Toetskwaliteit in objectieve zin en volgens het oordeel van studenten: een casestudy*

*Willem van Os & Marlies van Beek***

Dit artikel bevat een overzicht van psychometrische gegevens over de kwaliteit van ruim duizend meerkeuzetentamens over een periode van zes studiejaren. Die kwaliteit is niet al te hoog: de gemiddelde betrouwbaarheid is kleiner dan .70. Deze vaststelling komt overeen met de mening van studenten. Hun oordeel over de kwaliteit van het tentamen dat ze net hebben afgelegd is aanzienlijk lager dan hun waardering voor de docent of de cursus. In 193 gevallen was het mogelijk om tentamens waarvan psychometrische gegevens bekend zijn, te koppelen aan de evaluaties door studenten van deze tentamens. Deze casestudy kan worden gezien als een onderzoek naar de validiteit van het studentenoordeel over die toetskwaliteit. In verschillende opzichten blijkt het studentenoordeel indicatief te zijn voor de toetskwaliteit in psychometrische zin. Een eerste conclusie is dat de kwaliteit van de toetsing sterk voor verbetering vatbaar is. Daarnaast wordt aanbevolen meer aandacht te besteden aan de psychometrische analyse van tentamens met open vragen.

Inleiding

Het belang van de kwaliteit van de toetsing en, in het verlengde daarvan, het bewaken en verbeteren van die kwaliteit kan nauwelijks worden overschat. Het beoordelingskader van de Nederlands Vlaamse Accreditatie Organisatie (NVAO) formuleert als derde standaard dat de opleiding beschikt over een adequaat systeem van toetsing, met als criteria dat de toetsen valide, betrouwbaar en voor de studenten inzichtelijk zijn (november 2011). Voor wat betreft de kwaliteitsbewaking stelt de Wet op het Hoger Onderwijs d.d. 01-09-2010 in artikel 7.12b, lid 1 als eerste taak van de examencommissie 'het borgen van de kwaliteit van de tentamens en examens'. De betekenis van de toetsing voor de studievoortgang van de individuele student spreekt daarnaast voor zich.
Deze bijdrage bevat een overzicht van de stand van zaken van de kwaliteit van de toetsing, zowel in psychometrische zin als vanuit het standpunt van studenten. Dit overzicht blijft beperkt tot de Vrije Universiteit Amsterdam (VU) en heeft uit dien hoofde slechts de status van een casestudy. Wel is het denkbaar dat de hier gerapporteerde uitkomsten ook elders van toepassing zijn. Zekerheid daarover is er echter niet, in de eerste plaats omdat vergelijkbare overzichten van andere

* Dr. W. van Os is op 14 december 2012 overleden.

** Dr. W. van Os was verbonden aan de Faculteit der Psychologie & Pedagogiek van de Vrije Universiteit Amsterdam. Drs. M. van Beek (m.e.van.beek@vu.nl) is werkzaam bij de afdeling Onderwijs & Kwaliteitszorg; Onderwijsevaluatie, Toetskwaliteit en Institutional Research van de Vrije Universiteit Amsterdam

instellingen niet bekend zijn en de betreffende gegevens wellicht ook niet systematisch worden bijgehouden. Bovendien maken hogeronderwijsinstellingen zelden of nooit gebruik van evaluatievragenlijsten in gestandaardiseerde vorm, zodat bepaalde gegevens niet beschikbaar zijn. Dit geldt bijvoorbeeld voor de gegevens met betrekking tot de tweede onderzoeksvraag, namelijk het oordeel van de studenten over de door hen afgelegde tentamens.
Een tweede beperking is dat dit artikel zich richt op de kwaliteit van meerkeuzetentamens. Hoewel deze vooral in de bachelorfase in grote faculteiten in toenemende mate worden afgenomen, mag niet uit het oog worden verloren dat een groot deel van de toetsing, zeker in de masterfase, plaatsvindt met behulp van open vragen, opdrachten en werkstukken, en vermoedelijk zelfs niet zelden ook via mondelinge tentamens (zie hiervoor ook 'Conclusies en discussie').
De kernvraag van dit artikel (zie onderzoeksvraag 3) is de mate waarin de psychometrische kwaliteit van tentamens en het oordeel van studenten over de kwaliteit van deze tentamens overeenstemmen. In essentie is dit een vraag naar de mate waarin vragenlijstuitkomsten valide zijn. Eerder onderzoek over dit onderwerp is ons niet bekend, in tegenstelling tot onderzoek over de validiteit van het studentenoordeel over de docent, dat in de afgelopen decennia uitvoerig is onderzocht (zie bijvoorbeeld Abrami, d'Apollonia & Cohen, 1990; Marsh, 1987; Patrick, 2011).
De Tentamenservice van de VU biedt voor alle faculteiten de mogelijkheid afgenomen meerkeuzetentamens binnen afzienbare tijd te laten verwerken en analyseren.
Hoewel hierover geen volstrekte zekerheid bestaat, is het wel waarschijnlijk dat de Tentamenservice VU vrijwel alle aan de VU afgenomen meerkeuzetentamens verwerkt en analyseert. Deze Tentamenservice maakte tijdens de onderzoeksperiode deel uit van het Onderwijscentrum VU, net als de dienstverlening op het gebied van onderwijsevaluaties. Digitale toetsen en een onbekend aantal tentamens met een dermate gering aantal studenten dat de docent ze zelf met de hand nakijkt, vallen buiten dit overzicht.

Onderzoeksvragen

Zowel de resultaten van cursusevaluaties als de psychometrische gegevens van afgenomen meerkeuzetentamens worden systematisch bijgehouden en de eerder genoemde kernvraag is dan ook opgedeeld in de volgende onderzoeksvragen:

1 Wat is de gemiddelde psychometrische kwaliteit van meerkeuzetentamens die aan de Vrije Universiteit worden verwerkt en geanalyseerd?
2 Wat is het oordeel van studenten over de kwaliteit van de door hen afgelegde tentamens (zowel meerkeuze- als open vragen)?
3 In hoeverre en op welke punten komt die kwaliteit overeen met de door studenten gepercipieerde kwaliteit?

Onderzoeksmethoden

Onderzoeksvraag 1
Over een periode van zes opeenvolgende jaren is door de Tentamenservice van de faculteiten Aard- & levenswetenschappen, Bewegingswetenschappen, Economische wetenschappen & bedrijfskunde, Exacte wetenschappen, Letteren, Psychologie & Pedagogiek, Sociale wetenschappen, Tandheelkunde (Acta), Rechten, Geneeskunde (VUmc) en Wijsbegeerte een overzicht gemaakt van de verwerkte meerkeuzetentamens.
Van elk tentamen bevat het overzicht de betrouwbaarheid (*KR-20*), de gemiddelde *p*-waarde, de r_{ir}-waarden en het aantal vragen. Daarnaast wordt per tentamen het aantal vragen weergegeven waarop door de Tentamenservice commentaar is gegeven, bijvoorbeeld wegens een lage *p*-waarde en/of lage of negatieve r_{ir}-waarde. Bij de r_{ir}-waarden gaat het om waarden kleiner dan circa 0.15.
De analyse blijft beperkt tot de eerste tentamenkans (dus geen herkansingen) en tentamens met een minimaal aantal deelnemers van 25.

Onderzoeksvraag 2
Na afloop van het tentamen (alleen bij de eerste kans) vullen studenten doorgaans een evaluatievragenlijst in, waarin naast het oordeel over de docent en de cursus gevraagd wordt naar hun mening over het tentamen zelf. In alle gevallen wordt gebruik gemaakt van een vijfpunts Likertschaal. Het gaat bij de vragen over het tentamen om de representativiteit ervan, de mate waarin de studenten via oefenvragen en/of proeftentamens op de hoogte waren van hoe het tentamen er zou uitzien (de transparantie) en het totaaloordeel over het tentamen (zie voor de letterlijke tekst Appendix A). In één faculteit (Bewegingswetenschappen) wordt niet gevraagd naar de transparantie. Studenten moeten ten slotte aangeven hoe ze hun eigen slagingsverwachting inschatten: gezakt, twijfelachtig of geslaagd. Per cursus gaat het om drie percentages (samen 100%), die zijn omgezet naar een score op een driepuntsschaal (met 2 als minimum, 4 als maximum).
In de jaren tachtig is begonnen met deze manier van onderwijsevaluatie. Vanaf 2004 is overgegaan op een andere standaardvragenlijst die ook vragen bevat over het gebruik van ICT, opdrachten en dergelijke. De hier gerapporteerde gegevens hebben vrijwel steeds betrekking op de gemiddelde vraagscores van de recentere vragenlijst, berekend over 1229 verschillende cursussen uit de genoemde faculteiten. Van al die 1229 gemiddelde scores is opnieuw een gemiddelde berekend, zowel voor de universiteit als geheel als voor de verschillende faculteiten.

Onderzoeksvraag 3
Ten behoeve van de derde onderzoeksvraag, over de overeenkomst tussen de psychometrische en de door studenten gepercipieerde kwaliteit, is voor de laatste drie jaren (2008-2009, 2009-2010 en 2010-2011) een overzicht gemaakt van tentamens waarvan zowel een psychometrische analyse beschikbaar is, als de uitkomsten van de onderwijsevaluatie. Ook hier gaat het om meerkeuzetentamens met minimaal 25 deelnemers, en uitsluitend eerste kansen. Een extra beperking is dat het tentamen minstens uit veertig vragen moet bestaan, omdat de betrouw-

baarheid nu eenmaal sterk gerelateerd is aan het aantal toetsvragen. Daar komt bij dat tentamens met (duidelijk) minder vragen vaak meer het karakter hebben van diagnostische tussentoetsen dan van 'echte' eindtoetsen. Uiteindelijk leverde deze matching 193 tentamens op.

Uitkomsten

Onderzoeksvraag 1
Tabel 1 bevat een overzicht van de aan de Vrije Universiteit sinds het studiejaar 2005-2006 verwerkte meerkeuzetentamens. Het overzicht is beperkt tot de belangrijkste kwaliteitsindicatoren: de betrouwbaarheid (*KR-20*), het gemiddelde aantal items en de gemiddelde *p*-waarde.

Commentaar
Allereerst moet worden opgemerkt dat de faculteiten sterk verschillen in het aantal door de Tentamenservice verwerkte meerkeuzetentamens. Voor sommige faculteiten is dit aandeel laag, omdat ze geen of weinig grote opleidingen aanbieden, waarvoor meerkeuzetentamens nuttig of wenselijk zijn. De faculteit Geneeskunde (VUmc) verwerkt dan weer bijna alle tentamens zelf. De faculteit Wijsbegeerte heeft zelf kleine studentaantallen. De meerkeuzetentamens in deze faculteit zijn dan ook van studieonderdelen die door filosofiedocenten aan andere faculteiten worden gegeven.

Er zijn onderlinge verschillen tussen de faculteiten, maar die zijn niet erg groot. Acta (de faculteit Tandheelkunde) onderscheidt zich het sterkst door gemiddeld kleine aantallen items, omdat de meeste tentamens in deze faculteit een 'tussentoetskarakter' hebben. Het directe gevolg is een gemiddeld lage betrouwbaarheid, met uitzondering van het studiejaar 2010-2011. Deze faculteit is daarom niet helemaal vergelijkbaar met de andere faculteiten. Positieve uitschieters zijn Falw, Fbw, Fpp en Wijsbegeerte met een gemiddelde *KR-20* ≥ .70.
Over de hele VU berekend ligt de betrouwbaarheid in de onderzochte jaren gemiddeld op .66. Overigens moet hier wel worden opgemerkt dat de mogelijkheid bestaat dat het eindcijfer voor de cursus niet *uitsluitend* gebaseerd is op dit meerkeuzetentamen, maar dat er wellicht ook sprake was van andere toetsvormen (bijvoorbeeld open vragen en opdrachten), die samen het eindcijfer bepalen. Hierop wordt in de 'Conclusies en discussie' teruggekomen.
Ongeveer een kwart van de meerkeuzetentamens van de universiteit heeft een betrouwbaarheid die onder de .60 ligt. Een dergelijke betrouwbaarheid is te laag. Ter illustratie: in situaties waarin beslissingen over personen worden genomen – en de toetsing is zo'n situatie – is een betrouwbaarheid van minimaal .80 vereist, en van .90 of hoger wenselijk (Nunnally & Bernstein, 1994, p. 265). Een betrouwbaarheid boven de .90 is in de universitaire context, waarin meerkeuzetentamens zelden of nooit door professionals worden opgesteld, praktisch onhaalbaar. Bovendien hebben studenten via herkansingen steeds gelegenheid zich te 'revancheren'. Dit betekent dat beslissingen over studenten dus niet onherroepelijk zijn.

Tabel 1 Toetsgegevens over zes opeenvolgende jaren

Gegevens	Acta	Falw	Fbw	Few	Feweb	Fpp	Fsw	Let	Rec	Vumc	Wijs	VU
2005-2006	-			-								
N (= aantal tentamens)	-	5	12	-	32	24	21	-	7	1	-	**102**
Gemiddelde KR-20	-	.71	.73	-	.61	.70	.67	-	.62	.57	-	**.66**
% *KR 20* < .60	-	20%	17%	-	38%	17%	24%	-	43%	100%	-	**27%**
% *KR 20* ≥ .60 < .75	-	40%	25%	-	50%	38%	52%	-	43%	0%	-	43%
% KR 20 ≥ .75	-	40%	58%	-	13%	42%	24%	-	14%	0%	-	**28%**
Gemiddeld aantal items	-	50	53	-	35	52	34	-	36	70	-	**42**
Gemiddelde *p*-waarde	-	.69	.68	-	.65	.64	.61	-	.59	.70	-	**.64**
2006-2007												
N (= aantal tentamens)	11	18	14	-	35	28	21	-	9	1	-	**137**
Gemiddelde *KR-20*	.48	.71	.71	-	.66	.71	.65	-	.68	.47	-	**.67**
% *KR 20* < *.60*	64%	11%	7%	-	23%	14%	19%	-	22%	100%	-	**21%**
% *KR 20* ≥ .60 < .75	36%	33%	71%	-	43%	32%	62%	-	56%	0%	-	**45%**
% *KR 20* ≥ .75	0%	56%	21%	-	34%	54%	19%	-	22%	0%	-	**34%**
Gemiddeld aantal items	22	54	50	-	39	47	34	-	36	70	-	**42**
Gemiddelde *p*-waarde	.74	.68	.64	-	.63	.63	.63	-	.61	.67	-	**.64**
2007-2008												
N (= aantal tentamens)	25	21	14	2	29	33	27	-	11	3	-	**165**
Gemiddelde *KR-20*	.36	.73	.71	.59	.67	.74	.64	-	.64	.49	-	**.61**
% *KR 20* < .60	76%	14%	14%	50%	21%	12%	30%	-	27%	67%	-	**29%**
% KR 20 ≥ .60 < .75	20%	38%	43%	0%	45%	24%	44%	-	55%	0%	-	**35%**
% *KR 20* ≥ .75	4%	48%	43%	50%	34%	64%	22%	-	18%	33%	-	**35%**
Gemiddeld aantal items	22	54	52	47	38	48	32	-	28	47	-	**40**
Gemiddelde *p*-waarde	.71	.64	.64	.54	.64	.64	.59	-	.63	.69	-	**.63**
2008-2009												
N (= aantal tentamens)	26	21	13	4	38	32	28	-	10	1	-	**173**
Gemiddelde *KR-20*	.60	.74	.71	.66	.63	.69	.71	-	.65	.60	-	**.67**
% *KR 20* < .60	35%	0%	8%	50%	26%	19%	7%	-	20%	0%	-	**18%**
% *KR 20* ≥ .60 < .75	54%	52%	46%	25%	42%	41%	61%	-	60%	100%	-	**49%**
% *KR 20* ≥ .75	12%	48%	46%	25%	26%	41%	32%	-	20%	0%	-	**31%**

Vervolg Tabel 1

	Acta	Falw	Fbw	Few	Feweb	Fpp	Fsw	Let	Rec	Vumc	Wijs	VU
Gemiddeld aantal items	33	57	48	41	35	45	37		79	70	-	**43**
Gemiddelde *p*-waarde	.75	.65	.61	.57	.64	.64	.65		.64	.74	-	**.66**
2009-2010												
N (= aantal tentamens)	34	23	15	3	37	34	41		18	-	-	**205**
Gemiddelde *KR-20*	.45	.72	.71	.55	.64	.68	.72		.67	-	-	**.64**
% *KR 20* < .60	71%	17%	20%	33%	27%	21%	10%		22%	-	-	**28%**
% *KR 20* ≥ .60 < .75	6%	39%	33%	33%	54%	38%	54%		61%	-	-	**41%**
% *KR 20* ≥ .75	24%	43%	47%	33%	19%	41%	37%		17%	-	-	**31%**
Gemiddeld aantal items	28	52	52	45	36	48	39		36	-	-	**40**
Gemiddelde *p*-waarde	.74	.65	.64	.55	.64	.65	.64		.62	-	-	**.66**
2010-2011												
N (= aantal tentamens)	24	30	15	5	58	43	52	3	22	1	5	258
Gemiddelde *KR-20*	.63	.71	.67	.71	.63	.73	.70	.61	.68	.44	.74	.68
% *KR 20* < .60	38%	13%	40%	0%	36%	9%	13%	0%	9%	100%	0%	22%
% *KR 20* ≥ .60 < .75	33%	43%	20%	80%	31%	35%	63%	67%	64%	0%	40%	42%
% *KR 20* ≥ .75	29%	43%	40%	20%	33%	56%	23%	33%	27%	0%	60%	36%
Gemiddeld aantal items	42	52	46	38	38	47	39	36	32	20	51	42
Gemiddelde *p*-waarde	.75	.65	.66	.63	.63	.65	.62	.72	.65	.56	.71	.65
Totaal												
N (= aantal tentamens)	120	118	83	14	229	194	190	3	77	7	5	1040
Gemiddelde KR-20	.50	.72	.71	.65	.64	.71	.69	.61	.66	.51	.74	.66
% *KR 20* < .60	57%	12%	18%	29%	29%	15%	16%	0%	21%	71%	0%	24%
% *KR 20* ≥ .60 < .75	27%	41%	40%	43%	43%	35%	57%	67%	59%	14%	40%	43%
% *KR 20* ≥ .75	16%	47%	42%	29%	27%	50%	27%	33%	21%	14%	60%	33%
Gemiddeld aantal items	30	54	50	41	37	48	37	36	39	53	51	41
Gemiddelde *p*-waarde	.74	.66	.64	.58	.64	.64	.63	.72	.63	.68	.71	.65

Falw = Aard & levenswetenschappen; Fbw = Bewegingswetenschappen; Few = Exacte wetenschappen; Feweb = Economische wetenschappen & Bedrijfskunde; Fpp = Psychologie & Pedagogiek; Fsw = Sociale wetenschappen; Let = Letteren; Rec = Rechten; Wijs = Wijsbegeerte

Tabel 2 Vragenlijstgemiddeldes

Vragenlijstgemiddeldes										
	Falw	**Fbw**	**Few**	**Feweb**	**Fpp**	**Fsw**	**Let**	**Rec**	**Wijs**	**VU**
Cursus	3.82	3.78	3.71	3.72	3.88	3.82	3.99	4.12	4.25	3.88
Docent	3.78	3.91	3.70	3.76	3.93	3.92	4.07	4.22	4.22	3.89
Representatief	3.73	3.73	3.74	3.63	3.71	3.77	3.92	3.87	4.10	3.81
Transparant	3.28		3.48	3.06	3.10	3.49	3.52	3.49	3.71	3.46
Totaaloordeel	3.56	3.59	3.54	3.38	3.57	3.61	3.82	3.80	4.05	3.67
Gezakt	8.7	13.2	11.8	7.2	9.3	5.4	5.1	2.8	5.2	7.8
Twijfelachtig	41.1	50.8	35.2	24.1	43.0	28.6	30.1	31.2	30.3	33.0
Geslaagd	50.2	35.9	53.0	68.7	47.7	66.0	64.3	66.0	64.5	59.2
Slagingsverwachting	3.42	3.22	3.41	3.62	3.38	3.61	3.58	3.63	3.59	3.51
N cursussen	292	397	393	136	70	113	252	21	63	1229

Falw = Aard & levenswetenschappen; Fbw = Bewegingswetenschappen; Few = Exacte wetenschappen; Feweb = Economische wetenschappen & Bedrijfskunde; Fpp = Psychologie & Pedagogiek; Fsw = Sociale wetenschappen; Let = Letteren; Rec = Rechten; Wijs = Wijsbegeerte

Toch dient een betrouwbaarheid van .80 te worden nagestreefd, hetzij door het aantal (goede) toetsvragen te vergroten, hetzij door de toets te combineren met de score voor bijvoorbeeld open vragen en/of een eindwerkstuk.
Dousma, Horsten en Brants (1997) geven voor een toets met een betrouwbaarheid van .70 en een percentage gezakten van 30% een percentage foute beslissingen van 22%. Dit zijn dus studenten die ten onrechte, namelijk als gevolg van gebreken in het tentamen, gezakt of geslaagd zijn. Alleen al de rol die de toetsing in het nieuwe accreditatiestelsel inneemt (zie ook de 'Inleiding') maakt dat men hier niet onverschillig onder kan blijven.

Onderzoeksvraag 2
Tabel 2 bevat de gemiddelde scores per faculteit en voor de VU als geheel op de volgende evaluatievragen: Cursusoordeel; Docentoordeel; Tentamen representatief; Tentamen transparant; Moeilijkheidsgraad tentamen; Totaaloordeel tentamen; % gezakt; % twijfelachtig; % geslaagd; Slagingsverwachting (zie voor de exacte vraagformulering appendix A).
In het overzicht van de faculteiten ontbreken Tandheelkunde en Geneeskunde, omdat die in hoofdzaak gebruik maken van eigen, niet gestandaardiseerde vragenlijsten.

Commentaar
Zoals uit de gemiddelde waardering voor cursus en docent blijkt, zijn VU-studenten over het algemeen (erg) tevreden over het aan hen gegeven onderwijs. Faculteiten verschillen in dit opzicht ook weinig, al steekt vooral Wijsbegeerte er in positieve zin bovenuit. Hierbij moet wel worden aangetekend dat in deze facul-

Tabel 3 Correlaties tussen evaluatie-uitkomsten en toetsgegevens.

	N vragen	KR-20	KR-20 → 75	p-waarde	% geslaagd	N rapport	% rapport
Representativiteit	.31**	.22**	.18*	.39**	0,.41**	-.21**	-.35**
Transparantie	-.08	.18*	.16	.30**	.28**	-.32**	-.27**
Totaaloordeel	.08	.21**	.20**	.41**	.42**	-.32**	-.35**
Slagingsverwachting	-.09	.01	.14	.52**	.59**	-.41**	-.38**

*p < .05; **p < .01

teit, net als bij Letteren, vaak sprake is van kleine groepen, wat doorgaans scoreverhogend werkt.
Des te meer valt de relatief lage waardering voor het tentamen op. Dit geldt nog niet zozeer voor de representativiteit ervan, maar wel in hoge mate voor de ervaren transparantie: voor de VU als geheel is de score op die vraag 3.46, dat wil zeggen ongeveer vier tiende punt lager dan de waardering voor cursus of docent. Dit resulteert in een totaaloordeel dat steeds ongeveer twee tiende punt lager is dan de score voor docent en cursus. De correlatie tussen de ervaren transparantie en het totaaloordeel bedraagt .56, die tussen de representativiteit en het totaaloordeel zelfs .83. Het totaaloordeel hangt minder sterk samen met de gemiddelde slagingsverwachting: de correlatie bedraagt .51. Volledigheidshalve dient hier te worden opgemerkt dat de genoemde correlatiecoëfficiënten niet berekend zijn voor het databestand uit tabel 2, maar voor een subset (N = 825) met complete data uit het totale VU-bestand van 1229 cursusafnames waarin geen vragen voorkomen die niet van toepassing zijn voor de desbetreffende cursus.
De conclusie is dat de waardering van studenten over de kwaliteit van het tentamen veel lager is dan hun waardering voor docent of cursus. Dat minder positieve oordeel spitst zich toe op de ervaren transparantie, dat wil zeggen de mate waarin studenten vooraf wisten wat zij van het tentamen konden verwachten.

Onderzoeksvraag 3
Tabel 3 bevat de correlaties tussen de evaluatievragen over de representativiteit, de transparantie, de moeilijkheidsgraad, het totaaloordeel en de slagingsverwachting met de volgende psychometrische indices: het aantal vragen, de *KR-20*, de *KR-20* bij homogene testverlenging naar 75 vragen, de gemiddelde *p*-waarde, het slagingspercentage, het aantal vragen waarover door de Tentamenservice is gerapporteerd (wegens een lage *p*-waarde en/of r_{ir}-waarde), en het aantal vragen waarover is gerapporteerd als percentage van het aantal gestelde vragen.

Commentaar
Dat de ervaren representativiteit van het tentamen samenhangt met het aantal gestelde vragen is vermoedelijk het minst verrassend van de gevonden correlaties. Het verband met de andere indices ligt minder voor de hand, zeker als het gaat

om de gemiddelde *p*-waarde en het slagingspercentage. Desalniettemin is het duidelijk dat deze evaluatievraag een goede indicator is van de tentamenkwaliteit. Gezien de eerder vermelde hoge correlatie tussen de representativiteit en het totaaloordeel (r = .83) is het weinig verbazend dat de verbanden met de laatst genoemde evaluatievraag en de tentamengegevens nauw overeenkomen, behalve dan ten aanzien van het aantal gestelde tentamenvragen.
De ervaren transparantie correleert zwak met de betrouwbaarheid, maar weer redelijk met *p*-waarde, slagingspercentage, het aantal vragen waarover is gerapporteerd en het percentage vragen waarover is gerapporteerd. De laatste twee variabelen correleren eigenlijk met alle evaluatievragen redelijk. Het komt er op neer dat, hoe meer vragen er door de Tentamenservice zijn gerapporteerd als problematisch, hoe negatiever de studenten van hun kant ook zijn over de kwaliteit van het tentamen in kwestie. Het is van belang hier nog een keer op te merken dat de hier gerapporteerde evaluatie- en tentamengegevens onafhankelijk van elkaar zijn verkregen.
De slagingsverwachting, ten slotte, hangt niet samen met het aantal vragen en de betrouwbaarheid, maar wel – en zelfs sterk – met de gemiddelde *p*-waarde en het slagingspercentage. Daarnaast worden studenten negatiever met betrekking tot hun slagingsverwachting naarmate het tentamen volgens hen meer mankementen vertoont, zoals blijkt uit de negatieve correlaties tussen dat studentenoordeel met het aantal en percentage problematische tentamenvragen.

Conclusies en discussie

Een eerste conclusie is dat de kwaliteit van meerkeuzetentamens voor wat betreft de betrouwbaarheid onder de maat is. Bij enkele faculteiten wordt met enige regelmaat gevraagd om, voorafgaand aan het tentamen, te adviseren over de door de docent opgestelde tentamenvragen. Die adviezen blijven beperkt tot de meer formele aspecten van die vragen. Het verdient aanbeveling om, indien dit nog niet gebeurt, met collega's die al dan niet aan hetzelfde onderdeel verbonden zijn ook in inhoudelijk opzicht over het tentamen te overleggen.
Ten tweede is de betrouwbaarheid van meerkeuzetentamens sterk afhankelijk van het aantal tentamenvragen. Uitbreiding van het aantal vragen verdient dus altijd de voorkeur. Hierbij kan ook gedacht worden aan het combineren van het meerkeuzetentamen met een aantal open vragen – liefst meer dan vijf. Het gecombineerde cijfer is in de praktijk altijd een stuk betrouwbaarder dan de afzonderlijke deelcijfers. Een voorwaarde hierbij is wel dat die antwoorden op open vragen door twee beoordelaars worden nagekeken, die vooraf duidelijke afspraken hebben gemaakt over de criteria/puntentoekenning.
In sommige faculteiten bestaan tentamens uit meerkeuzevragen die vooral betrekking hebben op de stofbeheersing en uit een aantal cases uit de praktijk. De beantwoording van dergelijke cases kan ook worden geanalyseerd in psychometrische zin, en de correlatie tussen deze vragen en de meerkeuzevragen is een goede schatter van de betrouwbaarheid van het hele tentamen. De ervaringen hiermee zijn beperkt, maar positief.

Al eerder is gevonden dat studenten na afloop van het tentamen, zowel individueel als groepsgewijs, goed in staat zijn om aan te geven of ze (waarschijnlijk) gezakt of geslaagd zijn, dan wel of het 'erom hangt' (Van Os, 1999, p. 101-102). Ook in dit onderzoek komt dat tot uitdrukking. Een daarop aansluitende conclusie is dat de waardering van studenten voor de kwaliteit van het tentamen natuurlijk geen 'volmaakte' indicator is van de kwaliteit zoals die naar voren komt in de itemanalyse. Daar staat tegenover dat dit studentenoordeel bepaald niet onafhankelijk is van die objectieve kwaliteit. De gevonden en hier gerapporteerde verbanden zijn veelal niet erg groot (vanaf ongeveer 10% tot maximaal 35% verklaarde variantie) maar zeker niet verwaarloosbaar. Anders gezegd: wanneer studenten via een evaluatievragenlijst aangeven dat een tentamen kwalitatief niet acceptabel is, dan verdient het zonder meer aanbeveling dat tentamen eens nauwkeurig onder de loep te nemen. Voor de meerkeuzetentamens gebeurt dit al standaard door de Tentamenservice. Voor tentamens met open vragen is dat echter meestal niet het geval en er is reden om dit systematischer ter hand te nemen, mede gezien de ontwikkelingen op het gebied van het hoger onderwijs die aan het begin van dit artikel werden besproken. Recent is op de Vrije Universiteit dan ook begonnen met de analyse van tentamens met open vragen op een manier die analoog is aan die van meerkeuzetentamens.
Aan de eis dat een tentamen transparant moet zijn, kan gemakkelijk worden voldaan, bijvoorbeeld door middel van het bespreken van voorbeeldvragen of het beschikbaar stellen van een proeftentamen. Het is opmerkelijk hoe laag op deze vraag wordt gescoord, en de opmerkingen die studenten geven op de evaluatieformulieren betreffen dan ook dikwijls klachten hierover.
Ten slotte: zoals eerder opgemerkt is geen onderzoek bekend dat in meerdere of mindere mate overeenkomt met het onderhavige. Hier wordt dan ook een pleidooi gehouden voor ten eerste de systematische registratie van tentamengegevens zoals die in deze bijdrage zijn besproken en ten tweede de invoering en het consequente gebruik van gestandaardiseerde evaluatievragenlijsten. Dit kost tijd en dus geld. Het onderwerp in kwestie – de verbetering van de tentamenkwaliteit – is er echter belangrijk genoeg voor.

De auteurs zijn erkentelijk voor de bijdrage van drs. G.C. Reumer, zonder wiens inspanningen voor de opzet en uitvoering van zowel de tentamen- als de evaluatieadministratie dit artikel niet geschreven had kunnen worden. Hetzelfde geldt voor de vele student-assistenten die in de loop der jaren hebben bijgedragen aan de tentamenverwerking en -analyse, en aan het zorgvuldig en gewetensvol bijhouden van de administratie.

Referenties

Abrami, P.C., d'Apollonia, S. & Cohen, P.A. (1990). Validity of student ratings of instruction: What we know and what we do not. *Journal of Educational Psychology*, *82*, 219-231.

Dousma, T., Horsten, A. & Brants, J. (1997). *Tentamineren*. Groningen: Wolters-Noordhoff.

Marsh, H.W. (1987). Students' evaluations of university teaching: Research findings, methodological issues and directions for future research. *International Journal of Educational Rese*arch, *11*, 253-388.

Nunnally, J.C. & Bernstein, J.H. (1994). *Psychometric theory* (3rd ed.). New York: Mc Graw-Hill.

Os, W. van (1999). *Bruikbaarheid en Effectiviteit van Studentenoordelen over het Onderwijs*. Academisch proefschrift. Amsterdam: Vrije Universiteit.

Patrick, C.L. (2011). Student evaluation of teaching: Effects of the Big Five personality traits, grades and the validity hypothesis. *Assessment & Evaluation in Higher Education*, *36*, 2, 239-249.

Appendix A Tekst evaluatievragen

1 Cursus: Totaaloordeel over de inhoudelijke kwaliteit van dit studieonderdeel (*zeer slecht* tot zeer goed).
2 Docent: Totaaloordeel over de didactische kwaliteiten van de docent (*zeer slecht* tot zeer goed).
3 Representativiteit: Het tentamen bestreek voldoende de hele te bestuderen stof (*zeer oneens* tot zeer eens).
4 Transparantie: Via oefenvragen, proeftentamens, studieaanwijzingen en dergelijke wist ik vooraf goed wat ik op het tentamen kon verwachten (*zeer oneens* tot zeer eens).
5 Totaaloordeel: Het tentamen was een goede graadmeter van wat ik in deze cursus heb geleerd (*zeer oneens* tot *zeer eens*).
6 Slagingsverwachting: Mijn verwachting over het tentamen is: *gezakt; twijfelachtig; geslaagd*.