

Instrument voor collegiale screening van kennistoetsen

□ Karin Gerritsen-van Leeuwenkamp, Jan Adema en Gerard Straetmans

Toets- en itemanalyses laten zien dat de kwaliteit van kennistoetsen in het hoger onderwijs niet altijd toereikend is. Een manier om de kwaliteit van kennistoetsen vóór afname te evalueren is collegiale screening. Als collegiale screening vanuit het toetsbeleid van een opleiding systematisch wordt ingezet, is het verstandig om de criteria waarop gescreend wordt te standaardiseren. Op basis daarvan kunnen de resultaten gebruikt worden om generaliserende uitspraken te doen over de kwaliteit van toetsen. Dit artikel richt zich op de ontwikkeling van een instrument waarmee zowel de toetstechnische als de vakinhoudelijke kwaliteit van kennistoetsen kan worden beoordeeld.

Kwaliteit van kennistoetsen

Kennistoetsen in het hoger onderwijs voldoen niet altijd aan de kwaliteitseisen. Zo bleek bijvoorbeeld uit psychometrische analyses van ruim duizend meerkeuze-toetsen van de Vrije Universiteit Amsterdam dat de betrouwbaarheid van toetsen vaak 'onder de maat was' (Van Os & Van Beek, 2012). Als de toetskwaliteit tekort schiet kan dit leiden tot het nemen van onterechte zak- en slaagbeslissingen, vooral over studenten met scores dichtbij de cesuur (Sanders, 2011). Omdat deze kwaliteitsproblemen uiteindelijk invloed kunnen heb-

ben op het gerealiseerde eindniveau en de waarde van diploma's, dient de kwaliteit van toetsen vastgesteld en waar nodig verbeterd te worden. Eén van de manieren om de kwaliteit van kennistoetsen vóór afname te bepalen is collegiale screening.

Collegiale screening

Collegiale screening van toetsen is een vorm van formatieve evaluatie, uitgevoerd door een vakinhoudelijk deskundige collega van de toetsconstructeur, met als doel de kwaliteit van een toets te controleren en/of te verbeteren voordat die door studenten gemaakt gaat worden. Collegiale screening kan ook, aansluitend bij de aanbeveling van de commissie Externe validering examenkwaliteit hoger beroepsonderwijs, ingezet worden als een vorm van externe validering van examens (HBO-raad, 2012). Een bijkomend voordeel van collegiale screening is dat die kan bijdragen aan de professionalisering van docenten op het terrein van toetsing en examinering. Professionalisering van docenten op het gebied van toetsing en examinering is een andere aanbeveling die de commissie Externe validering examenkwaliteit hoger beroepsonderwijs heeft voorgesteld om de kwaliteit van examens in het hbo te verbeteren (HBO-raad, 2012). Deze aanbeveling is overgenomen en verder uitgewerkt door de Expertgroep BKE/SKE en heeft geleid tot de eis dat elke docent die eindverantwoordelijk is voor één of meer onderdelen van het toetsproces (ook wel examiner genoemd), moet beschikken over een basiskwalificatie examinering (BKE) (Expertgroep BKE/SKE, 2013).

Collegiale screening vindt al plaats in het hoger beroepsonderwijs (hbo), maar het ontbreekt docenten en de examencommissie vaak aan gestandaardiseerde instrumenten om de kwaliteit van toetsen te waarborgen en te verbeteren (Inspectie van het Onderwijs, 2015). Als collegiale screening vanuit het toetsbeleid van een opleiding systematisch wordt ingezet voor zowel externe validerings- als professionaliseringsdoelen, is het verstandig om de criteria waarop

Gestandaardiseerd instrument om zowel de toetstechnische als de vakinhoudelijke kwaliteit van kennistoetsen te beoordelen



gescreend wordt te standaardiseren, zodat alle toetsen op gelijke criteria worden beoordeeld. Alleen dan zijn de resultaten te gebruiken om generaliserende uitspraken te doen over de kwaliteit van de toetsen en de toetsdeskundigheid van de examinatoren en kan besloten worden om het toetsbeleid en/of de professionaliseringsactiviteiten van de opleiding bij te sturen. Deze afweging was bepalend voor het ontwerp van een instrument voor collegiale screening van kennis-toetsen.

Ontwikkeling en evaluatie van het instrument

Eerst is door middel van een literatuurstudie nagegaan wat volgens de (inter)nationale literatuur wordt verstaan onder de kwaliteit van kennistoetsen tijdens de constructiefase van het toetsproces. Zes bronnen zijn opgenomen in de literatuurstudie (American Educational Research Association, 2004; Brookhart & Nitko, 2007; Evers, Lucassen, Meijer, & Sijtsma, 2010; Hale & Astolfi, 2011; International test commissie, 2001; Van Berkel, Bax, & Joosten-ten Brinke, 2014).

Ontwikkeling en evaluatie prototype 1

Op basis van de resultaten uit de literatuurstudie zijn criteria en zogenoemde hulpvragen opgesteld. Daarbij golden twee belangrijke voorwaarden. In de eerste plaats of het criterium tijdens de constructiefase van de kennistoets (dus vóór afname) te beoordelen is en in de tweede plaats of het criterium observeerbaar is voor de screener. De criteria beschrijven de kwaliteit

van de producten en processen uit de volgende fasen van de BKE-toetscyclus: 1) basisontwerp; 2) construeren toetsmatrijs; en 3) construeren toets/normeren (Expertgroep BKE/SKE, 2013). De hulpvragen ondersteunen de screener bij het vormen van een oordeel over het betreffende criterium. Zie Figuur 1 voor een voorbeeld.

De geformuleerde criteria en hulpvragen werden via het digitale enquêteprogramma Parantion voorgelegd aan hogeschooldocenten van zes hogescholen en aan toetsexperts van vier hogescholen en drie bedrijven, ter beoordeling van de relevantie, volledigheid, juistheid en duidelijkheid. De respondenten zijn persoonlijk benaderd, waarna de vragenlijst is verstuurd. De respons van de docenten was 55% (n=12), van de toetsexperts 90% (n=18).

Vrijwel alle criteria en hulpvragen werden zowel door de docenten als de toetsexperts als belangrijk of cruciaal beoordeeld en deze werden geselecteerd voor opname in prototype 2. Er werden opmerkingen gemaakt over de betekenis van of het correct gebruik van toetstechnische termen. Daarnaast werd er feedback gegeven op de bruikbaarheid van het instrument. Feedback leidde tot het verwijderen, herordenen en herformuleren van criteria en hulpvragen binnen de twee hiervoor beschreven uitgangspunten. Onduidelijkheden en onjuistheden in teksten werden gecorrigeerd en toelichtingen in 'pop-ups' geplaatst bij toetstechnische begrippen (zie Figuur 2 voor een voorbeeld).

Criterium:

De toetsmatrijs is correct opgesteld.

Hulpvragen:

- Is het toetsdoel weergegeven in leerdoelen of leerstofonderwerpen?
- Is bij elk leerdoel/ leerstofonderwerp het cognitieve beheersingsniveau aangegeven?
- Geeft de matrijs per leerdoel/ leerstofonderwerp het percentage of aantal opgaven?
- Komen de mogelijke verschillen in vakinhoudelijk belang of complexiteit tussen leerdoelen/ leerstofonderwerpen tot uitdrukking in het aantal opgaven of scorepunten?

Figuur 1. formulering van een criterium en vier bijbehorende hulpvragen



Collegiale screening als manier om de kwaliteit van kennistoetsen vóór afname te bepalen

De geformuleerde criteria en hulpvragen werden via het digitale enquêteprogramma Parantion voorgelegd aan hogeschooldocenten van zes hogescholen en aan toetsexperts van vier hogescholen en drie bedrijven, ter beoordeling van de relevantie, volledigheid, juistheid en duidelijkheid. De respondenten zijn persoonlijk benaderd, waarna de vragenlijst is verstuurd. De respons van de docenten was 55% (n=12), van de toetsexperts 90% (n=18).

Vrijwel alle criteria en hulpvragen werden zowel door de docenten als de toetsexperts als belangrijk of cruciaal beoordeeld en deze werden geselecteerd voor opname in prototype 2. Er werden opmerkin-

gen gemaakt over de betekenis van of het correct gebruik van toetstechnische termen. Daarnaast werd er feedback gegeven op de bruikbaarheid van het instrument. Feedback leidde tot het verwijderen, herordenen en herformuleren van criteria en hulpvragen binnen de twee hiervoor beschreven uitgangspunten. Onduidelijkheden en onjuistheden in teksten werden gecorrigeerd en toelichtingen in 'pop-ups' geplaatst bij toetstechnische begrippen (zie Figuur 2 voor een voorbeeld).

Ontwikkeling en evaluatie prototype 2

Prototype 2 werd beschikbaar gesteld in een digitaal

40%

Instrument kwaliteit van kennistoetsen (5/11)

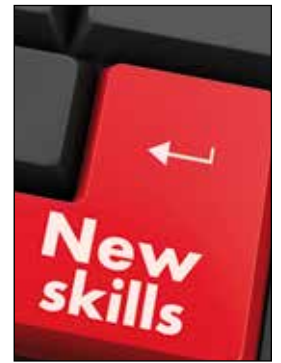
Hieronder krijgt u de vraagtypen (bv. open, meerkeuze, stelling, overig gesloten) te zien, waarvan u heeft aangegeven dat hiervan tenminste 10 (deel) opgaven zijn opgenomen in de toets.

Beoordeel steeds de eerste tien opgaven van elk aangegeven vraagtype. Geef **per hulpvraag** aan hoeveel van de tien vragen voldoen. Beoordeel **éerst** de **hulpvragen** (3.x) en kom aan de hand daarvan tot een oordeel over **criterium 3**. Dus vat jouw oordeel over de hulpvragen samen in een score op het criterium.

	Goed (9-10 vragen)	Voldoende (7-8 vragen)	Onvoldoende (0-6 vragen)	Ik weet het niet
3. De (deel)opgaven bevatten een duidelijke opdracht.	☒	☐	☐	☐
Open opgaven				
3.1 Is de opdracht concreet, zodat duidelijk is waaruit het antwoord moet bestaan?	☒	☐	☐	☐
3.2 Is duidelijk welke lengte het antwoord moet hebben?	☒	☐	☐	☐
3.3 Zijn de opgaven en eventuele deelopgaven duidelijk van elkaar te onderscheiden?	☒	☐	☐	☐
3.4 Is de opgave positief gesteld?	☒	☐	☐	☐
Meerkeuze opgaven:				
3.5 Is de opgave positief gesteld?	☒	☐	☐	☐
3.6 Bevat de opgave één specifiek probleem?	☒	☐	☐	☐

Dit gaat over de aanbeveling om negatieve formuleringen (bijvoorbeeld: nooit, geen, niemand) te vermijden. Negatief gestelde formuleringen verhogen de moeilijkheidsgraad van de opgaven ten onrechte.

Figuur 2. Screenshot van 'Prototype 2'. Een 'Pop-up' is geplaatst bij een toetstechnisch begrip



enquêteprogramma en bevatte 14 geselecteerde criteria en 54 hulpvragen (zie Figuur 2 en 3). Twee schalen werden in het instrument gebruikt, een checklist met twee schaalpunten (ja/nee) en een beoordelingsschaal met drie schaalpunten (onvoldoende, voldoende en goed). Beide schalen boden screeners ook de mogelijkheid om geen beoordeling te geven, middels het aanvinken van de button 'ik weet het niet'. In prototype 2 was een routing van de vragen actief, wat inhield dat op basis van gegevens over de te screenen toets (bijvoorbeeld open of gesloten vragen) de screener een selectie van relevante criteria en hulpvragen kreeg voorgelegd. Om de tijdsduur van het invullen van prototype 2 te beperken werd bij een aantal criteria ervoor gekozen om een steekproef van de opgaven uit de toets te laten beoordelen. Daarnaast bevatte dit prototype ook vijf open vragen om informatie te verkrijgen over de doelmatigheid, doeltreffendheid, bruikbaarheid, volledigheid en taalgebruik van het instrument. In totaal hebben 24 tweetallen van docenten (van

zeven verschillende opleidingen van drie instellingen) prototype 2 gebruikt om een kennistoets van een collega te screenen. De evaluatie van prototype 2 begon met een informatiebijeenkomst voor de deelnemende docenten. Daarna ontvingen de deelnemende docenten via de e-mail een uitnodiging vanuit het digitale enquêteprogramma Parantion, met na twee weken een herinnering. Kort daarna ontvingen alle deelnemende docenten een rapportage over de door hen gescreende toets. Deze rapportage bestond uit, de ingevulde gegevens van de te screenen toets, de samenvatting van de resultaten op criteriumniveau en de resultaten van alle beoordeelde criteria en hulpvragen.

De 48 screeningen zijn uitgevoerd door 25 docenten (vijf docenten hebben meer dan één kennistoets beoordeeld), waarvan zes met een basiskwalificatie examinering (BKE), vijf met een seniorkwalificatie examinering (SKE) en veertien met een onbekende toetsdeskundigheid. In 20 van de 48 screeningen

Prototype 2

De criteria

1. De toetsmatrijs is correct opgesteld.
2. De opgaven zijn uitwerkingen van de leerdoelen en het cognitieve beheersingsniveau zoals vermeld in de toetsmatrijs.
3. De (deel)opgaven bevatten een duidelijke opdracht.
4. De inleidende teksten, casussen of visueel materiaal zijn van goede kwaliteit.
5. Het beoordelingsmodel maakt een objectieve verwerking van de antwoorden tot scores mogelijk.
6. De alternatieven van de meerkeuze opgaven zijn effectief.
7. Het antwoord van de stelling opgaven is correct.
8. De alternatieven van de overige gesloten vragen zijn plausibel.
9. Het taalgebruik van de toets is correct en uniform.
10. De toets is zorgvuldig vormgegeven en samengesteld.
11. De scoring en de cesuur zijn methodisch verantwoord en helder.
12. De herkansing is een toets met andere opgaven maar van gelijkwaardig niveau.
13. De toets wordt veilig verspreid en bewaard.
14. Met de toets wordt vastgesteld of de student het toetsdoel bereikt heeft, zoals omschreven in het examenprogramma of de modulebeschrijving, of scorepunten?

Figuur 3: Criteria van prototype 2



Het instrument wordt vormgegeven in een digitale applicatie

was de docent vakdeskundig op het terrein van het te screenen onderwerp. De respondenten besteedden gemiddeld 38 minuten aan het invullen van het instrument. De doelmatigheid van het instrument beoordeelde het merendeel van de docenten als minimaal voldoende. Gegeven opmerkingen zijn: "Het geeft structuur in de controle, maar de interpretatie blijft uiteraard persoonsafhankelijk."; "Ik ga in elk geval weer beter en anders over vragen nadenken." Alle respondenten beoordeelden de bruikbaarheid van het instrument minimaal als voldoende. Ook werd feedback gegeven op de gebruikte schalen, twee docenten wilden de optie 'niet van toepassing' toevoegen. Bijna alle docenten gaven aan het instrument volledig te vinden en hadden geen aanvullingen. Sommige vonden ook dat alle opgaven van een toets beoordeeld zouden moeten worden in plaats van een steekproef. De formulering van de criteria werd beoordeeld als goed. De 'pop-up teksten' die de betekenis van bepaalde criteria moeten verduidelijken, vonden de respondenten zinvol.

Een belangrijk kwaliteitsaspect van het instrument is de betrouwbaarheid van de beoordelingen. Doordat

bij de evaluatie van prototype 2 met tweetallen is gewerkt was het mogelijk om de interbeoordelaarsbetrouwbaarheid te bepalen op grond van de overeenstemming in toegekende scores van de verschillende tweetallen. De resultaten zijn weergegeven in Tabel 1. Het percentage overeenstemming dat tweetallen gemiddeld bereikten was 92 voor de checklist en 75 voor de beoordelingsschalen. In de literatuur is consensus dat onafhankelijke beoordelaars op basis van puur toeval tot dezelfde beoordelingen kunnen komen. Een veel gebruikte overeenstemmingsmaat, die corrigeert voor deze kansovereenstemming is Cohen's Kappa (Cohen, 1960). Aangezien in deze beoordelaarsovereenstemmingsstudie de beoordelaars geen kennis hadden van de beoordelingsresultaten van andere beoordelaars, is een variant van Cohen's Kappa gebruikt, beschreven door Brennan and Prediger (1981).

In Tabel 1 zijn zowel de percentages beoordelaarsovereenstemming als de Kappa-waarden weergegeven. De waarden voor de checklist en de beoordelingsschaal zijn apart weergegeven. Tabel 1 laat zien dat de checklist bij 19 van de 24 tweetallen een goede

Tabel 1. Frequentieverdeling van de bereikte overeenstemming op de schalen van het instrument ($n = 24$ tweetallen van screeners)

Percentage overeenstemming			Kappa (Brennan & Prediger, 1981)		
%	Check list n	beoordelings-schaal n	Kappa	Check list n	beoordelings-schaal n
≤ 20		1	$\leq 0,20$		2
21-40			0,21 – 0,40	2	4
41-60		5	0,41 – 0,60	3	5
61-80	5	6	0,61 – 0,80	3	5
> 80	19	12	$> 0,80$	16	8

Noot: n = aantal tweetallen van screeners

Toelichting Kappa waarden:

$< 0,20$ slechte overeenstemming (poor/ slight agreement)

0,21-0,40 matige overeenstemming (fair agreement)

0,41-0,60 redelijke overeenstemming (moderate agreement)

0,61-0,80 goede overeenstemming (substantial agreement)

0,81-1,00 uitstekende overeenstemming

(almost perfect agreement)

In gewijzigde vorm overgenomen uit Landis and Koch (1977)

Examencommissies waarborgen de kwaliteit van toetsen



tot uitstekende overeenstemming laat zien. Voor de beoordelingsschaal bij 13 van de 24 tweetallen. De gevonden beoordelaarsovereenstemming is voldoende om een breder opgezette pilotstudie in het najaar van 2015 met vrouwen tegemoet te zien.

Tot slot

Wettelijk is vastgelegd dat leden van een examencommissie de kwaliteit van toetsen moeten waarborgen. Collegiale screening in de constructiefase kan hiervoor een effectieve methode zijn. Voorwaarde is dat de gebruikte criteria relevant zijn en de beoordelingsresultaten betrouwbaar. Dit vraagt om de inzet van een gestandaardiseerd instrument, waarvan de kwaliteit in de toetspraktijk van hbo-opleidingen is gebleken. Op basis van de resultaten van de evaluatie van prototype 2 wordt in de zomermaanden gestart met de constructie van een prototype 3 in de vorm van een digitale applicatie die gebruikt kan worden bij de beoordeling van de kwaliteit van kennistoetsen in de constructiefase.

Literatuurlijst

- American Educational Research Association. (2004). *Standards for educational and psychological testing*. Washington: American Educational Research Association.
- Brennan, R. L., & Prediger, D. J. (1981). Coefficient Kappa: some uses, misuses, and alternatives. *Educational and psychological measurement*, 41(3), 687-699. doi: 10.1177/001316448104100307
- Brookhart, S. M., & Nitko, A. J. (2007). *Assessment and Grading in Classrooms*. Upper Saddle River: Pearson Education.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1), 37-46. doi: 10.1177/001316446002000104
- Evers, A., Lucassen, W., Meijer, R., & Sijsma, K. (2010). COTAN *Beoordelingssysteem voor de kwaliteit van tests*. Amsterdam: Nederlands Instituut van Psychologen.
- Expertgroep BKE/SKE. (2013). *Verantwoord toetsen en beslissen in het hoger onderwijs, een voorstel voor een programma van eisen voor een basis- en seniorkwalificatie examinering (BKE/SKE)*. Den Haag: Vereniging Hogescholen.
- Hale, C. D., & Astolfi, D. (2011). Chapter 5 constructing tests and performance assessments *Measuring learning and performance: a primer*. St. Leo: Saint Leo University.
- HBO-raad. (2012). *Vreemde ogen dwingen: eindrapport commissie externe validering examenkwaliteit hoger beroepsonderwijs*. Den Haag: HBO-raad.
- Inspectie van het Onderwijs. (2015). *Verdere versterking onderzoek naar het functioneren van examencommissies in het hoger onderwijs*. Utrecht: Inspectie van het Onderwijs.
- International test commissie. (2001). Internationale richtlijnen voor het gebruik van tests, from <http://www.vub.ac.be/PE/Documenten/Internationale%20richtlijnen%20voor%20het%20gebruik%20van%20tests.pdf>
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159-174. doi: 10.2307/2529310
- Sanders, P. (2011). De betrouwbaarheid van toetsscores. In P. Sanders (Ed.), *Toetsen op school* (pp. 46-64). Arnhem: Cito.
- Van Berkel, H., Bax, A., & Joosten-ten Brinke, D. (2014). *Toetsen in het hoger onderwijs*. Houten: Bohn Stafleu van Loghum.
- Van Os, W., & Van Beek, M. (2012). Toetskwaliteit in objectieve zin en volgens het oordeel van studenten: een casestudy. *Tijdschrift voor het hoger onderwijs*, 30(4), 259-269.

□ Karin J. Gerritsen-van Leeuwenkamp MSc is kenniskringlid bij het Kenniscentrum Onderwijsinnovatie van Saxion en onderwijskundig adviseur bij de Academie gezondheidszorg van Saxion. Daarnaast is zij buitenpromovendus bij het Welten-instituut van de Open Universiteit.

E-mail: k.j.vanleeuwenkamp@saxion.nl

□ Jan Adema is kenniskringlid bij het Kenniscentrum Onderwijsinnovatie van Saxion en toetsdeskundig adviseur bij Cito, Arnhem. E-mail: j.j.adema@saxion.nl

□ Dr. Gerard J.J.M. Straetmans is lector Assessment bij het Kenniscentrum Onderwijsinnovatie van Saxion en medewerker van Cito, Arnhem. E-Mail: g.j.j.m.straetmans@saxion.nl