

De vraagtekenoptie bij juist/onjuist-items in toetsen: goed of fout?

A.M.M. Muijtjens, H. van Mameren, R.J.I. Hoogenboom, J.L.H. Evers, J.P.M. Geraedts, K.M.L. Leunissen, C.P.M. van der Vleuten

Samenvatting

Toevoeging van een vraagtekenoptie kan de meetfout tengevolge van gokken bij toetsen met juist/onjuist-items tegengaan. Zonder deze gokcorrectie vormt het aantal goed beantwoorde vragen de score, met gokcorrectie is de score het aantal goed-minus-fout-beantwoorde items. De gepubliceerde resultaten van vergelijkingen tussen scores met en zonder vraagtekenoptie zijn niet eensluidend. Om na te gaan welke scoringsmethode te verkiezen is voor kennistoetsen zijn in een bloktoets beide methoden gebruikt. Studenten maakten de toets eerst met vraagtekenoptie; vervolgens zetten zij de vraagtekens om in een juist/onjuist-antwoord. De vergelijking tussen de scores met en zonder vraagtekenoptie wees uit dat gokcorrectie bias veroorzaakt: minder gokken leidt tot een lagere score. Daarentegen is de betrouwbaarheid lager zonder gokcorrectie, al vermindert dit verschil naarmate meer toets-items beter op het blok aansluiten. Bij de keuze tussen beide methoden moet men een afweging maken tussen minder bias en een hogere betrouwbaarheid, maar ook tussen onderwijskundige voor- en nadelen van beide scoringsmethoden.

Inleiding

Toetsen bestaande uit juist/onjuist-items worden frequent gebruikt om het kennisniveau van medische studenten vast te stellen. Als de toets geen vraagtekenoptie bevat, ligt het voor de hand om de toetsuitslag vast te stellen aan de hand van het percentage goed beantwoorde

items, een methode die in dit artikel verder aangeduid zal worden als correctscore (Engels: number-right scoring). Een goed antwoord kan verkregen zijn door het aanwenden van (partiële) kennis maar ook door een gunstig uitgevallen gok. De goedscore verkregen door raden, staat los van het kennisniveau van de student en verhoogt dus de fout in de kennismeting. Met behulp van de score met gokcorrectie (Engels: formula scoring) wordt gestreefd naar reductie van de meetfout. Bij deze methode wordt de vraagtekenoptie als extra antwoordmogelijkheid toegevoegd en wordt het aantal goede minus het aantal foute antwoorden (G-F-score) als toetsscore gebruikt. Het gebruik van deze methode wordt vaak gebaseerd op de volgende redenering.^{1,2} Een student bezit complete kennis over een item en geeft dus het goede antwoord, of hij is volstrekt onwetend en gokt. Dan is het aantal foute antwoorden te beschouwen als een schatting van het aantal goede antwoorden behaald met gokken. De G-F-score schat dus het aantal goed gescoorde items op basis van kennis. Lord wijst erop dat de bij het gebruik van de G-F-score gehanteerde aanname het bestaan van partiële kennis negeert en daarom nauwelijks te verdedigen is.³⁻⁵ Lord laat echter ook zien dat voor de score met gokcorrectie volstaan kan worden met een andere aanname: de vraagtekens bij gebruik van instructies voor de score met gokcorrectie, worden bij het gebruik van correctscore-instructies vervangen door een willekeurige gok.

Op theoretische gronden kan worden berekend dat de score met gokcorrectie zou moeten leiden tot een verhoogde validiteit en

betrouwbaarheid.^{6,7} Empirische onderzoeken geven echter geen eensluidende resultaten te zien: bij toepassing van correctscore-richtlijnen wordt afname maar ook toename van betrouwbaarheid gevonden.⁸⁻¹⁰ Diverse empirische studies geven aan dat studenten in het algemeen geneigd zijn om met een vraagteken te antwoorden bij items waarvoor ze een grotere dan op puur toeval berustende kans hebben om het goede antwoord te kiezen.¹¹⁻¹³ Lord merkte op dat deze studies, vanwege inadequate invulinstructies, tekortschoten om zijn aanname met betrekking tot de score met gokcorrectie te testen.³ Cross & Frary en ook Bliss volgden de aanwijzingen van Lord, maar vonden desalniettemin een verhoogde kans dat een eerder met een vraagteken beantwoord item in tweede instantie goed beantwoord wordt.¹⁴⁻¹⁵ In meer recente publicaties worden deze resultaten bevestigd.¹⁶⁻¹⁸ Kennelijk zijn studenten niet goed in staat om onderscheid te maken tussen puur gokken en "geïnformeerd gokken", waarbij partiële kennis wordt aangewend.^{11-13, 16-19} Bliss en eerder ook al Votaw vonden dat de score met gokcorrectie vooral de goede studenten benadeelt, omdat die meer geneigd zijn de testinstructies nauwkeurig op te volgen.^{11, 15}

Naast psychometrische kenmerken zijn ook andere, meer onderwijskundige, aspecten van belang bij de vergelijking tussen de score met gokcorrectie en de correctscore. Bij testen met scoring met gokcorrectie wordt de student gestimuleerd om na te gaan of hij genoeg kennis over een item heeft. Daardoor zal hij eerder lacunes in zijn kennis op het spoor komen. Daarnaast biedt de vraagtekenoptie de docent informatie omtrent de kwaliteit van een item. Een relatief groot aantal vraagtekens in de antwoorden op een item is een indicatie dat het item niet behoort tot het geëxamineerde kennisdomein, dat het onderwerp ten onrechte niet in het onderwijs aan bod is gekomen, of dat de formulering van de vraag niet deugt. Bij score met gokcorrectie wordt een student niet

aangemoedigd om te gokken als hij het antwoord niet weet. Harden gaat in op de vraag of dit nou goed of slecht is. De tegenover elkaar staande meningen kunnen als volgt worden samengevat: "medici moeten niet reageren met gokken als ze met een lacune in hun kennis worden geconfronteerd" versus "in de medische praktijk moet een arts regelmatig een beslissing nemen op basis van gedeeltelijk onvolledige informatie".²⁰

In het licht van de controverse omtrent scoring met en zonder gokcorrectie is onderzocht welke methode de voorkeur verdient bij kennistoetsing in een medisch curriculum. De volgende vragen stonden daarbij centraal:

- Is er meer gevaar voor het optreden van bias in de kennismeting bij de score met gokcorrectie?
- Wordt partiële kennis beter gemeten met de correctscore?
- Is de toetsuitslag betrouwbaarder bij de score met gokcorrectie?
- Geeft de score met gokcorrectie meer informatie over de itemkwaliteit?
- Worden de betere studenten benadeeld bij gebruik van de score met gokcorrectie?

Bovenstaande vragen zijn onderzocht in een experiment waarbij beide scoringsmethoden toegepast zijn bij een kennistoets (bloktoets) van een onderwijsblok in het derde jaar.

Het effect dat een scoringsmethode heeft op het studiegedrag van studenten is eveneens van belang. Onderzoek hiernaar zou echter een longitudinale studie vergen, hetgeen buiten de kaders valt van het hier gepresenteerde onderzoek. Persoonlijkhedenkenmerken die mogelijk van invloed kunnen zijn op het invulgedrag van de studenten zijn eveneens niet betrokken in het huidige onderzoek.

Methode

Het experiment is uitgevoerd bij de reguliere bloktoets van het vijfde onderwijsblok in het derde jaar met als thema *Pijn*. De studenten

werd gevraagd om de items zowel op de gebruikelijke wijze (mèt vraagtekenoptie), als zonder gebruik van de vraagtekenoptie te beantwoorden.

Proefpersonen

Proefpersonen in het experiment waren de 151 derdejaars geneeskundestudenten van de Faculteit der Geneeskunde in Maastricht die in 1996 hebben deelgenomen aan de reguliere bloктоets van het vijfde onderwijsblok in het derde jaar.

Metingen

Score met gokcorrectie en correctscore

De gebruikte bloктоets is een kennistoets, bestaande uit 169 juist/onjuist-items. De studenten werd gevraagd om elk item in eerste instantie op de tot nu toe gebruikelijke wijze te beantwoorden, dat wil zeggen met gebruik van de vraagtekenoptie en wetend dat de uitslag van de toets gebaseerd wordt op de G-F-score. Daarnaast werd verzocht om in tweede instantie de met een vraagteken beantwoorde items alsnog met juist/onjuist te beantwoorden. Beide beantwoordingen vonden plaats in één afnamesessie, zodat tussentijdse kennisuitwisseling vrijwel uitgesloten was. In verband met de tweede beantwoording werd de tijdsduur van de toets met een uur verlengd. De tweede set antwoorden werd samengevoegd met de juist/onjuist-antwoorden van de overige items en de bijbehorende score werd beschouwd als het resultaat van de beantwoording van de toets bij gebruik van correctscore-instructies. Voor de twee scoringsmethoden werd als onvoldoende/voldoende grens het gemiddelde minus de standaarddeviatie van de bijbehorende score gebruikt. Om serieus invullen te bevorderen werd afgesproken dat de voor een student gunstigste van de twee uitslagen zou gelden als examenresultaat. Om meer inzicht te krij-

gen in het risico van bias en het meten van partiële kennis is nagegaan hoe de G-F-score per student verandert bij overgang van score met gokcorrectie naar correctscore. Om dezelfde reden is nagegaan hoe groot de kans is op een goed antwoord bij het beantwoorden van vraagteken-items.

Itemrelevantie bepaald door studenten

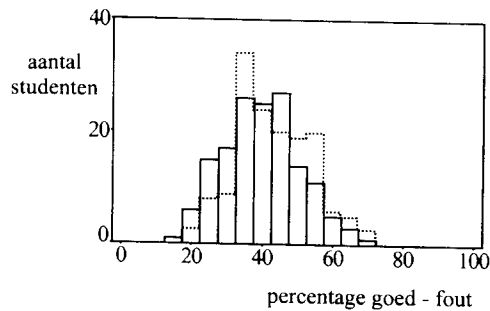
De studenten werd verzocht om aan te geven of een item naar hun mening al dan niet in het onderwijs was behandeld. Nagegaan is of de item-vraagtekenscore (het percentage vraagtekens in de beantwoording van een item) informatie geeft over deze itemrelevantie.

Itemrelevantie bepaald door docenten

Aan vier docenten die deel uitmaakten van de planningsgroep van het betreffende blok, werd gevraagd om onafhankelijk van elkaar voor elk item aan te geven of "een student die een voldoende verdient het goede antwoord zeker zou moeten weten". Het aantal positieve beoordelingen (0, 1, 2, 3 of 4) geeft aan in hoeverre een item tot de kern van het blok behoort. Een selectie van items met een hoge relevantiescore vormt een beter op het blok aansluitende deoltoets. Nagegaan is hoe de betrouwbaarheid bij score met gokcorrectie en bij correctscore verandert wanneer een dergelijke deoltoets gebruikt zou worden. De relevantiescore van de docenten is daarnaast gebruikt als gouden standaard voor de kwaliteit van een item. Nagegaan is in hoeverre de item-vraagtekenscore (bij score met gokcorrectie) en de item-foutscore (bij correctscore) informatie geven over de itemkwaliteit.

Algemeen kennisniveau studenten

Het kennisniveau van de studenten is bepaald op basis van de resultaten, behaald in acht opeenvolgende voortgangstoetsen (september



Figuur 1. Histogram van het percentage goed-minus-fout beantwoorde items bij toepassing van de twee onderzochte scoringsmethodes gokcorrectie en correctscore (getrokken lijn respectievelijk stippellijn).

1994 tot en met juli 1996). Per toets werden de studentscores gerangschikt en van een rangnummer voorzien. Het gemiddelde rangnummer per student is gebruikt als maat voor algemeen kennisniveau. Deze maat is gebruikt om na te gaan in hoeverre de betere studenten benadeeld worden door het toepassen van score met gokcorrectie.

Data-analyse

Voor het bepalen van de sterkte van de relatie tussen continue variabelen is de Pearson correlatiecoëfficiënt gebruikt. Deze werd als niet-significant (NS) beschouwd bij een overschrijdingskans $p > 0.05$. Hetzelfde significantieniveau is gehanteerd voor de paired t-test bij het toetsen van het verschil tussen de G-F-score bij correctscore en bij score met gokcorrectie. Nagegaan is hoe goed de item-vraagtekenscore en de item-foutscore onderscheid kunnen maken tussen items met een lage en een hoge itemrelevantie. Daartoe zijn de items opgesplitst in twee klassen: laagrelevante items (0, 1 of 2 positieve beoordelingen door docenten) en hoogrelevante items (3 of 4 positieve beoordelingen). Met logistische regressie is bepaald welke grenswaarde in de item-vraagtekenscore het beste onderscheid maakt tussen

de laag- en hoogrelevante items.²¹ Hetzelfde is gedaan voor de classificatie op basis van de item-foutscore. De kwaliteit van de resulterende classificatie komt tot uitdrukking in het percentage correct geclassificeerde items.

Resultaten

Verdeling itemrelevantie en juist-onjuist sleutel

De gebruikte toets bestaat uit 95 juist-items en 74 onjuist-items. Aan de hand van de relevantiescore van de docenten zijn de items verdeeld in twee categorieën: items met een lage en items met een hoge relevantie (relevantiescore respectievelijk ≤ 2 en ≥ 3). Van de items in de toets valt 69% in de categorie hoge relevantie (tabel 1, derde rij). Voor de juist en de onjuist versleutelde items is de verhouding hoog/laag vrijwel identiek (tabel 1, eerste en tweede rij, percentage hoog respectievelijk 68% en 70%).

Bias, meten partiële kennis

Van het percentage G-F per student, behaald bij de score met gokcorrectie wordt in figuur 1 de verdeling getoond (getrokken lijn). De verdeling van dezelfde variabele bij de correctscore is aangegeven met een stippellijn. Duidelijk is dat bij de correctscore in het algemeen hogere scores worden behaald. De G-F-score verschuift gemiddeld 3.4% (paired t-test, $p < 0.001$), van 39.9% bij score met gokcorrectie naar 43.3% bij correctscore. De standaarddeviatie van de verdelingen bedraagt respectievelijk 10.7% en 10.9%. De verschuivingen in de onvoldoende/voldoende kwalificaties bij overgang van score met gokcorrectie naar correctscore bedroegen 7.9% van onvoldoende naar voldoende, en 2.6% in omgekeerde richting. Het totaal aantal onvoldoende kwalificaties daalde daardoor van 18.5% naar 13.3%.

In totaal was 22.6% van de eerste antwoorden een vraagteken. In tabel 2 wordt de kruis-

Tabel 1. Frequentietabel van antwoordsleutel en relevantiescore van docenten, voor alle items in de toets (aantal en percentage van het totaal aantal items).

Sleutel	Relevantiescore ¹⁾		Totaal
	Laag (≤ 2)	Hoog (≥ 3)	
Juist	30 (18%)	65 (38%)	95 (56%)
Onjuist	22 (13%)	52 (31%)	74 (44%)
Totaal	52 (31%)	117 (69%)	169 (100%)

¹⁾Aantal positieve beoordelingen van het item door vier docenten.

Tabel 2. Kruistabel van de antwoordsleutel en het gegeven antwoord bij items die in eerste instantie met een vraagteken werden beantwoord, in percentages van het totaal aantal vraagtekenantwoorden (n=5763).

Sleutel	Antwoord		Totaal
	Juist	Onjuist	
Juist	38.0	21.0	59.1
Onjuist	21.3	19.6	40.9
Totaal	59.3	40.7	

tabel getoond van de antwoordsleutel en het gegeven antwoord voor alle vraagteken-items. Van de tweede antwoorden bij de vraagteken-items was 57.6% goed. Worden de percentages voor de juist en onjuist versleutelde vragen apart beschouwd, dan blijkt dat de G-F-winst behaald werd bij de juist versleutelde items, terwijl de onjuist-items een licht verlies opleverden. Dat houdt echter niet in dat bij de beantwoording van de juist-items informatie wordt gebruikt en bij de onjuist-items niet. Om dat vast te stellen moet de scheve verdeling van de marginalen in de tabel in aanmerking worden genomen. Immers alleen al op grond daarvan mag verwacht worden dat de G-F-winst bij de juist-items hoger is, zelfs al zou de beantwoording op puur toeval berusten. Voor een correcte interpretatie van de tabel dienen de in het experiment gevonden percentages te worden vergeleken met de op basis van toeval verwachte percentages. In de hierna volgende

analyse worden de resultaten van die aanpak gepresenteerd.

Het op grond van toeval, dat wil zeggen bij afwezigheid van enige associatie tussen sleutel en antwoord, te verwachten percentage goed bedraagt 51.7%. Dit percentage wordt verkregen door de betreffende marginale rij- en kolompercentages met elkaar te vermenigvuldigen ($0.593 \times 0.591 + 0.407 \times 0.409$), met andere woorden: er wordt rekening gehouden met de scheve verdeling van zowel de sleutel als de antwoorden.²¹ Er is dus sprake van 5.9% meer goede antwoorden dan verwacht mocht worden op grond van toeval. Uit tabel 2 valt daarnaast af te lezen dat voor de juist en onjuist versleutelde items het percentage goede antwoorden respectievelijk 64.4% (38/59.1) en 47.9% (19.6/40.9) bedroeg. De derde rij van tabel 2 laat zien dat vaker het antwoord juist is gekozen dan het antwoord onjuist (59.3 % vs. 40.7%). Bij blind gokken is de verwachting dat diezelfde verhouding bij zowel de juist als de

Tabel 3. Percentage goede antwoorden bij de in eerste instantie met een vraagteken beantwoorde items, voor alle items in de toets en voor items met een lage (≤ 2) resp. hoge (≥ 3) relevantiescore van docenten.

Items	Sleutel	Antwoord goed		
		Waargenomen	Verwacht ¹⁾	Vershil
Alle (n=169)	Juist	64.3	59.3	5.0
	Onjuist	47.9	40.7	7.2
Relevantiescore docenten: Laag (n=52)	Juist	66.6	60.8	5.8
	Onjuist	48.9	39.2	9.7
Relevantiescore docenten: Hoog (n=117)	Juist	62.3	58.2	4.1
	Onjuist	47.3	41.8	5.5

¹⁾Verwacht percentage bij statistische onafhankelijkheid tussen het gegeven antwoord en de antwoordsleutel

onjuist versleutelde items wordt gerealiseerd. Voor beide categorieën sleutels werd echter een percentage goed gevonden dat hoger was dan deze verwachting (tabel 3, eerste en tweede rij). Voor zowel de juist als de onjuist versleutelde items is er dus sprake van winst, die duidt op het gebruik van informatie bij de tweede beantwoording. De vergelijking van het waargenomen percentage goede antwoorden met de toevalsverwachting voor items met een lage respectievelijk hoge docentrelevantie levert vergelijkbare resultaten op (tabel 3, derde tot en met zesde rij).

Opmerking: De percentages in tabel 2 hebben betrekking op de antwoorden op 5763 vraagteken-items die door 151 studenten zijn gegeven, hetgeen inhoudt dat de antwoorden niet allemaal statistisch onafhankelijk zijn. Om die reden is een chi-kwadraat test voor deze gegevens ongeschikt. Bovenstaande interpretatie van de tabel op basis van verwachtingswaarden blijft echter wel valide.

Betrouwbaarheid

In tabel 4 zijn de betrouwbaarheden (Cronbachs alfa) van het percentage G-F voor score met gokcorrectie en correctscore met elkaar vergeleken. De eerste rij geeft de resultaten

voor de complete toets. Rij twee geeft de betrouwbaarheid voor de deelttoets van items (69% van de oorspronkelijke toets) die door ten minste drie van de vier geraadpleegde docenten beoordeeld werden als behorend tot de kern van het blok. Regel drie tenslotte toont de resultaten voor een deelttoets van 39% van de items waarvoor alle vier de docenten vonden dat het kernitems betrof. De getoonde betrouwbaarheidswaarden zijn gecorrigeerd voor verschillen in aantallen items en hebben betrekking op deelttoetsen ter grootte van de complete toets (n=169 items). De laatste kolom toont de relatieve toename van het aantal items dat nodig is om het verschil in betrouwbaarheid tussen score met gokcorrectie en correctscore teniet te doen (berekend volgens de Spearman-Brown formule).

Informatie over itemkwaliteit

De item-vraagtekenscore geeft informatie over het niet behandeld zijn van een item, zoals blijkt uit de significante correlatie tussen de item-vraagtekenscore en het aantal keer dat studenten voor het item "niet-behandeld" hebben aangegeven ($r=0.72$, $p<0.001$). De item-vraagtekenscore geeft ook informatie over de relevantiescore van docenten ($r=-0.45$,

Tabel 4. Betrouwbaarheid (Cronbach's alpha) voor de complete toets en voor de subtoetsen van items met een hoge relevantiescore van docenten, gecorrigeerd voor afgenomen aantal items naar n=169.

Relevantiescore	Betrouwbaarheid voor n=169		
	Score met gokcorrectie	Correctscore items ¹⁾	Fractie extra
Alle (n=169)	0.74	0.66	0.47
≥3 (n=117)	0.75	0.69	0.35
4 (n=66)	0.75	0.70	0.29

¹⁾Fractie extra items nodig om het verschil in betrouwbaarheid op te heffen (berekend met de Spearman-Brown formule)

$p < 0.001$). Bij de correctscore wordt deze rol niet overgenomen door de item-foutscore want die is niet of nauwelijks gecorreleerd met de relevantiescore ($r = -0.15$, NS). In hoeverre de item-vraagtekenscore (bij score met gokcorrectie) en de item-foutscore (bij correctscore) in staat zijn om de laagrelevante items te onderscheiden van de hoogrelevante (itemrelevantie bepaald door docenten ≤ 2 : laag, ≥ 3 : hoog) is onderzocht met logistische regressie. Met de item-vraagtekenscore als classificerende variabele leidde dat tot correcte classificatie van 27% van de laagrelevante en 87% van de hoogrelevante items. De item-vraagtekenscore is dus niet goed bruikbaar om de laagrelevante items te identificeren, want ze worden in slechts een kwart van de gevallen als zodanig herkend. De relatie tussen de itemrelevantie en de item-foutscore bij correctscore is zo zwak dat classificatie op basis van deze variabele helemaal niet lukt: alle items worden ingedeeld bij de klasse hoog.

Benadelen betere studenten

Een eventuele benadeling van de betere studenten door het hanteren van met gokcorrectie is onderzocht door na te gaan of de toename in G-F-score bij overgang op correctscore voor die groep studenten het grootst is. De toename in de G-F-score bleek echter niet significant

gecorreleerd te zijn met het kennisniveau van de student ($r = -0.09$, NS).

Discussie

Verdeling itemrelevantie en juist-onjuist sleutel

Naar het oordeel van de docenten bevat de gebruikte toets ongeveer 70% hoogrelevante items. Dit percentage blijkt voor de juist en onjuist versleutelde items vrijwel hetzelfde te zijn. Kennelijk is het voor beide categorieën items even moeilijk om hoogrelevante items aan te maken.

Er zitten meer juist versleutelde dan onjuist versleutelde items in de gebruikte toets. Indien dat algemeen het geval is bij de bloktoetsen en studenten daarvan op de hoogte zijn, dan leidt dat zeer waarschijnlijk tot strategisch gedrag: bij twijfel wordt gekozen voor het antwoord juist. Dit zou mede een verklaring kunnen zijn voor de voorkeur voor het antwoord juist bij het beantwoorden van de vraagteken-items. Nader onderzoek daarvan is gewenst maar valt buiten het kader van de hier gepresenteerde studie. Het hanteren van een vaste 50%-50% verdeling van de sleutel juist en onjuist in deze toetsen zou overigens ook weer strategisch gedrag kunnen uitlokken. Een juist-onjuist verdeling die random varieert rondom 50%-50% is waarschijnlijk geschikter om strategisch gedrag tegen te gaan.

Bias, meten partiële kennis

De verhoging van de G-F-score bij overgang van score met gokcorrectie naar correctscore (figuur 1) en de 57.6% goedscore bij beantwoording van de vraagteken-items zijn in overeenstemming met de resultaten in andere empirische onderzoeken.^{11-15 16 18} De resultaten geven aan dat gokken in het algemeen zal leiden tot een hogere toetsuitslag. Dat houdt in dat van twee studenten met eenzelfde kennisniveau degene die het meeste gokt doorgaans de hoogste score heeft. Verschillen tussen studenten wat betreft het aantal vragen waarop ze gokken in plaats van een vraagteken invullen, leiden dus tot systematische verschillen (bias) in de kennismeting met de score met gokcorrectie. Bij de correctscore is dit risico op bias niet aanwezig omdat iedere student dan alle items met juist/onjuist beantwoordt. Een indicatie voor de grootte van de bias is het verschil in gemiddelde score tussen de twee scoringsmethodes. Dit verschil is 3.4%, een niet geringe waarde in vergelijking met de standaarddeviatie van de scoreverdeling, die ongeveer 11% bedraagt. Dat het een betekenisvol scoreverschil betreft, blijkt ook uit de 10.5% wijzigingen in onvoldoende/voldoende kwalificatie bij overgang van score met gokcorrectie naar correctscore.

Het omzetten van vraagtekens in juist/onjuist-antwoorden blijkt winst op te leveren in de vorm van een hogere goedscore dan op grond van toeval te verwachten is (5% en 7.2% voor juist respectievelijk onjuist versleutelde items). De winst wordt geboekt doordat informatie wordt gebruikt bij de beantwoording. Die informatie kan bestaan uit (partiële) kennis van het betreffende onderwerp, maar ook uit onbedoelde hints in de vraag die leiden naar het goede antwoord. Ervan uitgaand dat de nodige zorgvuldigheid is betracht bij het samenstellen van de toets, geven de verkregen resultaten de indicatie dat partiële kennis vol-

lediger gemeten wordt bij gebruik van de correctscore.

Betrouwbaarheid

Wat de betrouwbaarheid betreft, blijkt in het onderhavige experiment dat de score met gokcorrectie hogere waarden oplevert dan de correctscore (0.74 vs. 0.66), maar ook dat het verschil kleiner wordt naarmate de items beter aansluiten op de blokinhoud (0.75 vs. 0.70 voor de deeltoets van items met de hoogste relevantiescore). Dat geldt ook indien de verschillen uitgedrukt worden in de hoeveelheid extra items die nodig is om de betrouwbaarheid voor de correctscore op gelijke hoogte met die van de score met gokcorrectie te brengen. Voor de totale toets komt dat neer op een uitbreiding met 47%, terwijl voor de deeltoets van hoogrelevante items volstaan kan worden met 29%

Informatie over itemkwaliteit

De significante correlatie van 0.72 van de itemvraagtekenscore en het percentage "niet-behandeld", wijst erop dat de itemvraagtekenscore informatie bevat over de itemkwaliteit aangegeven door de studenten. Met de relevantiescore van de docenten als gouden standaard blijkt de itemvraagtekenscore een betere indicator voor itemrelevantie te zijn dan de item-foutscore bij correctscore. Voor het identificeren van de laagrelevante items schieten echter beide variabelen tekort: met de itemfoutscore lukt dat helemaal niet en met de itemvraagtekenscore maar voor 27% van de laagrelevante items. Dat houdt dus in dat de informatie van een itembeoordeling door docenten slechts zeer ten dele verkregen kan worden met de itemvraagtekenscore bij score met gokcorrectie en in het geheel niet met de itemfoutscore bij correctscore.

Benadelen betere studenten

De in de literatuur gerapporteerde relatieve benadeling van de betere student bij de score met gokcorrectie wordt in ons experiment niet bevestigd.¹¹⁻¹⁵ Het verschil in de G-F-score voor de twee onderzochte scoringsmethodes bleek niet of nauwelijks gecorreleerd met het kennisniveau van de student.

Conclusie

De verkregen resultaten wijzen erop dat er bij de score met gokcorrectie een groter risico bestaat op bias in de kennismeting ten gevolge van verschillen tussen studenten wat betreft het aantal vragen waarop gegokt wordt in plaats van een vraagteken in te vullen. In vervolgonderzoek zal getracht worden meer zicht te krijgen op de verschillen tussen studenten wat betreft gokken in relatie tot toetsresultaten met de twee scoringsmethoden. De verkregen resultaten bevatten aanwijzingen dat de correctscore partiële kennis vollediger meet. Daarentegen blijkt de betrouwbaarheid van de toets hoger te zijn bij de score met gokcorrectie. Het verschil in betrouwbaarheid neemt echter af naarmate de toets meer bestaat uit items die beter aansluiten bij de blokinhoud. De itemvraagtekenscore blijkt weliswaar informatie te bevatten over itemrelevantie, maar vormt geenszins een betrouwbare basis om laagrelevante items te identificeren.

Uit oogpunt van biasbestrijding verdient de correctscore voor bloktoetsen de voorkeur boven de score met gokcorrectie. Wat de betrouwbaarheid betreft zou daarentegen de score met gokcorrectie de voorkeur verdienen. Er is een indicatie dat de betrouwbaarheid bij de correctscore verbeterd kan worden door meer en beter op het blok aansluitende items in de toets op te nemen, maar de vraag is of dat praktisch te verwezenlijken is. Bij de afweging tussen bias en betrouwbaarheid dient in aanmerking te worden genomen dat het eindoor-

deel over een student gebaseerd wordt op resultaten behaald in meerdere toetsen. De betrouwbaarheid van dat oordeel is groter dan die van de afzonderlijke toetsen (random fouten middelen uit). Bij bias gaat het om een systematische fout die, in tegenstelling tot de random fout, niet afneemt bij gebruik van meerdere metingen. Bij het bepalen van een voorkeur voor een van beide scoringsmethoden spelen naast psychometrische ook onderwijskundige aspecten een rol: bijvoorbeeld de doelstelling dat een arts moet leren onderkennen wat hij niet weet en het uitgangspunt dat een student niet gedwongen moet worden om zich te gedragen alsof hij alles weet. Alle aspecten (zowel psychometrische als onderwijskundige) dienen tegen elkaar afgewogen te worden bij het maken van een uiteindelijke keuze voor één van beide scoringsmethoden.

Literatuur

1. McCall WA. A new kind of school examination. *J Educ Res* 1920;1:33-46.
2. Diamond J, Evans W. The correction for guessing. *Rev Educ Res* 1973;43:181-91.
3. Lord FM. Formula scoring and number-right scoring. *J Educ Meas* 1975;12:7-11.
4. Gulliksen H. *Theory of mental tests*. 5e dr. New York: Wiley, 1965.
5. Coombs CH, Milholland JE. The assessment of partial knowledge. *Educ Psychol Meas* 1956;16:13-37.
6. Lord FM. Formula scoring and validity. *Educ Psychol Meas* 1963;23:663-72.
7. Mattson D. The effects of guessing on the standard error of measurement and the reliability of test scores. *Educ and Psychol Meas* 1965;25:727-30.
8. Keislar ER. Test instructions and scoring method in true-false tests. *J Exp Educ* 1953;21:243-49.
9. Swineford F, Miller PM. Effects of directions regarding guessing on item statistics of a multiple-choice vocabulary test. *J Educ Psychol* 1953;44:129-39.
10. Traub RE, Hambleton RK, Singh B. Effects of promised reward and threatened penalty on performance on a multiple-choice vocabulary test. *Educ Psychol Meas* 1969;29:847-61.
11. Votaw DF. The effect of do-not-guess directions upon the validity of true-false or multiple-choice tests. *J Educ Psychol* 1936;27:217-22.

12. Sheriffs AC, Boomer DS. Who's penalized by the penalty for guessing? *J Educ Psychol* 1954;45:81-90.
13. Slakter MJ. The effect of guessing strategy on objective test scores. *J Educ Meas* 1968;5:17-22.
14. Cross LH, Frary RB. An empirical test of Lord's theoretical results regarding formula scoring of multiple-choice tests. *J Educ Meas* 1977;14:313-21.
15. Blis LB. A test of Lord's assumption regarding examinee guessing behavior on multiple choice tests using elementary school students. *J Educ Meas* 1980;17:147-53.
16. Fleming PR. The profitability of "guessing" in MCQ examinations. *Med Educ* 1988;22:81.
17. Vries A. de. Minpunten van de negatieve scoring bij meerkeuzevragen. In: Houtkoop E, Pols J, Pollemans MC, Scherpbier AJJA, Verwijnen GE, redactie. *Gezond Onderwijs 3. Proceedings Gezond Onderwijs Congres 1993*. 's Gravenhage: Haagse Hogeschool, 1994:131-8.
18. Til CT van, Berkel HJM van. De invloed van de vraagtekenoptie op de toetsscore bij kennistoetsen met juist/onjuist vragen. In: Pols J, Cate Th J ten, Houtkoop E, Pollemans MC, Smal JA, redactie. *Gezond Onderwijs 4. Proceedings Gezond Onderwijs Congres 1994*. Houten/Zaventem: Bohn, Stafleu, Van Loghum, 1995:289-94.
19. Rowley GL, Traub RE. Formula-scoring, number-right scoring, and test-taking strategy. *J Educ Meas* 1977;14:15-22.
20. Harden RmcG, Brown RA, Biran LA, et al. Multiple choice questions: to guess or not to guess. *Med Educ* 1976;10:27-32.
21. Fienberg SE. The analysis of cross-classified categorical data. 2nd edition. London: MIT Press, 1980.

DE AUTEURS

Dr. A.M.M. Muijtjens, psychometricus, vakgroep Medische Informatica.

Dr. H. van Mameren, anatoom, vakgroep Anatomie en Embryologie, blokcoördinator.

R.J.I. Hoogenboom, statistisch medewerker, vakgroep Onderwijsontwikkeling en Onderwijsresearch.

Prof. dr. J.L.H. Evers, gynaecoloog, hoogleraar vakgroep Obstetrie en Gynaecologie, blokcoördinator.

Prof. dr. J.P.M. Geraedts, geneticus, hoogleraar vakgroep Moleculaire Celbiologie en Genetica, voorzitter Examencommissie.

Dr. K.M.L. Leunissen, internist/nefroloog, vakgroep Interne Geneeskunde, portefeuillehouder onderwijs van het faculteitsbestuur.

Prof. dr. C.P.M. van der Vleuten, psycholoog, hoogleraar vakgroep Onderwijsontwikkeling en Onderwijsresearch. Allen zijn verbonden aan de Faculteit der Geneeskunde, Universiteit Maastricht.

Correspondentieadres:

Dr. A.M.M. Muijtjens, Vakgroep Medische Informatica, Universiteit Maastricht, Postbus 616, 6200 MD Maastricht.