, 2011) de meest courante. Omwille van de complexiteit is een derde potentiële parameter, het al dan niet gebruik van de gokscore, buiten beschouwing gelaten. Het hanteren van de gokscore lijkt sterk verankerd te zijn in de praktijk.

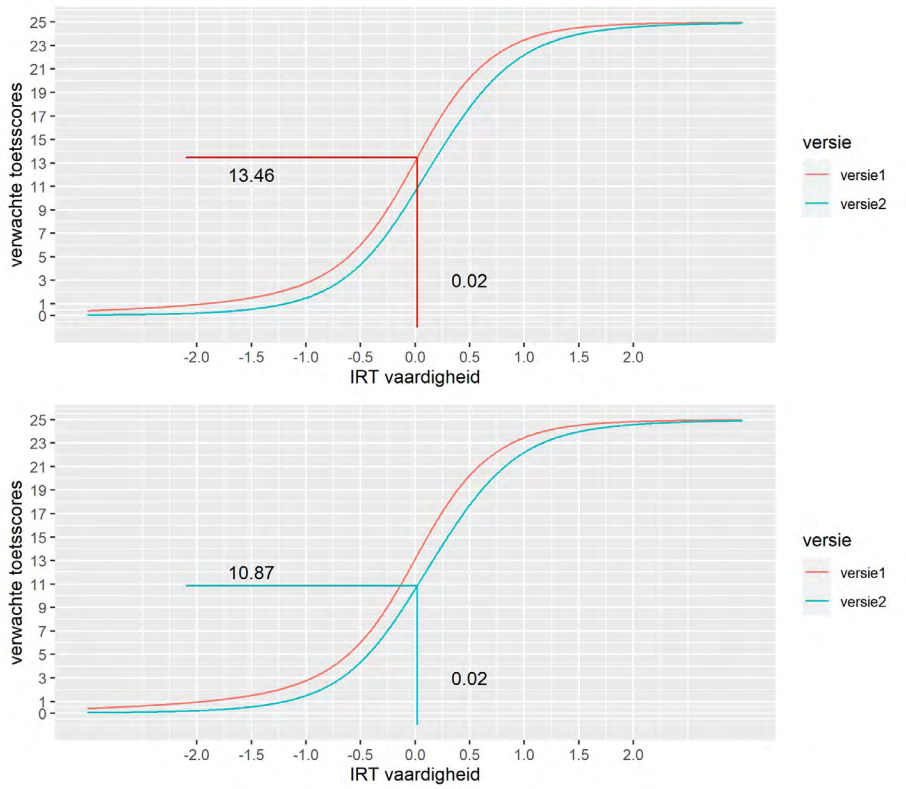
Onderzoeksmethode

Centraal in het experiment staan de LKT's van de lerarenopleidingen natuurkunde en wiskunde die in december 2019 zijn afgenomen. Naar de validiteit van de LKT wiskunde is eerder onderzoek gedaan (Drijvers et al., 2016). Er is voor natuurkunde en wiskunde gekozen vanwege het animo speciaal voor dit experiment een Angoffpanel te formeren. Omdat de samenstelling van het panel een groot effect op de cesuur kan hebben, zijn beide LKT's in tweeën gedeeld: zo kan hetzelfde panel in één dag naar twee toetsversies kijken. Er is voor ieder vak dus een panel en ieder panel kijkt naar twee toetsversies. In de samenstelling van deze toetsversies is geprobeerd de praktijk te benaderen: de versies zijn door een inhoudsdeskundige gemaakt in lijn met de toetsmatrijs, maar er is niet getracht de toetsen dezelfde moeilijkheid te geven.

Op alle vier de toetsversies (twee natuurkunde, twee wiskunde) wordt ook de cesuur bepaald met Cohenmethodes met verschillende parameters. Om deelvraag 2 over cesuren van verschillende Angoffpanels te beantwoorden, wordt daarnaast gekeken naar zogenoemde ankeropgaven: opgaven die ook al in eerdere LKT's zijn gebruikt en waarvan andere Angoffpanels al schattingen van de moeilijkheid hebben gemaakt.

Itemresponstheorie

Omdat in de beschreven aanpak verschillende toetsversies in het spel zijn, is de verwachting dat de cesuur per toetsversie verschilt: als de ene toetsversie bijvoorbeeld moeilijker is dan de andere, zal de cesuur bij de eerste ook lager moeten liggen. Om de cesuur van verschillende toetsversies te vergelijken, is het noodzakelijk de cesuur te vertalen in een genormaliseerde grootheid die onafhankelijk is van de toets. Deze genormaliseerde grootheid is de (latente) vaardigheid. De functie die een toetsscore omzet in een vaardigheid, wordt geleverd door de itemresponstheorie (IRT) (Veldkamp, 2019). We zullen kort uitleggen wat dit betekent.
In Figuur 1 staan twee voorbeelden van toetskrommen uit de itemresponstheorie bij twee fictieve toetsen. De grafiek geeft aan hoe de vaardigheid correspondeert met de toetsscore. Stel dat de cesuur van toetsversie 1 wordt vastgesteld op een toetsscore van 13.46. In de bovenste afbeelding is te zien dat hier een vaardigheid van 0.02 bij hoort. Toetsversie 2 is een moeilijkere toets en het is dan ook te verwachten dat de cesuur, als toetsscore, hier lager is. In termen van vaardigheid moet de cesuur echter stabiel zijn. Daarom zou de cesuur uit de onderstaande afbeelding moeten volgen: bij vaardigheid 0.02 hoort een toetsscore van 10.87.



Figuur 1. Voorbeeld hoe de cesuur van versie 1 naar versie 2 wordt overgebracht door middel van toetskrommen die geschat zijn met IRT.

De toetskromme ontstaat door eerst per toetsitem een soortgelijke itemkromme te construeren en vervolgens deze krommen bij elkaar op te tellen. Iedere itemkromme heeft een logistische vorm, die wordt vastgelegd door twee parameters: positie en steilheid. De waarde van deze parameters wordt bepaald door het specifieke toetsitem waarvoor de kromme wordt getekend. De positie wordt bepaald door de moeilijkheid: hoe moeilijker het item, hoe meer de kromme naar rechts komt te liggen. De steilheid wordt bepaald door de discriminatie-index: dit is een maat voor hoe goed het item onderscheid kan maken tussen lagere en hogere vaardigheid. Hoe steiler de grafiek, hoe groter het discriminerend vermogen van het item. De precieze methode om itemkrommen te schatten maakt gebruik van het één-parameter logistisch model (OPLM) (Verhelst et al., 1993).

Normeringspanels

Bij dit onderzoek zijn er twee normeringpanels: voor natuurkunde respectievelijk wiskunde. Ieder panel bestaat uit negen lerarenopleiders. Deze panelomvang is vergelijkbaar met de omvang die panels in het verleden hadden toen de Angoffmethode nog voor de cesuurstelling van de LKT werd gebruikt en is het resultaat van een compromis tussen organiseerbaarheid en betrouwbaarheid. De panelleden zijn aangedragen door de coördinatoren van de opleidingen. De eis was dat panelleden kennis moeten hebben van de kennisbasis, de toetsmatrijs en de toetsgids en ervaring moeten hebben met de studentenpopulatie. Deelname werd niet vergoed, maar het bood opleiders wel een kans om kennis te nemen van de LKT en feedback door te geven aan de redactie van de kennistoets.

De toetsversies

In december 2019 zijn de LKT natuurkunde en de LKT wiskunde afgenomen. Beide toetsen zijn op de gebruikelijke wijze geconstrueerd vanuit de opgavenbank op basis van de toetsmatrijzen. Een gedeelte van de toetsvragen bestaat uit ankeropgaven. In de analyses die volgen worden alleen de eerstekansers meegenomen: de studenten die de LKT voor het eerst maken. De herkansers worden in dit onderzoek genegeerd,

omdat het een atypische groep is: het zijn immers allen studenten die de toets minstens één keer, en mogelijk vaker, niet hebben gehaald.
De LKT wiskunde bestond uit 50 vragen en de LKT natuurkunde uit 60. Voor validatie van de panels zijn echter twee toetsen per opleiding nodig en daarom zijn de toetsen voor de normeringspanels in tweeën gedeeld door inhoudelijke experts die geen zitting in de panels hadden. Het streven was om de toetsen zodanig in tweeën te splitsen dat de inhoud van de twee delen zo gelijk mogelijk is en dat het verschil in moeilijkheid tussen de twee delen niet groter is dan tussen reguliere LKT's. Dit resulteerde in twee versies voor wiskunde van 25 vragen en twee versies van 30 vragen bij natuurkunde. Helaas bleek een vraag bij de natuurkundetoets in beide delen te zitten. Panelleden hebben deze vraag daarom twee keer genormeerd.

De normeringspanelprocedure
Voor wiskunde trof het panel elkaar op 24 januari 2020; voor natuurkunde op 7 februari. Beide dagen werd toetsversie 1 voor de lunch genormeerd en toetsversie 2 na de lunch. Op een korte kennismakingsronde na, is op beide dagen twee keer precies hetzelfde protocol doorlopen. Beide dagen werden afgerond met een voorlopige presentatie van de IRT-analyse en de ingeschatte cesuur bij de twee toetsversies. Daarnaast is er een discussie geweest over ervaringen met Angoffpanels.

Het protocol begint met een korte kalibratieronde, waarna panelleden per vraag individueel een schatting maken van de fractie studenten die de vraag goed heeft. Na alle vragen uit de toetsversie te hebben geschat, volgt een discussie waarin individuele schattingen kunnen worden bijgesteld. In die discussie worden ook de studentprestaties per vraag van de afname in december teruggekoppeld. De uiteindelijke cesuur wordt vastgesteld door te middelen over de panelleden (waarbij twee buitenste scores worden weggelaten) en vervolgens te sommeren over de toetsvragen.

Interne en externe validatie van de panelschattingen
Validatie van de panelschattingen wordt gedaan door itemresponstheorie (zie het voorbeeld in Vaheoja, 2019), waarbij het uitgangspunt is dat de vaardigheid bij de verschillende cesuurscores gelijk moet zijn.
Om de consistentie (betrouwbaarheid) van de panels te beoordelen, wordt generaliseerbaarheidstheorie gebruikt (Brennan & Lockwood, 1980). Bij de Angoffmethode maken panelleden per opgave een inschatting van de fractie studenten die de vraag goed heeft. De generaliseerbaarheidstheorie beschouwt zo'n inschatting als een stochastische variabele die de som is van vier componenten, te weten een parameter die de gemiddelde inschatting is voor de populatie panelleden en 'het universum aan opgaven', en drie stochasten: het effect veroorzaakt door de keuze van het panellid, het effect veroorzaakt door de keuze van de opgave en het residuele effect (ruis). Op basis van de data die het experiment genereert, kan een schatting worden gemaakt van de variantie van deze drie stochasten. Dit zal gebruikt worden om inzicht te geven in de consistentie van de inschattingen van de panelleden.
De derde validatiemethode is het vergelijken van de panelschattingen op de ankeropgaven met de panelschattingen uit het verleden. Deze vergelijking laat zien of de interpretatie van de grensstudent – die een centrale rol speelt in het Angoffprotocol – is veranderd. Als de schattingen nu lager zijn, geeft dit aan dat de panelleden milder zijn geworden in hun eisen voor de grensstudent; en omgekeerd. Omdat het aannemelijk is dat verschillende panels het globaal eens zijn over wat makkelijkere en moeilijkere vragen zijn, verwachten we enige correlatie tussen de schatting van het huidige panel en schattingen uit het verleden.

Eerlijkheid van de cesuurmethodes
Per vak en per cesuurmethode zullen de cesuren van de twee toetsversies worden vergeleken aan de hand van de voorwaardelijke kans op slagen bij gelijke vaardigheid, gegeven de cesuur. Een cesuurmethode is eerlijk als deze kans onafhankelijk is van de versie van de toets waarop de cesuur is gebaseerd. Hier is een methode per definitie eerlijk als dezelfde vaardigheid bij verschillende toetsen gelijke slagingspercentages geeft (Van der Linden, 1995).

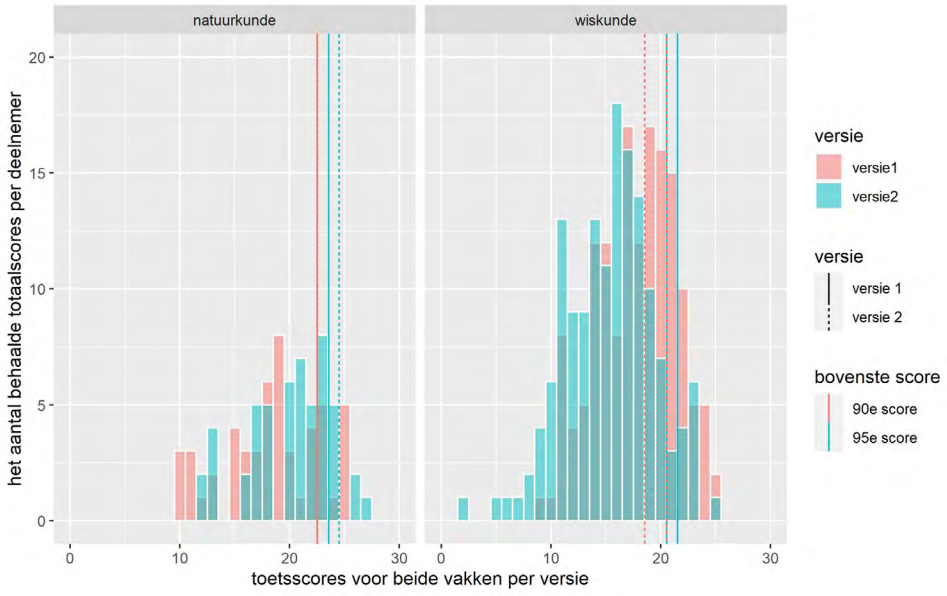
Resultaten

Geobserveerde toetsscores van de deelnemers op de toetsversies

In Tabel 1 zijn de gegevens van de toetsscores van de eerstekansers weergegeven, uitgesplitst naar de twee toetsversies. Per opleiding is op basis van IRT-schattingen de makkelijke en de moeilijke versie bepaald; te zien is dat de gemiddelde toetsscore en de 95ste-percentielscore ook lager is bij de moeilijkere versie. Merk echter op dat de 90ste-percentielscore bij natuurkunde voor beide toetsversies vrijwel gelijk is. In Figuur 2 zijn deze scores gevisualiseerd, waarbij de 90ste- en 95ste-percentielscores zijn weergegeven met lijnen.

	Wiskunde		Natuurkunde	
	versie 1 (makkelijker)	versie 2 (moeilijker)	versie 1 (moeilijker)	versie 2 (makkelijker)
Aantal eerstekansers	150	150	52	52
Aantal vragen	25	25	30	30
Cronbachs α	0.69	0.73	0.70	0.59
Gokscore	4.67	4.25	8	8
Toetsscores				
Minimum	9	2	10	12
1st kwartiel	15.00	12.25	15.75	17.75
Mediaan	18	16	19	21
Gemiddelde	17.76	15.35	18.33	20.06
90ste percentiel	20.55	18.54	22.52	22.56
95ste percentiel	21.53	20.52	23.55	24.51
Maximum	25	25	25	27

Tabel 1. Gegevens van de toetsscores van de eerstekansers.



Figuur 2. Histogrammen van totaalscores van eerstekansers, waarbij de 90ste- en 95ste-percentielscore voor beide versies zijn weergegeven.

De inschattingen en cesuren van de normeringspanels

In Tabel 2 zijn de normeringspanelschattingen van individuele panelleden weergegeven. Ook zijn de gemiddeldes en de standaarddeviatie per opleiding en toetsversie aangegeven. Vervolgens is de onafgeronde

cesuur weergegeven; deze is bepaald door in de berekening van het gemiddelde twee buitenste scores weg te laten. Voor de uiteindelijke cesuur wordt naar boven afgerond op gehelen (de toetsscore) en dit is ook in de tabel aangegeven. Merk op dat na afronding er alleen bij wiskunde versie 1 verschil is tussen het gemiddelde en de cesuur. Merk ook op dat de standaarddeviatie van de cesuur voor alle versies minder dan 1 scorepunt is, hetgeen aangeeft dat panelleden het redelijk met elkaar eens waren. In Tabel 2 is ook te zien dat er bij het natuurkundepanel drie panelleden (namelijk N5, N8 en N9) waren die in tegenovergestelde richting meebewogen met de moeilijkheid van de toets. Bij het wiskundepanel is een tegenovergestelde beweging niet te zien.

	Wiskunde (25 vragen per versie)			**Natuurkunde (30 vragen per versie)**	
Deelnemer	**versie 1 (makkelijker)**	**versie 2 (moeilijker)**	Deelnemer	**versie 1 (moeilijker)**	**versie 2 (makkelijker)**
W1	13.65	13.25	N1	15.50	17.45
W2	11.60	10.70	N2	16.12	16.35
W3	13.80	13.20	N3	15.97	16.75
W4	13.95	12.90	N4	14.76	16.85
W5	15.05	12.65	N5	15.41	15.25
W6	15.00	13.70	N6	14.08	14.37
W7	13.57	11.95	N7	15.48	18.15
W8	15.15	13.00	N8	18.20	16.75
W9	13.30	12.90	N9	15.10	14.40
Gemiddelde ± SD	13.90 ± 0.35	12.69 ± 0.28		15.62 ± 0.36	16.26 ± 0.41
Cesuur ± SD	14.05 ± 0.21	12.84 ± 0.09		15.48 ± 0.12	16.26 ± 0.35
Afgeronde cesuur	15	13		16	17
Beoordelaars-overeenstemming	0.91	0.90		0.92	0.85
$\hat{\sigma}^2(\bar{X}\|I)$	0.55	0.35		0.48	0.64
95%-betrouwbaarheids-Interval van het gemiddelde	[13.17; 14.63]	[12.11; 13.27]		[14.87; 16.37]	[15.40; 17.12]

Tabel 2. De cesuren per normeringspanellid en per opleiding en toetsversie, en de interne consistentie, de verwachte variantie en het 95%-betrouwbaarheidsinterval van de cesuur.

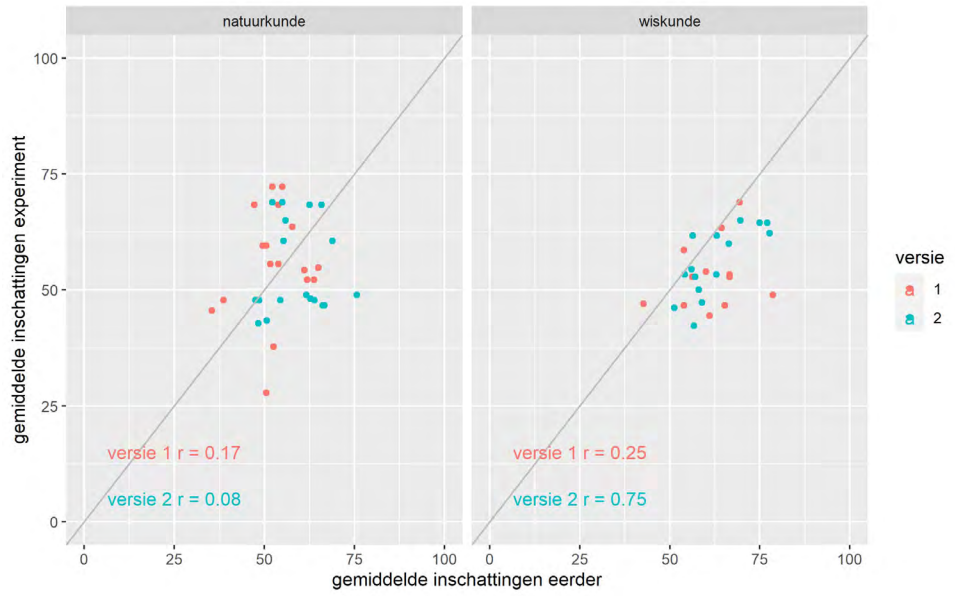
Interne consistentie van de normeringspanels

Tabel 2 geeft de verwachte variantie $\hat{\sigma}^2(\bar{X}|I)$ van de normeringspanelschatting $\bar{X}$ voor de vragen uit een toetsversie, wanneer wordt gegeneraliseerd over verschillende panelsamenstellingen en een vaste verzameling opgaven (zie voor de berekeningswijze Brennan & Lockwood, 1980). Combineren we dit met de gemiddelde panelschatting en de schatting van de variantie, dan geeft dit per toetsversie een 95%-betrouwbaarheidsinterval die ook in Tabel 2 is opgenomen. De gemiddelde variantie tussen de panelleden is klein ($\hat{\sigma}^2$=0.55 en $\hat{\sigma}^2$=0.35 voor versie 1 en 2 bij wiskunde, en bij natuurkunde $\hat{\sigma}^2$=0.48 en $\hat{\sigma}^2$=0.64). Deze kleine variantie geeft aan dat ze een grote mate van overeenstemming hebben (minder dan 1 scorepunt verschil) over de cesuur per toetsversie binnen het panel.

Externe consistentie van de normeringspanels

In Figuur 3 zijn de schattingen van de panelleden weergegeven op de ankeropgaven. Dit zijn de vragen die in het verleden ook door andere normeringspanels zijn ingeschat. Bij wiskunde betrof het 27 en bij natuurkunde 35 toetsvragen. De gemiddelde schatting bij wiskunde was eerder 62.19 en op dezelfde vragen werd door het panel uit het experiment gemiddeld 54.67 geschat. Dit verschil (D = –7.52, SE = 1.53) is significant lager ($t(26)$ = 4.91, p < 0.001). Bij natuurkunde was de gemiddelde schatting eerder 56.08 en deze keer 54.98. Dat is 0.4 scorepunten lager op 35 vragen (D = 1.10, SE = 2.21) en niet significant ($t(34)$ = 0.50, $p > 0.5$). De berekende correlaties tussen de inschattingen uit het verleden en de inschattingen van nu, laten echter bij wiskunde een

hogere samenhang in inschattingen zien: 0.25 voor versie 1 en 0.75 voor versie 2. Bij natuurkunde is er geen samenhang in de inschattingen op de ankeropgaven: 0.17 en 0.08 voor versie 1 en versie 2, respectievelijk.

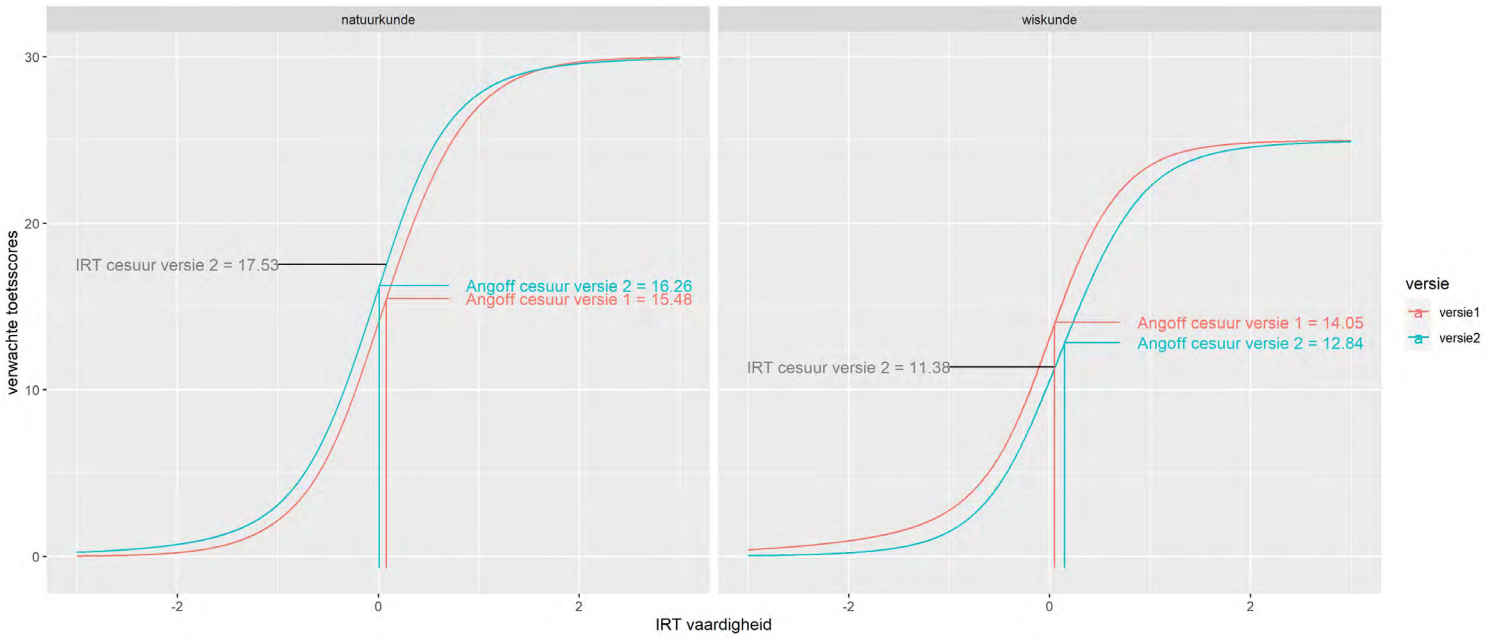


Figuur 3. Samenhang tussen gemiddelde schattingen van de moeilijkheid voor een grensstudent per toetsvraag op de ankeropgaven vanuit het experiment en de gemiddelde schattingen uit eerdere normeringspanels per toetsversie en vak.

Consistentie van de cesuur per cesuurmethode over de toetsversies

Voor beide opleidingen is er apart een model geschat (OPLM). Een statistische toets laat zien dat dit voor beide opleidingen een passend model is (natuurkunde R^*_{1c}=49.83, df=59, p=0.80; wiskunde R^*_{1c}=112.13, df=147, p=0.98; zie Glas, 1989).

In Figuur 4 zijn voor iedere toetsversie de krommen uit de itemresponstheorie (IRT) getekend die toetsscores en vaardigheid met elkaar in verband brengen. De onafgeronde cesuren van het panel zijn voor beide versies aangegeven. Ook is in de figuur aangegeven wat de cesuur bij de tweede versie had moeten zijn als de panelcesuur van de eerste versie consistent zou zijn. Het verschil tussen de aldus overgebrachte cesuur en de geschatte cesuur van het panel bij versie 2 is een maat voor de stabiliteit van het panel. Te zien is dat bij beide opleidingen de ingeschatte cesuur van het panel niet gelijk is aan de scorewaarde die via de toetskrommen wordt overgebracht. Beide panels wijken 1 scorepunt (onafgerond: wiskunde 1.5, natuurkunde 1.28) af.



Figuur 4. Toetskrommen berekend met itemresponstheorie (OPLM) op basis van alle deelgenomen studenten, waarbij de Angoffcesuren van versie 1 (rode lijn) is overgebracht naar versie 2 (zwarte lijn). De blauwe lijnen geven de cesuur aan die was geschat door de panelleden bij versie 2.

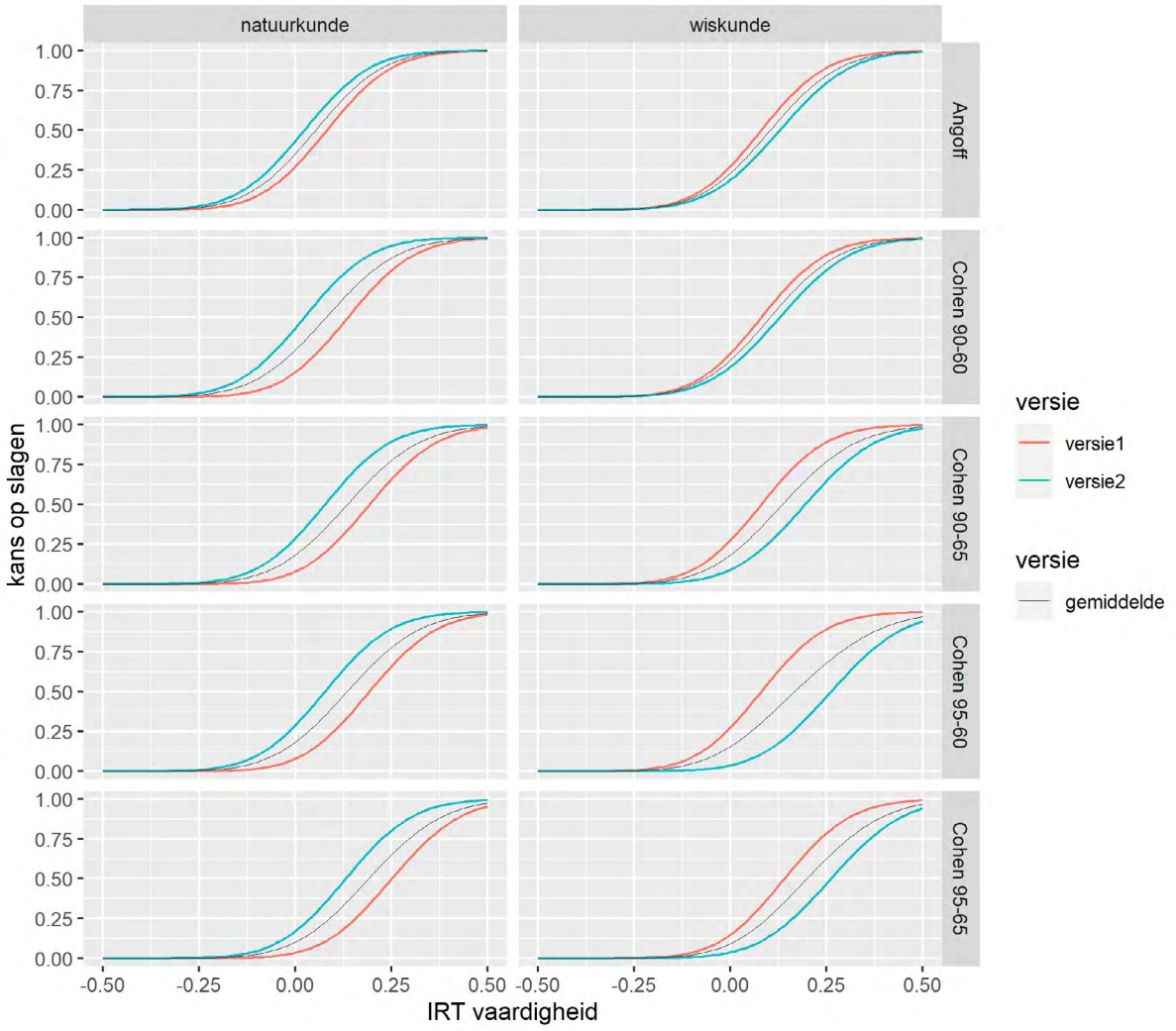
De Angoffcesuren zijn weergegeven in Tabel 3. Daarnaast bevat deze tabel vier cesuren die zijn berekend met de Cohenmethode met verschillende parameters. Ook hier is ter vergelijking weer de cesuur van de eerste versie via de IRT-kromme overgezet naar de tweede versie en vergeleken met de cesuur die het Angoffpanel bij de tweede versie stelde. Uit het verschil tussen die twee waarden blijkt dat geen van de methodes een stabiele cesuur als vaardigheid op de tweede versie bepaalt. Binnen de toetsversies is wel te zien dat zowel bij wiskunde als bij natuurkunde de cesuren die bepaald zijn met de Cohenmethode met parameter 90–60 (90ste-percentielscore en vermenigvuldigingsfactor 60%) en de Angoffmethode het meest overeenstemmen. Voorts lijken de Cohenmethodes met parameter 95–60 en 90–65 het meest op elkaar.

	Cohen 95–65	*Cohen 95–60*	*Cohen 90–65*	*Cohen 90–60*	*Angoff*
Wiskunde					
Onafgeronde cesuur	15.63	14.78	14.99	14.19	14.05
Bijbehorende IRT-vaardigheid	0.15	0.09	0.11	0.06	0.05
Cesuur voor versie 1	**16**	**15**	**15**	**15**	**15**
Onafgeronde cesuur	14.83	14.01	13.54	12.82	12.24
Bijbehorende IRT-vaardigheid	0.29	0.23	0.20	0.15	0.15
Cesuur voor versie 2	**15**	**15**	**14**	**13**	**13**
Onafgeronde cesuur van versie 1, overgebracht naar versie 2 met IRT	12.83	11.92	12.23	11.47	11.38
Eerlijke IRT-cesuur voor versie 2	**13**	**12**	**13**	**12**	**12**
Verschil tussen IRT-cesuur en de cesuur gezet voor versie 2	**-2**	**-2**	**-3**	**-1**	**-1**
Natuurkunde					
Onafgeronde cesuur	18.11	17.33	17.44	16.71	15.48
Bijbehorende IRT-vaardigheid	0.22	0.18	0.19	0.14	0.08
Cesuur voor versie 1	**19**	**18**	**18**	**17**	**16**
Onafgeronde cesuur	18.73	17.90	17.46	16.73	16.26
Bijbehorende IRT-vaardigheid	0.15	0.10	0.08	0.04	0.01
Cesuur voor versie 2	**19**	**18**	**18**	**17**	**17**
Onafgeronde cesuur van versie 1, overgebracht naar versie 2 met IRT	20.03	19.34	19.51	18.63	17.53
Eerlijke IRT-cesuur voor versie 2	**21**	**20**	**20**	**19**	**18**
Verschil tussen IRT-cesuur en de cesuur gezet voor versie 2	2	2	2	2	1

Tabel 3. Cesuren bepaald door de Cohen- en Angoffmethode voor alle versies, waarbij ook de cesuur van versie 1 door middel van itemresponstheorie is overgebracht naar tweede versie.

Eerlijkheid van de cesuur per cesuurmethode voor de deelnemers

Om de eerlijkheid van de cesuurmethodes te kunnen beoordelen, is de voorwaardelijke kans op slagen, gegeven de gestelde cesuur, als functie van de vaardigheden berekend en weergegeven in Figuur 5. De rode krommen geven de voorwaardelijke kansen aan voor versie 1 en de blauwe krommen voor versie 2. De zwarte kromme is het gemiddelde. Het punt waarbij de kans op slagen gelijk is aan 0.5 hoort bij de cesuurwaarde; immers, per definitie verwacht je dat de scores van de grensstudent symmetrisch verdeeld zijn rondom de cesuurwaarde.



Figuur 5. Voorwaardelijke kans op slagen, gegeven de cesuur op verschillende toetsversies, voor alle methodes om de eerlijkheid (gelijke slagingspercentage voor dezelfde vaardigheid) te kunnen beoordelen.

Bij een cesuur die eerlijk gesteld is, zou de kans op slagen alleen afhankelijk moeten zijn van de vaardigheid van de student en niet van de toetsversie. In de figuur is een cesuurmethode dus eerlijk als de rode en blauwe krommen samenvallen. Dit is bij geen enkele methode het geval. Bij wiskunde lijken de Cohenmethode met parameters 90–60 en de Angoffmethode het eerlijkst. Bij natuurkunde is de Angoffmethode het eerlijkst.

Impact van de verschillende cesuren op slagingspercentage

In Tabel 4 zijn het aantal geslaagde studenten met de gegeven cesuur uit Tabel 3 weergegeven om de impact van verschillende cesuren op de slagingspercentages te beoordelen. De oneerlijkheid van de verschillende cesuurmethodes komt tot uiting in het verschil van geslaagde studenten – aangezien in dit geval de twee toetsversies bij dezelfde groep studenten zijn afgenomen, zou er bij een consistente cesuur geen verschil in slagingspercentages tussen de toetsversies moeten zijn. Bij wiskunde resulteert dit, afhankelijk van de gekozen methode, in een verschil van 9 à 13 van de 150 eerstekansers die onterecht zakken wanneer de cesuur uit versie 1 als maatgevend wordt beschouwd. Bij natuurkunde gebeurt het omgekeerde: daar slagen onterecht 5 à 10 studenten van de 52.

		Cohen 95–65	*Cohen 95–60*	*Cohen 90–65*	*Cohen 90–60*	*Angoff*
versie 1	aantal geslaagd	107	119	119	119	119
(*n*=150)	% geslaagd	71%	79%	79%	79%	79%
versie 2	aantal geslaagd	90	90	103	112	112
(*n*=52)	% geslaagd	60%	60%	69%	75%	75%
IRT-cesuur	aantal geslaagd	112	121	112	121	121
(versie 1 naar versie 2)	% geslaagd	75%	81%	75%	81%	81%
verschil versie 2	aantal geslaagd	-22	-31	-9	-9	-9
en IRT-cesuur	% geslaagd	-15%	-21%	-6%	-6%	-6%
versie 1	aantal geslaagd	27	33	33	36	39
	% geslaagd	52%	63%	63%	69%	75%
versie 2	aantal geslaagd	34	39	39	44	44
	% geslaagd	65%	75%	75%	85%	85%
IRT-cesuur	aantal geslaagd	28	34	34	34	39
(versie 1 naar versie 2)	% geslaagd	54%	65%	65%	65%	75%
verschil versie 2	aantal geslaagd	6	5	5	10	5
en IRT-cesuur	% geslaagd	12%	10%	10%	19%	10%

Tabel 4. Het aantal geslaagde studenten bij de gegeven cesuur uit Tabel 3.

Eindgesprek

In het gesprek dat bij de afronding van de dag plaatsvond, gaven zowel bij wiskunde als bij natuurkunde panelleden aan dat ze het doorlopen van het Angoffprotocol een leuke en leerzame activiteit vonden. Ze doelden daarbij niet zozeer op het precieze protocol voor de cesuurbepaling, maar op de gelegenheid die het bood om de inhoud van de toets te leren kennen. Panelleden gaven aan dat dit hun een gevoel van transparantie gaf. Daarnaast vonden ze het waardevol om met collega's te kunnen discussiëren over de toetsvragen.

Conclusie en discussie

In hoeverre zijn de cesuren die bepaald zijn door Cohenmethodes met diverse parameters en de cesuren van de Angoffpanels over verschillende toetsversies vergelijkbaar, wanneer ze zijn omgezet in vaardigheid door middel van itemresponstheorie?

De cesuurmethodes leverden allemaal verschillende cesuurscores op. Ook was te zien dat per methode de cesuur die via IRT van toetsversie 1 werd overgebracht op toetsversie 2 verschilde van de cesuur die rechtstreeks op toetsversie 2 was vastgesteld. Dit heeft een aantal redenen, die alle te maken hebben met schattingsfouten. Ten eerste zijn de toetsen nooit volledig valide en betrouwbaar te krijgen (systematische fout); zie ook Cronbachs alfa in Tabel 1. Maar zelfs als dit wel zo zou zijn, zijn er steekproeffluctuaties in de wijze waarop studenten de toetsvragen beantwoorden (toevalsfout). Daarnaast doet IRT, zoals ieder model, niet volledig recht aan de complexe werkelijkheid (modelfout). Ten slotte vindt in de vaststelling van de cesuur een afronding plaats, die verschillen kan uitvergroten of juist verbergen (systematische fout); dit is onontkoombaar, omdat de toetsscore een discrete grootheid is.

Maar hoewel de cesuren verschillen, blijkt de Angoffmethode het eerlijkst te zijn; en, alleen bij wiskunde, ook de Cohenmethode 90–60. De kans voor een student om te slagen hangt daar het minst af van de toetsversie. Dit impliceert overigens niet dat de cesuur daarmee ook 'terecht' is. Deze cesuurbepalingen hebben veel impact. Bij wiskunde gaat het om het al dan niet slagen van 9 à 31 studenten; bij natuurkunde om 5 à 21 studenten. Er is op dit moment echter geen argumentatie gevonden om één methode te verkiezen boven een andere. Wat wel kan worden geconstateerd, is dat de aanname over de moeilijkheid van de LKT die in de Cohenmethode wordt gebruikt, in lijn is met de inschattingen van het Angoffpanel.

In hoeverre verschillen de cesuren van verschillende Angoffpanels op dezelfde toets?
Om de panelinschattingen nader te beoordelen is generaliseerbaarheidstheorie gebruikt. Dit liet zien dat elk panel consistent was in hun inschattingen per toetsversie. Dit komt tot uiting in de hoge beoordelaarsbetrouwbaarheid en de kleine betrouwbaarheidsgrens van de gemiddelde inschatting van het panel. Echter, hoewel de inschattingen per toetsversie consistent zijn, is de samenhang tussen de cesuur en de moeilijkheid van de toets laag. Dat impliceert dat verschillen in moeilijkheid van een toets de cesuur en de bijbehorende slagingspercentages kan beïnvloeden. Bij wiskunde is er wel samenhang te zien bij de inschattingen op de ankeropgaven, maar bij natuurkunde was er geen samenhang. Interessant is ook dat bij natuurkunde drie panelleden de tweede versie moeilijker inschatten, terwijl deze volgens de IRT-analyse juist makkelijker was. Dat de panelleden niet consistent meebewogen met de moeilijkheid van de toetsversies, was te zien bij zowel wiskunde als natuurkunde. Bij wiskunde was versie 2 moeilijker; daar was het panel strenger waardoor de cesuur minder laag kwam te liggen dan de cesuur die uit een IRT-vergelijking volgt. Bij natuurkunde was versie 2 makkelijker, maar het panel werd ook milder. Beide panels lijken dus met hun cesuur meer naar het midden te neigen. Of dit een systematisch effect is, valt op grond van dit onderzoek niet te zeggen; daarvoor zou je nog meer versies moeten onderzoeken.

In hoeverre zijn de schattingen van Angoffpanels intern consistent?
Bij natuurkunde wijken de inschattingen op de ankeropgaven niet significant af van inschattingen van eerdere panels. Bij wiskunde is dit wel het geval: panels uit het verleden hebben de cesuur hoger gelegd dan nu. Daarentegen was de spreiding bij de ankeropgave bij het panel wiskunde kleiner dan bij natuurkunde. Het oordeel van het wiskundepanel is dus stabiel, maar het nulpunt is verlaagd: wiskunde is minder streng geworden. Omdat de panelinstructie nagenoeg onveranderd is, kan dit niet het effect verklaren. Een oorzaak zou eerder gevonden kunnen worden in het samengaan van twee ontwikkelingen. Ten eerste is er onder lerarenopleiders de afgelopen jaren onvrede geuit over te hoge cesuurstelling. Het zou kunnen dat als gevolg hiervan in de Angoffpanels opgaven snel als lastig werden ingeschat. Men heeft dan een absolute cesuur, bijvoorbeeld 60%, in het hoofd en baseert de inschatting van de moeilijkheid van een vraag op de moeilijkheid van de toets als geheel, hoewel dit niet in overeenstemming met de voorschriften van de Angoffmethode is. Ten tweede is de toets, mede als gevolg van de onvrede, de afgelopen tijd mogelijk makkelijker geworden, doordat de hoeveelheid onderwerpen die wordt getoetst is gereduceerd en in de toetsconstructie preciezer wordt gekeken naar de complexiteit van toetsvragen. Dat betekent dat de ankeropgaven in relatie tot de complete set toetsopgaven nu relatief moeilijker kunnen zijn dan in oude LKT's. Omdat het panel, een hele toets overziend, opgaven mogelijk snel te moeilijk inschat kan dit verklaren waarom de ankeropgaven nu als moeilijker worden ingeschat. Bij natuurkunde heeft de eerste ontwikkeling ook plaatsgevonden, maar de tweede niet of in ieder geval minder systematisch – er zijn daar bijvoorbeeld nog steeds vragen waaraan studenten gemiddeld meer dan vier minuten besteden om te antwoorden – en dit kan verklaren waarom het effect bij natuurkunde niet is opgetreden. Dit is echter allemaal speculatie; en het is zelfs niet gezegd dat er sprake is van een reëel, systematisch verschil. Om daar meer duidelijkheid over te krijgen, zouden meer toetsversies moeten worden vergeleken.

Afrondende conclusie
Er zijn veel methodes voor cesuurstelling. In dit onderzoek zijn de methodes van Angoff en Cohen onderzocht, waarbij bij die laatste diverse parameters zijn gehanteerd. Met name bij natuurkunde lijkt de Angoffmethode eerlijker dan de Cohenmethodes. We constateerden echter ook dat de panelinschattingen op de ankeritems bij natuurkunde weinig samenhang hadden met de eerdere inschattingen. De methodekeuze heeft invloed op de uiteindelijke cesuur. Een verschil van enkele scorepunten kan grote impact hebben op individuele kandidaten; dit moet bij de keuze voor een cesuurmethode in overweging worden genomen. Een voorzichtige conclusie van dit onderzoek is wel dat, hoewel geen enkele cesuur uitblinkt in stabiliteit, de consistentie van alle methodes redelijk lijkt te zijn.

De conclusie betreft de LKT wiskunde en natuurkunde. Belangrijke contextfactoren die van invloed zijn op deze conclusie zijn de duidelijkheid van de kennisbasis en toetsmatrijs, de validiteit van de toets en de kwaliteit van het panel van inhoudsdeskundigen.
Kenmerk van de Cohen- en Angoffmethode is dat er naar een afzonderlijke toetsversie wordt gekeken. Er is een andere optie om de cesuur te bepalen, waarin meerdere toetsversies worden betrokken: via itemresponstheorie. Hierbij wordt van één toets, bijvoorbeeld met de Angoffmethode, de moeilijkheid vastgesteld, waarna de cesuur via IRT wordt overgebracht op alle volgende toetsen. Deze methode blijkt het meest stabiel, ook bij kleine steekproeven (Vaheoja, 2019). Dit alternatief vraagt wel om een strikt opgavenbankbeheer, met name met betrekking tot ankeropgaven. Op basis van de antwoorden van studenten op ankeropgaven kan verschil worden gemaakt tussen de moeilijkheid van de toetsvraag en de vaardigheid van de studenten. Dit moet echter nauwkeurig worden bijgehouden: zowel de antwoorden op alle vragen als de gegevens over in welke toetsen verschillende vragen zijn afgenomen. Een administratieve fout daarin kan grote invloed hebben op de parameterschattingen.
Met itemresponstheorie kan de cesuur als consistente vaardigheid over de toetsversies gehanteerd worden, wat vervolgens bepalend is voor de interne kwaliteit van een cesuurmethode. Cizek en Bunch (2007) onderscheiden daarnaast nog twee soorten kwaliteitscriteria: procedurele en externe. In de probleemstelling is al de organiseerbaarheid genoemd; op dit procedurele criterium scoort de Cohenmethode duidelijk beter dan de Angoffmethode. Voor de externe kwaliteit lijkt de Angoffmethode daarentegen weer beter. Dat wordt ondersteund door de reacties die de panelleden in de afsluitende discussieronde van het experiment gaven. Het kunnen analyseren van een toets draagt in de beleving van opleiders bij aan de transparantie en de redelijkheid van de zak-slaagbeslissing. Betrokkenheid van opleiders bij de cesuurbepaling past bij hun rol als poortwachter: bewaker van de toegang tot het beroep van leraar (Lunenberg et al., 2013). Daarnaast komt betrokkenheid de verbinding tussen de LKT en de rest van de opleiding, en daarmee de ecologische validiteit van de toets, ten goede. Tot slot kan het panel ook beschouwd worden als instrument voor professionalisering en instituutsoverstijgend intercollegiaal overleg. In de IRT-methodiek kan betrokkenheid van opleiders in de cesuurbepaling worden bewerkstelligd door periodiek, bijvoorbeeld eens in de drie jaar, de moeilijkheid van een toets te herijken met behulp van een panel. Op deze manier lijkt IRT een gulden middenweg te bieden, waarbij de transparantie van de toets enerzijds, en de eerlijke cesuurstelling voor de studenten anderzijds, elkaar ontmoeten.

Met dank aan: de normeringspanelleden, Noor van Gils, Karel Langendonck, de deelraad LKT van de Raad van Kwaliteitsborging, Dato de Gruijter en Norman Verhelst.

Monika Vaheoja
Monika Vaheoja werkt als psychometricus bij 1ovoordeleraar. Daarnaast is zij een promotieonderzoek aan het afronden bij Universiteit Twente op het toepassen van itemresponstheorie op de landelijke kennistoetspraktijk en cesuurstelling.

vaheoja@1ovoordeleraar.nl

Theo van den Bogaart
Theo van den Bogaart is lerarenopleider wiskunde. Hij is aangesloten bij het lectoraat Wiskundig en analytisch vermogen van professionals van Hogeschool Utrecht.

theo.vandenbogaart@hu.nl

Referenties

1ovoordeleraar (2020a). *Kennisbases*. Geraadpleegd op 12 mei 2020, van https://kennisbases.1ovoordeleraar.nl

1ovoordeleraar (2020b). *Praktische informatie landelijke kennistoetsen lerarenopleidingen*. Geraadpleegd op 12 mei 2020, van https://lkt.1ovoordeleraar.nl

Angoff, W. (1971). Scales, norms, and equivalent scores. In R. Thorndike (Red.), *Educational Measurement* (pp. 508-600). American Council on Education.

Brandon, P. K (2004). Conclusions about frequently studied modified Angoff standard-setting topics, *Applied Measurement in Education, 17*(1), 59-88.

Brennan, R. L, & Lockwood, R. E. (1980). A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory, *Applied Psychological Measurement, 4*(2), 219-240.

Cizek, G. J., & Bunch, M. B. (2007). Standard setting. *A guide to establishing and evaluating performance standards on tests*. Sage.

Cohen-Schotanus, J., & Van der Vleuten, C. (2010). A standard setting method with the best performing students as point of reference: Practical and affordable. *Medical Teacher, 32*(2) 154-160.

De Gruijter, D. N. M. (2008). *Toetsing en toetsanalyse*. Universiteit Leiden.

Drijvers, P. H. M., Straat, H., & Wools, S. (2016). Wiskunde valide getoetst? De digitale landelijke kennistoets wiskunde van de tweedegraads lerarenopleiding vergeleken met de instituutstentamens. *Tijdschrift voor lerarenopleiders, 37*(3), 27-38. www.velon.nl

Glas, C. A. W. (1989). *Contributions to estimating and testing Rasch models*. Cito.

Lunenberg, M., Dengerink, J., & Korthagen, F. (2013). *Het beroep van lerarenopleider. Professionele rollen, professioneel handelen en professionele ontwikkeling van lerarenopleiders*. Reviewstudie in opdracht van NWO/PROO. Vrije Universiteit Amsterdam.

OECD. (2018, mei). *Knowledge bases for initial teacher education in the Netherlands*. Geraadpleegd op 1 oktober 2020, van http://www.oecdteacherready.org/promising-practice/knowledge-bases-for-initial-teacher-education-in-the-netherlands

Taylor, C. (2011). Development of a modified Cohen method of standard setting. *Medical Teacher,* 678-682.

Vaheoja M. (2019). Finding equivalent standards in small samples. In B. Veldkamp & C. Sluijter (Red.), *Theoretical and practical advances in computer-based educational measurement. Methodology of Educational Measurement and Assessment* (pp. 175-185). Springer.

Van der Linden, W. (1995). A conceptual analysis of standard setting in large-scale assessment. *Proceedings of the joint conference on standard setting for large scale assessments of the National Assessment Governing Board (NAGB) and the National Center for Educational Statistics (NCES)* (pp. 97-117). U.S. Government Printing Office.

Veldkamp, B. P. (2019). Het wiskundige fundament van toetsen en examens. *Nieuw Archief voor Wiskunde, 5*(20), 161-168.

Verhelst, N. D., Glas, C. A. W., & Verstralen, H. H. F. M. (1993). *OPLM: One parameter logistic model. Computer program and manual.* Cito.