OVO
Artikel

De cesuurmethode Cohen-Schotanus bij de opleiding Geneeskunde in Leiden

Gijs van Duijn, Arnout Jan de Beaufort en Roeland van der Rijst

De beslissing omtrent de cesuur, de grens onvoldoende/voldoende, is een belangrijke en moeilijke beslissing. In de literatuur worden verschillende methoden voorgesteld voor het bepalen van de cesuur. In deze studie wordt de invloed onderzocht van het elimineren van vragen en het verkleinen van de referentiegroep op de uitkomsten van toetsen bij gebruik van de cesuurmethode Cohen-Schotanus. De keuze van de methode van cesuurbepaling heeft voor studenten ingrijpende gevolgen. Het is dus verstandig om goed na te denken over de methode die gekozen wordt in de opleiding.

Inleiding

Grofweg worden er drie type cesuurmethoden onderscheiden: absolute-, relatieve- en compromismethoden. Bij absolute bepaling van de cesuur wordt voorafgaand aan de toets bepaald wat de slaaggrens is. Het laagste aantal punten waarbij een student nog een voldoende heeft, is de cesuur. Wijnen (1971) presenteerde in zijn proefschrift een relatieve normering, ook wel de methode Wijnen genoemd. Daarin wordt de scheidslijn voldoende/onvoldoende achteraf bepaald aan de hand van de gemiddelde score van de studenten die meedoen aan de toets (Wijnen, 1971). De cesuurmethode van Cohen-Schotanus is een compromismethode waarbij een absolute cesuur wordt toegepast op de hoogst *bereikte* score (in tegenstelling tot de theoretisch hoogst mogelijke score) van de toets. Bij deze methode wordt rekening gehouden met het kennispercentage van de beste presteerders op een toets (Cohen-Schotanus, van der Vleuten & Bender, 1996; van Duijn, 2010). De methode Cohen-Schotanus gaat uit van een in principe absolute normering. Als correctie op deze absolute norm wordt een relatief referentiepunt in de normbepaling meegenomen. Een belangrijke reden waarom de cesuurmethode een relatief deel heeft, is dat toetsen een wisselende moeilijkheidsgraad hebben en dat voorafgaand onderwijs een wisselende mate van volmaaktheid kent (de Gruijter, 2008; Hambleton & Pitoniak, 2006). De redenering achter de cesuurmethode Cohen-Schotanus is dan dat de best presterende studentengroep laat zien waar de beste prestatie ligt die onder de gegeven omstandigheden, qua onderwijs en toets, kon worden geleverd. In de praktijk is dit de hoogst haalbare score gebleken. Hoewel de tentamens met veel zorg worden gemaakt, is het nooit uitgesloten dat er vragen bij zitten waarvan de docent bij de analyse achteraf vindt dat ze niet zo fraai zijn. Is de docent niet tevreden over die gemiddelde score van de studenten, dan moet, volgens Wijnen, de docent dat vooral aan zichzelf wijten: dan deugde of zijn onderwijs niet, omdat de studenten te weinig zijn gestimuleerd, of de toets deugde niet omdat de vragen niet goed bij de verworven kennis aansloten (Herraets, 1999; Wijnen, 1986). Het kan ook zijn dat bepaalde onderwerpen tijdens het onderwijs, bij nader inzien, onderbelicht zijn gebleven terwijl daar wel tentamenvragen over gaan. Door de (5%) beste studenten als referentiepunt te nemen wordt gecorrigeerd voor dit soort onvolkomenheden in de toets en voor verschillen in moeilijkheidsgraad tussen toetsen (Hofstee, 1977).

Het kennispercentage en de referentiegroep

Het kennispercentage van een toets is het percentage behaalde punten (goede antwoorden) dat op een toets gehaald moet worden voor een voldoende, rekening houdend met de gokcorrectie. Uitgangspunt bij het bepalen van de cesuur van toetsen, is dat de student aangetoond moet hebben een percentage van de stof te beheersen. In de absolute cesuurmethode betekent dit concreet dat wordt uitgegaan van een vast kennispercentage, gecorrigeerd voor gokken. Bij het berekenen van het kennispercentage wordt dan uitgegaan van de *maximaal* te behalen score. Het betreft een kennispercentage gebaseerd op de maximaal te behalen punten op die toets. Deze maximaal te behalen score is een score die in de praktijk vrijwel nooit door iemand gehaald wordt. Dit kan bijvoorbeeld komen door onvolkomenheden in onderwijs en toetsvragen. Onder andere om die reden is de compromis cesuurmethode van Cohen-Schotanus ontwikkeld. De cesuurscore wordt daarbij berekend met behulp van een vast percentage van de maximaal *behaalde* score gecorrigeerd voor gokken. Nadat de cesuurscore is bepaald, wordt deze score omgerekend naar een percentage van de maximaal *te behalen* score gecorrigeerd voor gokken: het kennispercentage behorend bij de cesuur. Hiermee wordt vergelijking mogelijk met kennispercentages van andere toetsen (Cohen-Schotanus, van der Vleuten & Bender, 1996). Daarnaast wordt dit kennispercentage gebruikt om de cesuur voor de herkansing van het vak vast te stellen en zo de moeilijkheidsgraad van de toets gelijk te houden.

De keuze om uit te gaan van de hoogst *bereikte* score bij het berekenen van de cesuur, in tegenstelling tot het theoretische te behalen maximum, heeft implicaties voor de interpretatie van het kennispercentage behorend bij de cesuur. Omdat in de praktijk de *bereikte* score van de referentiegroep vrijwel steeds aanmerkelijk lager ligt dan het theoretisch maximum worden in de praktijk bij lagere cesuurscores ook lagere kennispercentages gevonden. Een laag percentage voor dit kengetal wordt door sommigen geïnterpreteerd als zou 'de lat' voor de studenten te laag liggen, terwijl dit echter meer aanleiding zou moeten

geven om ons zorgen te maken over het feit dat zelfs de beste studenten niet het maximale aantal punten op de toets halen. Mogelijk werkt de cryptische wijze waarop dit kengetal in de analyses geformuleerd wordt, een dergelijke interpretatie in de hand. (cf. Cohen-Schotanus, 1995)

Elimineren van vragen

De moeilijkheidsgraad van een item in een toets wordt gedefinieerd als de proportie studenten die het juiste antwoord hebben gegeven. Dit wordt de p-waarde van een toetsvraag genoemd. Slecht scorende vragen zijn vragen met een p-waarde kleiner dan de gokkans. Een vraag waarvan de p-waarde onder de gokkans ligt, discrimineert erg slecht. De oorzaken van het feit dat de vraag zo slecht discrimineert tussen goede en zwakke studenten kunnen verschillen. Wellicht is de vraag extreem moeilijk, wellicht is de stof niet behandeld of is een van de overige antwoordalternatieven plausibeler in de ogen van studenten. Op basis van de psychometrische redenen zou het dus verstandig zijn om een vraag met een p-waarde kleiner dan de gokkans te verwijderen, maar in een toets kunnen niet ongelimiteerd vragen verwijderd worden. Een toets zal ook tot op bepaald niveau representatief moeten zijn voor de behandelde collegestof. Bij het elimineren van vragen zal dan ook altijd een afweging gemaakt moeten worden tussen de psychometrische kwaliteit en de representativiteit van de toets.

Onderzoeksvragen

In deze studie wordt de invloed onderzocht van het elimineren van vragen en het verkleinen van de referentiegroep op de uitkomsten van toetsen bij gebruik van de cesuurmethode Cohen-Schotanus en op de slagingspercentages. Deze studie geeft een beeld van de effecten op het percentage studenten dat slaagt voor een tentamen als gevolg van de veranderingen in de cesuur. De volgende onderzoeksvragen staan hierbij centraal:
Hoe groot is het effect van het elimineren van slecht scorende vragen (i.e. vragen met een p-waarde kleiner dan de gokkans) op de slagingspercentages?
Hoe groot is het effect van het verkleinen van de referentiegroep van 5% naar 1% best scorende studenten op de bereikte kennispercentages en op de slagingspercentages?

Methode van onderzoek

Context

De toetsen in het eerstejaars blokonderwijs van de opleiding Geneeskunde werden geëvalueerd volgens de cesuurmethode Cohen-Schotanus. Men heeft daarbij een aantal eigen accenten aangebracht, namelijk met betrekking tot samenstelling van de groep beste presteerders (het 95e percentiel als referentiegroep) en met betrekking tot de hoogte van het kennispercentage, namelijk 60%, die gebruikt worden bij de bepaling van de cesuur. In de huidige situatie wordt de cesuurmethode toegepast ongeacht de samenstelling van het tentamen, dus ook bij tentamens met open vragen. Momenteel gebruikt men bij de opleiding de gemiddelde score van de 5% beste studenten als referentiegroep. Niet het theoretisch haalbare maximale aantal punten, maar het in de praktijk behaalde gemiddeld aantal punten van de 5% hoogst scorende studenten wordt als referentiepunt genomen bij de cesuurbepaling. Onderzocht is het effect van de keuze voor het gemiddeld aantal punten van de 1% hoogst scorende studenten op het kennispercentage en op het slagingspercentage.

De opleiding heeft geen bovengrens geformuleerd voor wat betreft het percentage studenten dat zakt bij een tentamen. In theorie maakt het niet uit hoe slecht een cohort een tentamen heeft gemaakt, er wordt niet ingegrepen op de bepaling van de cesuur door de coördinator. De beslissing tot het schrappen van slecht discriminerende vragen ligt bij de blokcoördinator, evenals andere beslissingen, bijvoorbeeld om bij nader inzien een tweede antwoordmogelijkheid ook als goed te rekenen.

Veranderingen in de cesuur heeft effect op het percentage studenten dat slaagt voor een tentamen.

Steekproef en analyse

Voor dit onderzoek zijn alle bloktoetsen (n=39) uit het studiejaar 2008–2009 van de opleiding Geneeskunde van het Leids Universitair Medisch Centrum (LUMC) betrokken, voor zover het de eerste gelegenheid van een toets betrof. Herkansingen zijn buiten beschouwing gelaten omdat de cesuurbepaling volgens de methode Cohen-Schotanus uitsluitend plaatsvindt bij de eerste gelegenheid en bij voldoende grote aantallen studenten. Iedere toets werd meegenomen in de analyse. Voor het onderzoek zijn de tentamens in drie categorieën onderverdeeld, namelijk: 1) tentamens met uitsluitend open vragen, 2) tentamens met uitsluitend gesloten vragen en 3) tentamens die bestaan uit een combinatie van open en gesloten vragen. Tijdens de analyse van de tentamens werden naast de p-waarde (proportie goede antwoorden) ook andere de psychometrische kenmerken bepaald, zoals, rir-waarde (item-restcorrelatie) en alfa (schatting van de betrouwbaarheid). Op deze manier werd bekend wat de moeilijkheidsgraad van een vraag was en in hoeverre de vraag samenhing met de andere vragen uit de toets.

Resultaten

Eliminatie van slecht discriminerende vragen

In de huidige situatie is het de blokcoördinator die, op basis van de gegevens uit de analyses van de vragen, beslist of een 'slechte' vraag geschrapt wordt of gehandhaafd blijft. In deze studie worden de gevolgen geschetst wanneer als regel wordt ingesteld: *altijd eliminatie van items waarbij p-waarde lager is dan gokkans.* Van de 39 onderzochte toetsen hadden 23 toetsen een vraag waarvan de p-waarde lager was dan de gokkans. Deze vragen waren als volgt verdeeld: zeven toetsen waarbij één vraag aan het criterium voldeed, acht toetsen met twee, zeven met drie en één met vier vragen. Het ging hierbij om vragen waarvan de docent besloot de vraag, ondanks de lage p-waarde, niet te laten vervallen.

Het verplicht laten vervallen van vragen met een p-waarde lager dan de gokkans leidt in de praktijk tot een daling met gemiddeld twee vragen van een toets. Daardoor krijgen sommige studenten ook een of twee punten minder. Wanneer vragen met een p-waarde kleiner dan de gokkans verplicht verwijderd worden in een tentamen met gesloten vragen, stijgt het kennispercentage behorend bij de cesuur met gemiddeld 1,25% en het percentage studenten dat slaagt, daalt met 0,24%. Analyse van

Verplichte eliminatie van vragen met een p-waarde lager dan de gokkans leidt tot een lichte daling van het aantal geslaagde studenten.

tentamens met een mengvorm van open en gesloten vragen, laat een vergelijkbaar beeld zien, namelijk het kennispercentage stijgt met gemiddeld 0,49% en het percentage studenten dat slaagt voor het tentamen daalt met 0,41%.

Samenvattend, verplichte eliminatie van vragen met een p-waarde lager dan de gokkans zal leiden tot een lichte daling van het aantal geslaagde studenten.

Referentiegroep

De groep beste presteerders wordt momenteel gedefinieerd als het 95e percentiel. Tabel 1 geeft de gemiddelde kengetallen van tentamens Geneeskunde (2008-2009) met verschillende vraagtypen. De kolommen 2 en 3 geven de situatie weer van de huidige werkwijze: met een referentiegroep die bestaat uit 5% best presterende studenten. De kolommen 4 en 5 geven het resultaat indien de cesuur bepaald wordt op grond van het gemiddelde van de 1% best presterende.
Wanneer de cesuur bepaald wordt op grond van de 1% best presterende studenten, blijkt de cesuur in één à twee ruwe scorepunten omhoog te gaan. De consequenties daarvan verschillen per toets, en worden mede bepaald door afrondingsverschillen. Een verandering in de referentiegroep van 5% naar 1% best presterende studenten, zorgt ervoor dat in verhouding een groter aantal studenten (ruwweg tussen 7 en 14 procent) voor de toets zakt. In de meeste gevallen zal dat, al het andere onveranderd gelaten, als consequentie hebben dat het percentage 'gezakt' boven de 30% uitstijgt.

Twee praktijkvoorbeelden

Een toets met gesloten vragen
De toets bevat drie vragen met drie antwoordalternatieven, 41 vragen met vier alternatieven en vier vragen met vijf alternatieven. Een tweetal vragen is achteraf door de blokcoördinator verwijderd. Bij een tiental vragen konden psychometrische kanttekeningen geplaatst worden. Deze toets is afgenomen bij 350 studenten.

Wat bij deze toets opvalt, is dat er een verschil bestaat tussen het theoretische maximum en de maximumscore die in de praktijk behaald werd (Tabel 2). Eveneens valt op dat het kennispercentage laag is. De cesuur wordt bepaald door het gemiddelde van de 19 best presterende studenten: 39,32. Deze waarde maakt duidelijk dat in dit specifieke geval er een groot verschil gevonden wordt tussen de gemiddelde score van de beste 5% studenten en het theoretisch maximum dat bijna negen punten hoger ligt. Wanneer een strengere waarde voor het referentiepunt wordt aangehouden, namelijk het 99e percentiel, dan wordt de gemiddelde score van de topgroep de waarde 42, gebaseerd op 4 best presterende studenten. De cesuur stijgt daarbij naar 30 en het kennispercentage naar 49,93%. Het percentage gezakt stijgt echter met ruim 14% tot boven de 40%.

In dit concrete voorbeeld is er een aanzienlijk gat tussen de (theoretisch) maximaal te behalen score van 48 punten en de score van de best scorende student: 43. Eveneens is er een aanzienlijk verschil tussen het gemiddelde van het 95e percentiel en dat van het 99e percentiel (resp. 39,32 en 42,0 punten). De omvang van deze verschillen is onder meer het gevolg van de aanwezigheid van een aanzienlijk

Tabel 1: *Gemiddelde kengetallen van tentamens Geneeskunde (2008-2009) met open vragen, gesloten vragen of een mengvorm van open en gesloten vragen.*

Soort vragen	Gezakt bij cesuur 95e percentiel (%)	Kennispercentage bij cesuur 95e percentiel (%)	Gezakt bij cesuur 99e percentiel (%)	Kennispercentage bij cesuur 99e percentiel (%)
Open	15,55	49,26	23,28	52,62
Gesloten	22,98	50,96	30,76	54,49
Mixed	21,08	50,79	25,63	53,09

Tabel 2: *Een toets met 50 gesloten vragen.*

Referentiegroep	*5%*	*1%*
Totaal aantal punten	50	50
De max. haalbare score na eliminatie	48	48
Max. gehaalde score	43,0	43,0
Gokscore	12,1	12,1
Aantal in referentiegroep	19	4
Gemiddelde referentiegroep	39,42	42,00
Cesuur	28	30
Kennispercentage	44,37	49,93
Percentage geslaagden	73,37	59,21

aantal 'slechte' vragen in deze toets. Dit betreft geen vraaginhoudelijke kritiek, maar heeft betrekking op de psychometrische kenmerken van de vragen. De consequentie van een keuze voor een strengere selectie van de groep best scorende studenten resulteert voor deze concrete toets in een slaagpercentage van minder dan 60%.

Een toets met gesloten en open vragen

De gesloten vragen kennen vier antwoordalternatieven. Met de open vragen zijn in totaal 40 punten te halen. Het aantal punten varieerde per open vraag tussen de vier en acht punten. Er zijn geen vragen verwijderd, bij twee vragen konden psychometrische kanttekeningen gemaakt worden. Deze toets is afgenomen bij 300 studenten
In dit voorbeeld stijgt het 'kennispercentage' naar 54,84 en het percentage 'gezakt' stijgt met 12 studenten (bijna 4%) tot boven de 30% wanneer voor de bepaling van de cesuur het 99e percentiel referentiepunt wordt aangehouden (Tabel 3). Ook in dit voorbeeld wordt het theoretische maximum van de toets niet gehaald, maar zijn er aanzienlijk minder vragen met psychometrische kanttekeningen. Het effect van de verhoging van de cesuur met twee ruwe score punten is hier minder dramatisch, hierdoor zullen een twaalftal studenten extra zakken.
Deze twee praktijkvoorbeelden illustreren dat de keuze voor een strengere referentiegroep (het 99e percentiel) een stijging van het kennispercentage oplevert. Gemiddeld over de geanalyseerde toetsen stijgt het kennispercentage met circa 3%.

Tabel 3: *Een toets met 50 gesloten vragen en zeven open vragen.*

Referentiegroep	***5%***	***1%***
Totaal aantal punten	90	90
De max. haalbare score na eliminatie	90	90
Max. gehaalde score	85,5	85,5
Gokscore	12,5	12,5
Aantal in referentiegroep	16	4
Gemiddelde referentiegroep	80,81	83,75
Cesuur	53	55
Kennispercentage	52,26	54,84
Percentage geslaagden	72,29	68,47

Discussie

In dit onderzoek presenteren we een aantal kengetallen van de Geneeskunde toetsen uit het studiejaar 2008-2009. Inzichtelijk wordt gemaakt welke verschuivingen optreden in de cesuur, het kennispercentage en het percentage studenten dat slaagt, wanneer één van de parameters van de cesuurbepaling wordt gewijzigd.

Eliminatie van vragen

Beschreven wordt wat de gevolgen zijn van het standaard verwijderen van alle vragen met een p-waarde die lager is dan de kanswaarde van de vraag. Het elimineren van slecht scorende vragen, zal op de cesuur en op het kennispercentage een beperkt effect hebben doordat het gemiddelde van de referentiegroep niet of nauwelijks verandert. Gemiddeld vervallen twee vragen. Wanneer de cesuurbepaling onveranderd blijft op het 95e percentiel, zal het effect op de cesuur beperkt zijn. Het is immers niet te verwachten dat het gemiddelde van een groep van ongeveer 16 studenten substantieel verandert door verwijdering van twee vragen van dit type. Het verwijderen van dit type vragen heeft in de meeste gevallen nauwelijks tot geen effect op de scores van de best presterende studenten omdat zij op deze 'slechte' vragen net als hun medestudenten voor de verkeerde afleiders kiezen. In de analyse van toetsvragen blijft het weloverwogen oordeel van de coördinator van belang omdat er meerdere overwegingen een rol spelen bij het al dan niet achteraf verwijderen van een vraag. Hier dient bij aangetekend te worden dat er meerdere criteria zijn waarop de kwaliteit van een vraag kan worden beoordeeld dan uitsluitend de p-waarde.

Referentiegroep

Ook wordt beschreven wat de gevolgen zijn wanneer de 1% beste studenten de referentiegroep bepaalt in vergelijking met de huidige referentiegroep. Wanneer de cesuur bepaald wordt op grond van de 1% best presterende studenten, blijkt dat de scores van vier of minder studenten de waarde van de cesuur bepalen tegen circa zestien studenten uit de huidige referentiegroep. In deze kleine 'kopgroep' van studenten komen bij verschillende toetsen deels dezelfde uitblinkers terug. De cesuur in ruwe scorepunten gaat dan één à twee ruwe scorepunten omhoog.

De consequenties verschillen per toets, en kunnen, mede door afrondingsverschillen, aanzienlijk zijn. Een keuze voor strengere selectie van de referentiegroep 'best presterende studenten' zal resulteren in een hogere cesuur. Deze verhoging van de zak-/slaaggrens brengt met zich mee dat een in verhouding groter aantal studenten (ruwweg tussen 7 en 14 procent) voor de toets zakt. In de meeste gevallen zal dat, al het andere onveranderd gelaten, als consequentie hebben dat het percentage 'gezakt' boven de 30% uitstijgt (zie de beschreven praktijkvoorbeelden). De hogere cesuur werkt door in het kengetal kennispercentage, zij het beperkt (gemiddeld ruim drie procent).

Afsluitende opmerking

Het kennispercentage wordt berekend om de moeilijkheidsgraad (cesuur) van de herkansing van de toets te bepalen. De interpretatie van de hoogte van het percentage dient echter met zorgvuldigheid te geschieden: afgezet tegen het absolute kennispercentage van 60% geeft het inzicht in de mate van de 'perfectie' van de toets; in de mate waarin de beste studenten erin zijn geslaagd de maximale score te benaderen. De verwarring die bij het gebruik van deze term op de loer ligt, geeft wellicht aanleiding om over een naamswijziging na te denken. Voor gebruik in een discussie of 'de lat' voor studenten hoger gelegd moet worden is het kengetal ongeschikt. Het kennispercentage geeft wel de mogelijkheid om toetsen onderling te vergelijken.

De keuze van de methode van cesuurbepaling heeft voor studenten ingrijpende gevolgen. Het blijkt dat de keuze voor de cesuur een wezenlijke invloed heeft op de doorstroming van studenten in hun studie, terwijl het geen invloed lijkt te hebben op de kwaliteit van de afgestudeerden (Cohen-Schotanus & van der Vleuten, 2010). Het is dus verstandig om goed na te denken over de cesuurmethode die gekozen wordt in de opleiding.

Personalia

Drs. Gijs van Duijn is onderwijskundig adviseur aan de Universiteit Leiden, ICLON, Afdeling Hoger Onderwijs.
Dr. Arnout Jan de Beaufort is opleidingscoördinator geneeskunde bij het Leids Universitair medisch Centrum.
Dr. Roeland van der Rijst is universitair docent aan de Universiteit Leiden, ICLON, Afdeling Hoger Onderwijs.

Literatuur

Cohen-Schotanus, J., Vleuten, C.P.M. van der, & Bender, W. (1996). Een betere cesuur bij tentamens. *Onderzoek van Onderwijs, 25*, 54-55.

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

Duijn, G. van (2010). *De cesuurmethode Cohen-Schotanus binnen de opleiding Geneeskunde: De invloed van een aantal factoren op de uitkomsten van toetsten.* Rapportnummer 152. Leiden: Universiteit Leiden/ ICLON.

Gruijter, D.N.M. de (2008). *Toetsing en toetsanalyse.* Leiden: Universiteit Leiden/ ICLON.

Hambleton, R.K., & Pitoniak, M.J. (2006). Setting performance standards. *Educational Measurement, 4*, 433-470.

Herraets, J. (1999). Het onverstoorbare gelijk van Wynand Wijnen, onderwijsvernieuwer. *Tijdschrift voor Hoger Onderwijs, 2.*

Hofstee, W.K.B. (1977). Cesuurprobleem opgelost. *Onderzoek van Onderwijs, 6*, 6-7.

Wijnen, W.H.F.W. (1971). *Onder of boven de maat. Een methode voor het bepalen van de grens voldoende/onvoldoende bij studietoetsen.* Proefschrift. Groningen: Universiteit Groningen.

Wijnen W.H.F.W. (1986). Toetsprogrammering en rendementsverlies. *Onderzoek van Onderwijs, 15*, 10–5.