Tijdschrift voor Medisch Onderwijs, juni 2009 | Vol. 28, nr. 3, p. 124-129

Meerkeuzevragen met drie, vier of vijf alternatieven: wat is beter?

E.C. Paes, O. ten Cate

Samenvatting

Inleiding: *Er lijkt bij veel docenten een gewoonte te bestaan om in studietoetsen vierkeuzevragen te stellen. In de literatuur wordt echter geadviseerd driekeuzevragen te hanteren, omdat het in de praktijk vaak moeilijk is om drie of meer plausibele alternatieven te formuleren. De definitie van een 'implausibele afleider', namelijk een afleider die door 5% of minder van de studenten wordt gekozen, lijkt niet helemaal toereikend. In een retrospectieve studie is onderzocht of met een strengere definitie het advies om te kiezen voor driekeuzevragen nog steeds houdbaar is.*

Methode: *Op grond van de itemanalyses van 935 vragen in 21 meerkeuzetoetsen die in 2006/2007 afgelegd zijn door studenten geneeskunde van het UMC Utrecht, werden alle implausibele afleiders geïdentificeerd. Een implausibele afleider werd gedefinieerd als a) door 5% of minder van het aantal deelnemers gekozen en b) ten minste met een factor 4 minder vaak gekozen dan de meest gekozen afleider.*

Resultaten: *Van de vier- en vijfkeuzevragen bleek 55.9% respectievelijk 66.6% minstens één implausibele afleider te bevatten. Bijna een kwart respectievelijk een vijfde van deze afleiders werd door geen enkele deelnemer gekozen. Van de driekeuzevragen bevatte 18.8% een implausibele afleider en werd 16.7% van deze afleiders door niemand gekozen.*

Discussie en Conclusie: *Om een vraag betrouwbaarder en niet te eenvoudig te maken gaat de voorkeur uit naar meerdere keuzemogelijkheden. Echter, de afleiders moeten dan ongeveer even plausibel zijn, wat blijkens dit onderzoek lang niet altijd lukt. Onze bevindingen ondersteunen dan ook het advies aan driekeuzevragen de voorkeur te geven boven vragen met vier of vijf keuzemogelijkheden. (Paes EC, Cate O ten. Meerkeuzevragen met drie, vier of vijf alternatieven: wat is beter? Tijdschrift voor Medisch Onderwijs 2009;28(3):124-129.)*

Inleiding

Het formuleren van een goede meerkeuzevraag vergt kennis en ervaring. Elke goede meerkeuzevraag bestaat uit een stam (de introducerende vraag of een incomplete stelling), gevolgd door twee of meer keuzemogelijkheden. Deze keuzemogelijkheden bestaan naast het correcte antwoord uit enkele zogenaamde 'afleiders'. Het aantal afleiders hangt af van het type meerkeuzevraag.

Een driekeuzevraag heeft twee afleiders en een vierkeuzevraag drie. De afleiders moeten zo geformuleerd zijn dat het correcte antwoord op grond van *inhoudelijke* kennis herkend wordt en niet omdat de alternatieven heel onwaarschijnlijk zijn.[1] Vierkeuzevragen worden veel gebruikt en vaak aangeraden.[2] Vierkeuzevragen zouden onder meer betrouwbaarder zijn dan twee- of driekeuzevragen, omdat met meer alternatieven de raadkans afneemt.[3-4] Echter, in een review van literatuur over aanbevolen regels voor toetsconstructie concluderen Haladyna et al.[5] dat de kwaliteit van de afleiders be-

langrijker is dan het aantal en dat twee plausibele afleiders in de meeste gevallen voldoende is. In eigen onderzoek hadden zij geconstateerd dat bij slechts 1-8% van de vierkeuzevragen daadwerkelijk sprake was van drie effectieve afleiders en dat gemiddeld de overige vragen slechts één functionele afleider bevatten.[6] Ook Rodriguez[7] en Vyas en Supe[8] concluderen na een review van de bestaande relevante literatuur over toetsvragen, dat vragen met drie keuzemogelijkheden de voorkeur verdienen.

Afleiders moeten plausibel zijn, dat wil zeggen voldoende vaak gekozen worden. Implausibele afleiders hebben een flatterend effect op het eindcijfer. Immers, wanneer één van de afleiders in een vierkeuzevraag door studenten direct als zeer onwaarschijnlijk wordt beschouwd, zal deze al snel weggestreept worden en dus niet als volwaardige afleider fungeren. Er is dan in feite sprake van een driekeuzevraag met een hogere raadkans, namelijk 0.33 ten opzichte van 0.25 bij een vierkeuzevraag. Deze raadkans wordt in een toets normaliter gecompenseerd door aan elk goed antwoord bij een drie- en vierkeuzevraag respectievelijk 0.66 en 0.75 punt toe te kennen. Aan een vierkeuzevraag met een implausibele afleider – welbeschouwd dus een driekeuzevraag – wordt dan ten onrechte 0.75 punt toegekend. Omdat het in de praktijk moeilijk blijkt om drie of meer effectieve afleiders te formuleren wordt in de literatuur geadviseerd vragen met drie keuzemogelijkheden te hanteren. Implausibele afleiders beïnvloeden de kwaliteit van een meerkeuzevraag nadelig.[5-8]

De hypothese dat meerkeuzevragen met meer dan twee afleiders relatief veel implausibele alternatieven bevatten, is in ons onderzoek nogmaals empirisch onderzocht, echter met een strenger criterium. In de literatuur wordt een implausibele afleider gedefinieerd als een afleider die door 5% of minder van de kandidaten gekozen wordt.[6-7 9] Het bezwaar van dit criterium is dat een vierkeuzevraag met een p-waarde van 0.85 of hoger – een 'gemakkelijke vraag' dus – volgens dit criterium altijd minimaal één implausibel alternatief zal opleveren. Immers, de drie afleiders samen worden dan door 15% van de kandidaten gekozen. De verdeling zou dan bijvoorbeeld als volgt kunnen zijn: juiste alternatief gekozen door 85%, afleider a) door 7%, afleider b) door 4% en afleider c) door 4%. Wij hebben er daarom voor gekozen een tweede criterium voor implausibiliteit toe te voegen: een afleider moet aanvullend ten minste met een factor vier minder vaak gekozen worden dan de meest gekozen afleider. Bij een vierkeuzevraag met eenzelfde p-waarde van 0.85 zou dan één afleider implausibel zijn bij de volgende verdeling: juiste alternatief gekozen door 85%, afleider a) door 8%, afleider b) door 6% en afleider c) door 1%. In dat geval beschouwen wij afleider c) wel als implausibel, maar afleiders a) en b) niet. Het gebruik van uitsluitend ons criterium als alternatief voor de 5% regel zou overwogen kunnen worden, maar voldoet weer niet bij vragen met een zeer lage p-waarde. Een 'moeilijke' vraag, met een p-waarde van bijvoorbeeld 0.15, zou drie afleiders kunnen hebben die door respectievelijk 14%, 14% en 57% van de kandidaten gekozen wordt. Het lijkt ons in dit geval niet juist dan twee van de drie afleiders als implausibel te bestempelen, ook al is 14% minder dan een kwart van 57%. Wij hebben daarom gekozen voor de combinatie van beide criteria: een afleider is implausibel indien hij én door 5% of minder van de kandidaten gekozen wordt én ten minste met een factor vier minder vaak wordt gekozen dan de meest gekozen afleider.

Er zijn ook andere redenen om een afleider af te keuren. Een positieve point-bi-

serial correlatie van de afleider met de somscore van de toets, ook wel positieve z-waarde genoemd, geeft aan dat deze afleider overwegend gekozen wordt door de betere kandidaten. Daar het ons gaat om plausibiliteit van afleiders voor de gehele groep kandidaten is dit meer kwalitatieve criterium buiten beschouwing gelaten.

In deze retrospectieve studie is onderzocht of meerkeuzevragen met drie keuzemogelijkheden inderdaad minder implausibele afleiders bevatten en daarom de voorkeur verdienen boven vragen met vier of vijf keuzemogelijkheden.

Methode

Van alle schriftelijke toetsen voor eerste-, tweede- en derdejaarsstudenten geneeskunde van het UMC Utrecht in het studiejaar 2006/2007, werden 21 toetsen geïdentificeerd die zich leenden voor dit onderzoek. Alle toetsen waren afgelegd door het gehele jaarcohort, bestaande uit circa 300 studenten. Herkansingstoetsen werden uitgesloten. In alle toetsen betrof het de meerkeuzevragen met de 'gedwongen raden' vorm: de instructie daarbij is geen vragen onbeantwoord te laten. Eerst werd gekeken naar het voorkomen van vragen met twee, drie, vier en vijf keuzemogelijkheden. Ervan uitgaande dat bij tweekeuze- en juist/onjuist-vragen geen onderscheid gemaakt kan worden tussen een vraag met één implausibel alternatief en eenvoudigweg een 'gemakkelijke vraag', werden deze buiten beschouwing gelaten. Dit gold ook voor de vragen waarin tweekeuzestellingen gecombineerd worden tot een vierkeuzevraag van het type 'A juist + B onjuist; A onjuist + B juist; beide juist; beide onjuist'. Bij dergelijke vragen is namelijk geen sprake van een extra, mogelijk implausibel, alternatief maar van een combinatie van twee alternatieven, in feite dus een gecombineerde tweekeuzevraag. Vervolgens werden de itemparameters van de 21 toetsen onderzocht. De vragen waarvan één of meer afleiders door 5% of minder van het aantal deelnemers werden gekozen en bovendien met een factor > 4 minder vaak werden gekozen dan de meest gekozen afleider, werden gemarkeerd.

Resultaten

Per toets waren er gemiddeld 49.6 meerkeuzevragen en 297 kandidaten. In totaal werden 935 meerkeuzevragen geanalyseerd, waarvan 6.8% met drie, 88.7% met vier en 4.5% met vijf keuzemogelijkheden. Tien toetsen (47.6%) bevatten uitsluitend vierkeuzevragen. De overige 11 toetsen (52.4%) waren opgebouwd uit verschillende soorten meerkeuzevragen. De gemiddelde gecorrigeerde moeilijkheidsgraad (p-waarde) van alle vragen was p=0.61. De p-waarde geeft de moeilijkheidsgraad weer, gecorrigeerd voor de raadkans. Van de drie-, vier- en vijfkeuzevragen bleek respectievelijk 18.8%, 55.9% en 66.7% minimaal één implausibele afleider te bevatten. Dit verschil is sterk significant ($Chi2=3.58$, $df=2$, $p<0.001$). De gemiddelde p-waarde was p=0.64. Respectievelijk 16.7%, 23.8% en 21.4% van deze vragen bevatte een implausibele afleider die zelfs door geen enkele deelnemer gekozen werd. Hiervan was de gemiddelde p-waarde p=0.78. (zie Tabel 1).

Discussie en Conclusie

Vragen met vier keuzemogelijkheden lijken in de praktijk sterk de voorkeur te hebben. In deze steekproef werd een onevenredige verdeling gevonden van het aantal drie-, vier- en vijfkeuzevragen; respectievelijk 6.3%, 81.8% en 4.1%. Tevens bleek dat vier- en vijfkeuzevragen significant meer implausibele afleiders bevatten dan driekeuzevragen. Een afleider werd in deze studie als implausibel gedefinieerd indien deze door 5% of minder van

Tabel 1. *Het voorkomen van vragen met implausibele afleiders in 21 toetsen.*

	Driekeuze-vragen	Vierkeuze-vragen	Vijfkeuze-vragen
Aantal onderzochte vragen	64 (6.8%)	829 (88.7%)	42 (4.5%)
Gemiddelde p-waarde*	0.64	0.61	0.59
Aantal vragen met één of meer implausibele afleiders	12 (18.8%)	463 (55.9%)	28 (66.7%)
Gemiddelde p-waarde*	0.58	0.67	0.66
Aantal vragen waarvan afleider door geen enkele deelnemer gekozen**	2 (16.7%)	110 (23.8%)	6 (21.4%)
Gemiddelde p-waarde*	0.72	0.79	0.83

** p-waarde: moeilijkheidsgraad gecorrigeerd voor de raadkans*
*** percentage van het aantal vragen met één of meer implausibele afleiders*

de kandidaten gekozen werd en met een factor > 4 minder dan de meest gekozen afleider. De keus voor een factor 4 is arbitrair en tot stand gekomen op basis van de interpretatie van het bestudeerde materiaal. Een factor 4.5 of 3.5 zou ook mogelijk zijn geweest en zou wellicht iets andere resultaten hebben gegeven, maar geen wezenlijke andere conclusie.

Bij het formuleren van een kwalitatief goede meerkeuzevraag zijn plausibele afleiders van groot belang. Dat blijkt in de praktijk met name bij vier- en vijfkeuzevragen vaak niet goed te lukken. Docenten veronderstellen vaak dat vierkeuzevragen betrouwbaarder zijn dan driekeuzevragen. Dat lijkt terecht gezien de geringere raadkans. Echter, Rogers en Harley[10] concluderen in een empirische studie dat in het algemeen bij driekeuzevragen juist minder vaak geraden wordt dan bij vierkeuzevragen. Rodriguez[7] beschrijft dat wanneer men vijfkeuzevragen vervangt door driekeuzevragen de moeilijkheidsgraad slechts met p=0.07 daalt en het discriminerend vermogen en de betrouwbaarheid van de vraag onveranderd blijven. Wanneer men driekeuzevragen met vierkeuzevragen vergelijkt, neemt de moeilijkheidsgraad met gemiddeld p=0.04 af en nemen het discriminatoir vermogen en de betrouwbaarheid van de vraag zelfs toe met respectievelijk p=0.03 en p=0.02. In deze studie zien we dat op het totale aantal onderzochte vragen de gemiddelde gecorrigeerde moeilijkheidsgraad (p-waarde) van de vragen met een afleider die door geen enkele deelnemer gekozen werd, hoger ligt, namelijk p=0.78 ten opzichte van p=0.61. De vragen zijn dus relatief gemakkelijker. Rodriguez[7] beschrijft ook dat het minder tijd kost om driekeuzevragen te ontwerpen en in eenzelfde tijdsbestek meer vragen beantwoord kunnen worden. Hierdoor komen meer verschillende onderwerpen aan bod. Tevens zouden meerdere keuzemogelijkheden, ook wanneer deze alle plausibel zijn, het risico met zich meedragen belangrijke informatie in de afleiders bloot te geven over onderwerpen die verder in de toets nog aan bod moeten komen. Soortgelijke bevindingen worden ook door Haladyna en Downing[4] beschreven. De moeilijkheidsgraad, het discriminerende vermogen en de validiteit van een vier- of vijfkeuzevraag worden volgens hen niet substantieel beïnvloed wanneer men minder afleiders gebruikt. Ook stellen ze dat het maken van driekeuzevragen juist

effectiever is, daar er meer vragen ontworpen kunnen worden, zodat de betrouwbaarheid van de toets uiteindelijk positief beïnvloed wordt. In 2008 concluderen ook Vyas en Supe[8] in een literatuurreview omtrent de optimale hoeveelheid alternatieven dat driekeuzevragen de voorkeur verdienen. Ze beschrijven dat het makkelijker en sneller is om twee plausibele afleiders te formuleren dan meer en dat fouten daarmee dus voorkomen worden. In combinatie met het feit dat driekeuzevragen sneller gelezen worden, kunnen er zo meer vragen in een toets verwerkt worden. Hierdoor kan in eenzelfde tijdsbestek een grotere variëteit aan leerstof getoetst worden, wat tot een verhoogde inhoudsvaliditeit leidt. Ook zij stellen dat meerkeuzevragen zelden meer dan drie nuttige opties bevatten.

Budesco en Nevo[11] zijn de enigen die de voorkeur voor driekeuzevragen betwisten. In hun studie in 1985 verwerpen ze de zogenaamde wet van de proportionaliteit van Grier[12] uit 1975, die stelt dat de totale toetstijd evenredig is met de hoeveelheid alternatieven per vraag. Ze concluderen dat er juist sprake is van een negatieve relatie en dat driekeuzevragen dus over het algemeen insufficiënt zouden zijn. Ze geven echter geen advies voor de optimale hoeveelheid alternatieven. Daarentegen bevestigen Bruno en Dirkzwager[13] in 1995, gebaseerd op de wet van de proportionaliteit, dat driekeuzevragen optimaal zijn.

Andere studies die alleen vier- en vijfkeuzevragen met elkaar vergeleken hebben, kwamen tot de conclusie dat vierkeuzevragen de voorkeur verdienen.[14-15]

De resultaten van onze studie bevestigen de eerdere bevindingen van Rodriguez[7], Haladyna en Downing[5-6], ook indien men een strikter criterium van implausibiliteit hanteert. Het blijkt dat toetsen met vier- en vijfkeuzevragen, gebaseerd op beide criteria, inderdaad een groter percentage vragen met één of meerdere implausibele afleiders bevatten dan toetsen met driekeuzevragen.

Implausibele afleiders zouden bewust vermeden moeten worden. Men kan beter niet koste wat kost een extra alternatief bedenken met de gedachte dat dit de betrouwbaarheid en kwaliteit van de vraag verhoogt. Dit is tijdrovend en brengt vaak juist een tegenovergesteld effect teweeg. In algemene zin gaat de kwaliteit van de afleiders boven de kwantiteit. Ten slotte zou de auteur ter lering de vragen na het afnemen van de toets op onvoldoende functionerende afleiders kunnen controleren.

Het resultaat van dit onderzoek, in samenhang met de kennis uit de literatuur, ondersteunt het advies aan vragen met drie keuzemogelijkheden de voorkeur te geven boven vragen met vier of vijf keuzemogelijkheden.

Literatuur

1. Kehoe J. Writing multiple-choice test items. Practical Assessment, Research and Evaluation 1995;4(9). http://pareonline.net/getvn.asp?v=4&n=9.[Bezocht op 14 september 2008].
2. De Groot AD, Van Naerssen RF. Studietoetsen construeren, afnemen, analyseren. Den Haag: Mouton Publishers;1969.
3. Haladyna TM, Downing SM. A Taxonomy of Multiple-Choice Item-Writing Rules. Appl Meas Educ 1989a;2(1):37-50.
4. Haladyna TM, Downing SM. A Taxonomy of Multiple-Choice Item-Writing Rules. Appl Meas Educ 1989b;2(1):51-78.
5. Haladyna TM, Downing SM, Rodriguez MC. A review of multiple-choice item-writing guidelines for classroom assessment. Appl Meas Educ 2002;15:309-34.
6. Haladyna TM, Downing SM. How many options is enough for a multiple choice item? Educ Psychol Meas 1993;53:999-1010.
7. Rodriguez MC. Three options are optimal for Multiple-Choice Items: A Meta-Analysis of 80 Years of Research. Educ Meas: Issues and Practice 2005;24(2):3-13.
8. Vyas R, Supe A. Multiple Choice questions: A literature review on the optimal number of options. The National Medical Journal of India 2008; 3(21):21-4.

9. Wakefield JA. Does the fifth choice strengthen a test item? Public Personnel Review 1958;19:44-8.
10. Rogers WT, Harley D. An empirical comparison of three-and four-choice items and tests: Susceptibility to testwiseness and internal consistency reliability. Educ Psychol Meas 1999;59:234-47.
11. Budesco DV, Nevo B. Optimal number of options: An investigation of the assumption of proportionality. J Educ Meas 1985;22:899-13.
12. Grier JB. The number of alternatives for optimum test reliability. J Educ Meas 1975;12:109-13.
13. Bruno JE, Dirkzwager A. Determining the optimal number of alternatives to a multiple-choice test item: An information theoretic perspective. Educ Psychol Meas 1995;55:959-66.
14. Hodson D. Some effects of changes in question structure and sequence on performance in a multiple choice chemistry test. Res Sci Technol Educ 1984;2:177-85.
15. Ramos RA, Stern J. Item behaviour associated with changes in the number of alternative in multiple choice items. J Educ Meas 1973;10:305-10.

De auteurs:
Emma Paes is medisch student in het UMC Utrecht en was ten tijde van dit project onderwijsstagiaire.
Prof. dr. Th.J. (Olle) ten Cate is hoogleraar medische onderwijskunde, directeur van het Expertisecentrum voor Onderwijs en Opleiding van het UMC Utrecht en coördinator van de onderwijsstages.

Correspondentieadres:
Emma Paes, Willemstraat 3, 3511 RJ Utrecht.
Tel.: 06-51567362; e-mail: emmapaes@gmail.com

Belangenconflict: geen gemeld
Financiële ondersteuning: geen gemeld

Summary

Aim: *Many teachers seem to prefer the use of four-option items in closed format tests. In the literature three-option items are recommended because it is considered difficult to formulate more than three plausible distractors. However, the common definition of an implausible distractor, i.e. one that is selected by 5% or less of candidates, may not be fully adequate. We examined the validity of this recommendation in a retrospective study, using a stricter definition of an implausible distractor.*
Method: *The results of 21 multiple choice tests (935 items) of medical students at UMC Utrecht in 2006-2007 were examined to identify all the items with one or more implausible distractors. An aimplausible distractor was defined as a distractor selected by 5% or less of the candidates and with a frequency of selection that was lower by a factor 4 than that of the most frequently selected distractor.*
Results: *Of all four option items 55.9% contained at least one implausible distractor of which 25% was not selected by any of the participants. For five-option items the corresponding percentages were 66.6% and 20% and for three items they were 18.8% and 16.7%.*
Discussion and Conclusion: *Items with many distractors are generally considered preferable because they are assumed to enhance test reliability and ensure that test items are not too easy. All distractors must be equally plausible but this study showed that this was not the case. Our study supports the recommendation to use three-option items rather than four- or five-option items. (Paes EC, Ten Cate O. Multiple choice questions with three, four of five answer options: which is preferable? Dutch Journal of Medical Education 2009;28(3):124-129.)*