Dit artikel van Boom Lemma Tijdschriften is gemaakt voor Vrije Universiteit 180209

De lage respons bij digitale onderwijsevaluaties: een overschat probleem?

Willem van Os & Marlies van Beek[*]

Onderwijsevaluaties worden steeds vaker digitaal uitgevoerd – via e-mail of een link naar een website. Dit heeft duidelijke logistieke voordelen. Het belangrijkste nadeel betreft de respons die doorgaans aanzienlijk lager is dan bij 'papieren' evaluaties. In dit artikel wordt ingegaan op de vraag of die lagere respons ook consequenties heeft voor de antwoordpatronen (systematisch hoger of lager), dan wel of de resultaten vergeleken met eerdere papieren evaluaties sterk afwijken. Hiertoe zijn de uitkomsten van 21 cursussen die eerst op papier zijn geëvalueerd vergeleken met die van (dezelfde) cursussen waarbij de afname bij een latere gelegenheid digitaal plaatsvond. Gemiddeld waren er nauwelijks verschillen. Per cursus waren die verschillen er onder bepaalde voorwaarden wel. Die voorwaarden hebben primair betrekking op het minimaal vereiste aantal respondenten: ligt dat onder de vijftien, dan neemt de kans op onverwachte en ook slecht verklaarbare uitkomsten toe.

Inleiding

De opkomst van digitale evaluaties

Onderwijsevaluatie is sinds lang gemeengoed aan vrijwel alle instellingen voor hoger onderwijs, in Nederland zowel als elders. Nu internet in toenemende mate beschikbaar is, wordt ook de digitale afname van vragenlijsten steeds aantrekkelijker. Johnson (2003) noemt als voordelen het gemak waarmee ze kunnen worden afgenomen, langere en meer doordachte antwoorden van studenten (bij open vragen), lagere kosten en een snellere verwerkingstijd. Met betrekking tot die lagere kosten komen Bothell en Henderson (2003) bijvoorbeeld op een bedrag van $ 0,47 voor een digitaal formulier, tegenover $ 1,06 voor een papieren afname. Hierbij hebben zij rekening gehouden met de vaak onderschatte ontwikkelingskosten van een digitaal evaluatiesysteem. Deze bedragen zullen niet overal geldig zijn, maar wel kan worden aangenomen dat de operationele kosten van digitaal evalueren in het algemeen aanzienlijk lager zullen liggen dan die van papieren evaluaties.

Het responsprobleem

Het meest genoemde nadeel van digitale onderwijsevaluaties heeft betrekking op de respons: die is consistent (veel) lager dan bij evaluaties op papier. Johnson

* Dr. W. van Os (w.van.os@vu.nl) is werkzaam bij de Taakgroep Onderwijsevaluatie & Kwaliteitszorg, Onderwijscentrum Vrije Universiteit Amsterdam. Drs. M.E. van Beek is werkzaam bij de Taakgroep Onderwijsevaluatie & Kwaliteitszorg, Onderwijscentrum Vrije Universiteit Amsterdam.

Dit artikel van Boom Lemma Tijdschriften is gemaakt voor Vrije Universiteit 180209

(2003) meldt een respons van 71% voor papieren en 50% voor digitale afnames, berekend over 74 cursussen. Dit zijn beide erg hoge percentages vergeleken met die van Sax, Gilmartin en Bryant (2003) en van Porter en Whitcomb (2003) met responspercentages van respectievelijk 21,5% en 15,2%. Ook bij louter papieren evaluaties kan de respons overigens erg verschillen. Aan de Vrije Universiteit Amsterdam varieerde in 2009-2010 het responspercentage bijvoorbeeld van 4,8% tot 100% (berekend over 648 cursussen met een gemiddelde respons van 71,4%). Van belang is vooral de mate waarin studenten van de zin van evalueren zijn overtuigd en de moeite die de faculteit doet om studenten te informeren over de maatregelen die naar aanleiding van onderwijsevaluaties worden genomen (Nulty, 2007). In een overzicht van 9 studies berekent Nulty een digitale respons van 33% versus een respons van 56% bij papieren surveys. Zie in dit verband ook Cook, Heath en Thompson (2000). Alles bijeen genomen wijzen de gegevens erop dat, vergeleken met situaties waarin vragenlijsten 'in de klas' worden afgenomen, de respons bij digitale afnames (via e-mail of internet) beduidend lager is. Bij Nulty scheelt het ruim 20%. De vraag is allereerst hoe erg dat is.

Is een lage respons een probleem?

Een lage respons kan om twee redenen problematisch zijn.

1. Vergeleken met eerdere papieren evaluaties zijn de uitkomsten van digitale afnames systematisch hoger of lager. Dit is het geval wanneer vooral studenten responderen die meer of minder tevreden zijn over de cursus of de docent. De respondenten vormen dan een niet-representatieve steekproef uit de populatie, en hun waardering voor het onderwijs vertegenwoordigt niet de mening van alle studenten.
2. Het aantal respondenten is zo klein dat toevallige uitschieters, in positieve of negatieve zin, te sterk de uitkomsten van de evaluatie in kwestie bepalen. Dit hoeft niet te leiden tot verschillen in de gemiddelde waardering, berekend over een aantal cursussen tezamen. Immers, bij de ene cursus kunnen de uitkomsten hoger zijn, bij de andere juist lager. Per cursus leidt het echter wel van jaar tot jaar tot uiteenlopende resultaten, ook al is de cursus of de docent in wezen hetzelfde gebleven.

Gezien de doelen van onderwijsevaluatie zijn beide mogelijkheden hoogst ongewenst. In aansluiting op Marsh (1984) zijn die doelen in de eerste plaats het aanbrengen van onderwijsverbeteringen, en in de tweede plaats – en dat geldt de laatste jaren in toenemende mate – het gebruik van de uitkomsten in het kader van het personeelsbeleid. Docenten zullen minder geneigd zijn hun onderwijs te veranderen wanneer zij het gevoel hebben dat de kritiek alleen of vooral afkomstig is van ontevreden studenten, en dat studenten die wél tevreden zijn niet hebben gerespondeerd. Beslissingen op personeelsgebied zullen daarnaast ook als minder gerechtvaardigd worden gezien wanneer zou blijken dat dezelfde docent met dezelfde – feitelijk onveranderde – cursus in het ene jaar heel anders wordt gewaardeerd dan in het andere jaar.

Dit artikel van Boom Lemma Tijdschriften is gemaakt voor Vrije Universiteit 180209

Enkele gegevens uit de literatuur

Leung en Kember (2005) vonden weinig verschillen in interne structuur tussen de beantwoording via papier of digitaal; ook de betrouwbaarheid van de schalen (Cronbach's alpha) kwam overeen, alsmede de hoogte van de beoordelingen. Bij Carini, Hayek, Kuh, Kennedy en Quimet (2003) waren er geringe inhoudelijke antwoordverschillen tussen digitale en papieren beantwoording van de National Survey of Student Engagement. De antwoordpatronen waren echter niet identiek: op alle acht schalen scoorden respondenten digitaal positiever dan op papier. Ook Johnson (2003) vermeldt gemiddeld hogere scores bij digitale beoordelingen. De door hem berekende correlatie tussen de papieren en digitale overall-beoordeling van docenten bedraagt 0,84, en die van de cursus 0,86. In alle voorgaande gevallen gaat het om parallelonderzoek waarbij studenten voor de keuze worden gesteld om hetzij op de ene, hetzij op de andere wijze te responderen. Bij herhaalonderzoek betreft het situaties waarin overgegaan wordt op digitaal evalueren, en de uitkomsten van jaar tot jaar worden vergeleken. Dit is uiteraard alleen zinvol wanneer de vragenlijst net als het onderwijs gelijk is gebleven. Het uitgangspunt is dus dat alleen de onderzoeksgroep (studenten) en het medium (papier versus digitaal) zijn veranderd. Dergelijk onderzoek heeft tot op heden slechts sporadisch plaatsgevonden.

Bij een vergelijking binnen de Faculteit der Aard- & Levenswetenschappen aan de Vrije Universiteit tussen papieren en (een jaar later) digitale afnames rapporteert Van Os (2010) hogere overall-scores voor docent, cursus en tentamen voor de afname op papier. De verschillen waren klein, behalve bij die voor het tentamen. Overigens betrof het slechts acht cursussen, zodat de zeggingskracht beperkt is. De hertest-correlaties – waarbij per cursus wordt nagegaan in hoeverre de vraagscores overeenkomen wanneer hetzij op papier, hetzij digitaal wordt geëvalueerd – waren hier gemiddeld hoog. Wanneer het ging om een papieren afname versus (een jaar later) een papieren afname was die 0,73, en bij een papieren versus (eveneens een jaar later) een digitale afname zelfs 0,78. De sterkere en zwakkere punten van docent en cursus worden kennelijk ook door de studenten van het latere cohort onderkend, en hierbij maakt het weinig uit via welk medium er wordt geëvalueerd. Dit komt overeen met de bevindingen van Van Os (1999, p. 187), waarin correlaties werden gevonden van ongeveer 0,80 wanneer het gaat om herhaalevaluaties van dezelfde cursus met dezelfde docent, en van 0,20 bij dezelfde cursus met een andere docent. Marsh (1984, p. 718) meldt in dit verband correlaties van gemiddeld 0,69 (dezelfde docent, dezelfde cursus), en 0,49 (andere docent, dezelfde cursus). Zowel bij Marsh als bij Van Os gaat het overigens wel om evaluaties op papier.

Zo op het oog lijkt er dus weinig aan de hand te zijn. Dat neemt niet weg dat er ook signalen zijn dat met enige regelmaat sterk afwijkende uitslagen worden gevonden die niet kunnen worden verklaard door bijzondere gebeurtenissen binnen de cursus en/of met de docent. Zo spreken in een recent intern onderzoek aan de Vrije Universiteit met betrekking tot de dienstverlening diverse docenten hun twijfels uit over de waarde van digitale onderwijsevaluaties die gebaseerd zijn

Dit artikel van Boom Lemma Tijdschriften is gemaakt voor Vrije Universiteit 180209

op slechts enkele studenten (Hollander & Van Os, 2010). Daarom is besloten om het onderzoek bij de Faculteit der Aard- & Levenswetenschappen in een andere context en op wat grotere schaal te herhalen.

Vraagstellingen

De primaire vraagstelling luidt of het gebruik van papieren of digitale evaluaties (via e-mail of een link naar een website) tot andere uitkomsten leidt in termen van systematisch hogere of lagere scores. De tweede vraag is of de hierboven vermelde stabiliteit van onderwijsevaluaties – hier gedefinieerd als de hertest-betrouwbaarheid – in gelijke mate opgaat voor digitale en papieren afnames en zo niet, van welke factoren dat afhankelijk is.

Onderzoeksgroep en methoden
Aan de Vrije Universiteit Amsterdam worden jaarlijks cursussen gegeven in het kader van HOVO: Hoger Onderwijs Voor Ouderen. Ze zijn bedoeld voor cursisten vanaf vijftig jaar en in die zin niet toegankelijk voor studenten. Doorgaans gaat het zowel in het voorjaars- als in het najaarsprogramma om ongeveer veertig verschillende cursussen (met elk meestal circa tien bijeenkomsten van twee uur), plus een zomerprogramma in de maand juli met vier à vijf bijeenkomsten. Er is een grote mate van afwisseling in het cursusaanbod, al worden sommige cursussen over een reeks van jaren gegeven. Enkele voorbeelden zijn 'Art Nouveau – Jugendstil', 'Oorsprong en evolutie van de mens', 'Actualiteiten Ruimtelijke Ordening' en 'Kopstukken van de moderne filosofie'. De waardering is gemiddeld bijzonder hoog: voor de cursusinhoud en voor de docent (overall-vragen) respectievelijk 4,37 en 4,45 op een vijfpuntsschaal. Voor een belangrijk deel gaat het om gepensioneerde of nog actieve (universitaire) docenten die, voor zover ze van de VU afkomstig zijn, ook door reguliere studenten erg positief werden of worden beoordeeld. Voor de cursisten geldt in dit verband dat gemiddeld 40% een afgeronde hbo-opleiding heeft, en 39% een wo-opleiding. Het overgrote deel is dus hoogopgeleid.

De cursussen worden na afloop geëvalueerd met behulp van een vragenlijst die tot voor enkele jaren via de post werd opgestuurd (zie voor de tekst bijlage 1). De betrouwbaarheid van de vragenlijst (Cronbach's alpha) bedraagt 0,97. De respons op deze postale enquêtes was steeds relatief hoog: 60,7%, berekend over 466 cursussen sinds najaar 2001. Dit is weliswaar lager dan de in de inleiding genoemde percentages bij afname 'in de klas' of op het tentamen, maar daar moet het ook niet mee worden vergeleken. De respons op postale enquêtes ligt gemiddeld op ongeveer 50% (Richardson, 2005). Vooral om administratieve redenen is besloten om met ingang van zomer 2009 over te gaan op digitale afnames. Van 21 HOVO-cursussen is inmiddels zowel een papieren als een digitale evaluatie beschikbaar. Hiervoor geldt het volgende. Het aantal deelnemers aan de op papier geëvalueerde cursussen varieert van 21 tot 81. Het aantal respondenten loopt uiteen van 10 tot 41. De responspercentages variëren van 44% tot 76%. Bij de digitale afna-

Dit artikel van Boom Lemma Tijdschriften is gemaakt voor Vrije Universiteit 180209

mes bedraagt het aantal deelnemers 20 tot 78, en is het aantal respondenten 9 tot 36. Hier ligt het responspercentage tussen de 24% en 58%. De gemiddelde papieren respons is 59%, de digitale 44%. Het verschil, 15%, is geringer dan de door Nulty (2007) vermelde 20%, maar gaat wel in die richting.

Voor de beantwoording van de primaire vraagstelling is nagegaan of de gemiddelde waardering voor cursus en docent op de desbetreffende overall-vragen (vraag 8 en 14 - zie bijlage 1) verschillend is, afhankelijk van de wijze van evalueren (op papier of digitaal). Voor de beantwoording van de tweede vraag is van de 21 cursussen die zowel op papier als (bij een volgende gelegenheid) digitaal zijn geëvalueerd, de hertest-correlatie berekend tussen enerzijds de papieren, anderzijds de digitale afname. Tevens is nagegaan of, en zo ja in hoeverre, deze correlaties gemiddeld afwijken van die welke uit de literatuur bekend zijn, en van hertest-correlaties bij HOVO-cursussen vóórdat er digitaal werd geëvalueerd.

Resultaten

Gemiddelde overall-score voor cursus en docent
Tabel 1 bevat de gemiddelde score op de overall-vragen over de cursus en de docent voor de 21 cursussen die zowel op papier als digitaal zijn geëvalueerd.

Tabel 1: Gemiddelde score papieren versus digitale afname

	Cursus		**Docent**		***N***
	gem	***s.d.***	***gem***	***s.d.***	
Papier	4,42	0,36	4,48	0,41	21
Digitaal	4,39	0,33	4,50	0,43	21
T	*0,39 (ns)*		*-0,20 (ns)*		

Gemiddelde op een schaal van 1 tot 5 waarbij 1 = zeer slecht en 5 = zeer goed

Papieren en digitale afnames leveren, gemiddeld over deze 21 cursussen, vrijwel identieke scores op. Het verschil is verwaarloosbaar en niet significant. Een wat ander beeld ontstaat wanneer per cursus gekeken wordt naar het verschil tussen de papieren en digitale afname en men het teken veronachtzaamt (zodat het niet uitmaakt of de score op de digitale afname nu hoger of lager is). In absolute zin bedraagt het gemiddelde verschil dan namelijk 0,27 bij de overall-score over de cursus, en 0,31 over de docent. Dit zijn naar verhouding grote verschillen. Dit komt ook tot uitdrukking in de gemiddelde correlatie tussen papieren en digitale afnames (berekend over alle vragen) van deze 21 cursussen: r = 0,60, duidelijk lager dus dan de eerder genoemde correlaties van achtereenvolgens 0,80 (Van Os, 1999) en 0,69 (Marsh, 1984). Om na te gaan of dit inderdaad moet worden toegeschreven aan het gebruikte medium, is bij 77 HOVO-cursussen die meermaals (soms wel zes keer) op papier zijn geëvalueerd de overeenkomstige hertest-correlatie berekend.

Dit artikel van Boom Lemma Tijdschriften is gemaakt voor Vrije Universiteit 180209

Correlaties tussen opeenvolgende afnames

Wanneer cursussen meer dan twee keer zijn geëvalueerd hebben de correlaties betrekking op de eerste keer versus de tweede keer, en de tweede keer versus de derde keer, maar niet op de eerste versus de derde keer. Het gaat dus om *opeenvolgende* afnames. Op die manier zijn er in totaal 131 correlatiecoëfficiënten, met als gemiddelde r = 0,73. Hertest-correlaties bij evaluaties op papier zijn onderling gemiddeld dus hoger dan bij evaluaties die eerst op papier, later digitaal zijn uitgevoerd. In een vervolganalyse is in de eerste plaats gekeken naar de verstreken tijd tussen de opeenvolgende afnames. De veronderstelling hierbij is dat naarmate die tijd groter is, de kans op afwijkende uitkomsten (ten opzichte van de voorgaande afname) ook zal toenemen, met als gevolg een lagere hertest-correlatie. Het tijdsverschil tussen een papieren en een digitale afname bedraagt gemiddeld 2,3 jaar, die tussen papieren afnames onderling 1,3 jaar. Daar staat echter het opmerkelijke feit tegenover dat naarmate het tijdsverschil tussen een papieren en een digitale evaluatie van een cursus groter is, de correlatie tussen beide afnames óók toeneemt. Het verband is zelfs tamelijk sterk: r tijd-hertestcorrelatie = 0,49! De hierboven geformuleerde veronderstelling (hoe groter tijdsverschil, hoe groter de kans op afwijkende uitkomsten en hoe lager de hertest-correlatie) blijkt dus onjuist te zijn. Het tijdsverschil kan daarmee geen reden zijn voor de lagere hertest-correlaties bij digitale afnames.

Een tweede vervolganalyse heeft daarom betrekking op de respons, meer in het bijzonder het verschil in respons tussen papieren afnames onderling en papieren afnames ten opzichte van digitale afnames. De gedachte is dat naarmate het verschil in respons tussen twee afnames groter is en/of naarmate de respons bij één of beide afnames in absolute zin kleiner is, de kans op lagere hertest-correlaties toeneemt. Uit een uitgevoerde regressieanalyse ('stepwise') blijkt dat geen van de variabelen (gemiddelde *N*, gemiddelde respons, verschil in *N*, verschil in respons en de *N* bij de tweede afname) een significante voorspeller van de hertest-correlatie is, wanneer het gaat om papieren evaluaties onderling. Gaat het om de hertest-correlatie tussen papieren afname enerzijds en digitale afname anderzijds, dan is vooral het aantal respondenten bij de digitale afname (dit is hier immers steeds de tweede) van invloed op de hoogte van de hertest-correlatie, gevolgd door het responsverschil tussen papieren en digitale afname (zie tabel 2).

In model 1 fungeert het aantal respondenten bij de tweede afname als voorspeller, in model 2 wordt het responsverschil tussen eerste en tweede afname aan de voorspelling toegevoegd. Dit verhoogt de *R* van 0,45 tot 0,58 ('Adjusted multiple *R*' respectievelijk 0,17 en 0,29). De gestandaardiseerde *bèta*-coëfficiënten van achtereenvolgens de *N* bij de tweede afname en het responsverschil zijn respectievelijk 0,411 en -0,374. Bij dit alles fungeert een aantal van ongeveer vijftien respondenten als een soort ondergrens. Althans: wanneer er minder dan vijftien respondenten zijn, bedraagt de hertest-correlatie 0,50; bij een aantal groter dan vijftien is die 0,69. Die laatste coëfficiënt wijkt amper af van de eerder genoemde waarde van 0,73 bij papieren afnames onderling.

Dit artikel van Boom Lemma Tijdschriften is gemaakt voor Vrije Universiteit 180209

Tabel 2: Stapsgewijze regressievergelijking met de hertest-correlatie als afhankelijke variabele, en de N bij de tweede afname en het responsverschil als voorspellers

Model		Som of squares	df	Mean square	F	P
1	Regression	0,313	1	0,313	6,928	0,014
	Residual	1,265	28	0,045		
	Total	1,578	29			
2	Regression	0,532	2	0,266	6,868	0,004
	Residual	1,046	27	0,039		
	Total	1,578	29			

Conclusie en discussie

De gerapporteerde bevindingen sluiten goed aan op het door Nulty (2007) verstrekte overzicht van het verband tussen groepsgrootte, steekproef en responspercentage, zoals weergegeven in tabel 3.

Tabel 3: Verhouding tussen aantal cursisten, vereist aantal respondenten en vereist responspercentage (overgenomen uit Nulty, 2007)

Aantal cursisten	Vereist aantal respondenten	Vereiste respons (%)
10	7	75
20	12	58
30	14	48
40	16	40
50	17	35
60	18	31
70	19	28
80	20	25
90	21	23
100	21	21

De tabel geldt voor 'liberale condities' – dat wil zeggen een 'sampling error' van 10% en een 80% betrouwbaarheidsniveau. Onder die randvoorwaarden dient men minimaal 16 respondenten te hebben, wil men redelijkerwijze uitspraken kunnen doen over de hele groep van bijvoorbeeld 40 cursisten. Dit komt overeen met een responspercentage van 40%. Het is ook mogelijk om andersom te redeneren, en uit te gaan van de bij digitale evaluaties vrij gebruikelijke responspercentages van hoogstens 35 à 40%. Uit de tabel valt op te maken dat de totale groep dan minstens 40 tot 50 cursisten moet bedragen. Uit de tabel wordt overigens óók duidelijk dat bij groepen die de 100 gaan benaderen (zeer) lage responspercenta-

Dit artikel van Boom Lemma Tijdschriften is gemaakt voor Vrije Universiteit 180209

ges nauwelijks een probleem vormen om toch uitspraken te kunnen doen over de groep als geheel.

Het lijkt er dus op dat digitale vragenlijsten gemiddeld zeker niet zonder meer tot hogere of lagere uitkomsten leiden, maar wel dat de hertest-betrouwbaarheid in bepaalde omstandigheden minder hoog is. Waar de correlatie tussen herhaalafnames van papier versus papier gemiddeld 0,73 is (in overeenstemming met waarden die bekend zijn uit de literatuur), is die bij papier versus digitaal niet hoger dan 0,60. Een lagere respons is op zichzelf dus niet zo erg, maar daar zijn in absolute zin wel grenzen aan omdat de kans op een niet-representatief sample dan sterk toeneemt. Voor alle duidelijkheid: dit geldt zowel voor papieren als voor digitale evaluaties! Bij een responspercentage van 35 tot 40% en zestien tot zeventien respondenten moet de groepsgrootte toch wel ongeveer 40 tot 50 cursisten bedragen. Is de groep kleiner dan 40, dan wordt het vereiste responspercentage in de praktijk vrijwel altijd onrealistisch hoog om toch maar tot het gewenste aantal respondenten te komen. Het gevolg daarvan is dat, wil men in dergelijke gevallen toch digitaal evalueren, over de betrouwbaarheid van de uitslag in toenemende mate onzekerheid gaat bestaan. Gezien de in de inleiding genoemde doelen van onderwijsevaluatie – onderwijsverbetering en het gebruik bij beslissingen over aanstelling, bevordering of contractverlenging – is dat problematisch.

Een eerste hieruit voortvloeiende aanbeveling spreekt voor zich, te weten al het mogelijke doen om de respons te verhogen. Daarbij moet primair worden gedacht aan de al eerder genoemde aansporing om voor studenten zichtbaar te maken dat hun mening ertoe doet. Zo kan men elk studiejaar de cursus beginnen met een korte uitleg aan studenten wat aan de cursusopzet is veranderd op grond van de evaluatie-uitkomsten van voorgaande jaren. Studenten stellen dit erg op prijs. Uiteraard dient de evaluatie gemakkelijk toegankelijk te zijn, en moeten studenten verzekerd zijn van de anonimiteit van hun reacties. Ten slotte: houd de vragenlijsten kort. Bij het Instituut voor Psychologie van de Erasmus Universiteit Rotterdam wordt het responsprobleem opgelost door het invullen van evaluatievragenlijsten verplicht te stellen: pas daarna mogen studenten aan de toets deelnemen (De Koning, Loyens, Smeets & Van der Molen, 2010). Het is een misschien wel voor de hand liggende maatregel, maar niet één waartoe docenten zelfstandig kunnen overgaan – op zijn minst zou het een facultair of instellingsbesluit moeten zijn. Het is ook niet zeker of een dergelijke verplichting bevorderlijk is voor de serieuze invulling van die vragenlijsten. Richardson noemt het uiteindelijk ook een ethische kwestie, mede omdat in veel beroepsorganisaties wordt benadrukt dat deelnemers aan een onderzoeksproject te allen tijde in staat moeten zijn zich terug te trekken. In elk geval geldt dat 'It will be important for institutions to clarify whether the collection of feedback is a formal part of the teaching-learning process or whether it is simply tantamount to institutional research' (Richardson 2005, p. 406).

Een tweede aanbeveling is om bij groepsgroottes van duidelijk minder dan 40 studenten wegens de te verwachten respons alleen digitaal te evalueren wanneer het

Dit artikel van Boom Lemma Tijdschriften is gemaakt voor Vrije Universiteit 180209

echt niet anders kan. Is die noodzaak er, dan gelden de hierboven genoemde suggesties in versterkte mate, en dienen de uitkomsten met voorzichtigheid te worden geïnterpreteerd. Dit vooral wanneer er voor de docent zelf ook daadwerkelijk iets van afhangt. Het onderzoek waarover hier wordt gerapporteerd, kent uiteraard zijn beperkingen. De onderzoekspopulatie wijkt sterk af van reguliere hbo- en wo-studenten. Niet alleen qua leeftijd en daarmee vermoedelijk ook de mate waarin ze ingevoerd zijn in de wereld van het internet, maar ook ten aanzien van het doel waarmee het onderwijs wordt gevolgd. HOVO-cursisten schrijven zich immers louter en alleen op grond van hun belangstelling in voor een bepaalde cursus. Aan het einde volgt ook geen tentamen. Het zou daarom te ver voeren om de bovenstaande bevindingen zonder meer te generaliseren naar het hele hoger onderwijs, en een en ander pretendeert ook niet meer dan een eerste aanzet tot verder onderzoek op dit terrein. Het onderwerp is er belangrijk genoeg voor. Uiteindelijk ontlenen onderwijsevaluaties hun betekenis toch vooral aan de mate waarin ze door docenten en onderwijsmanagers 'vertrouwd' kunnen worden.

Om te besluiten met een positieve boodschap: vooralsnog lijken er geen redenen te zijn om zich zorgen te maken over die betekenis. In vergelijking met evaluaties op papier laten digitale evaluaties geen wezenlijk verschillende uitkomsten zien. Het is waar dat de respons aanzienlijk lager ligt. Net als bij papieren evaluaties geldt echter dat dit geen kwaad kan zolang men meer dan ongeveer vijftien respondenten heeft, ook al is de groep waarover men uitspraken wil doen een veelvoud daarvan.

Referenties

Bothell, T.W. & Henderson, T. (2003). Do online ratings of instruction make $ense? *New Directions for Teaching and Learning*, 96, 69-79.

Carini, R.M., Hayek, J.C., Kuh, G.D., Kennedy, J.M. & Quimet, J.A. (2003). College student responses to web and paper surveys: does mode matter? *Research in Higher Education*, 44, 1-19.

Cook, C., Heath, F. & Thompson, R.L. (2000). A meta-analysis of response rates in web or internet-based surveys. *Educational & Psychological Measurement*, 60, 821-836.

Hollander, A.P. & Os, W. van (2010). *Rapportage Evaluatie Dienstverlening Onderwijscentrum VU*. Amsterdam: Vrije Universiteit.

Johnson, T.D. (2003). Online student ratings: Will students respond? *New Directions for Teaching and Learning*, 96, 49-59.

Koning, B. de, Loyens, S., Smeets, G. & Molen, H. van der (2010). De student ontcijferd. *HO Management oktober 2010*, 16-18.

Leung, D.Y.P. & Kember, D. (2005). Comparability of data gathered from evaluation questionnaires on paper and through the internet. *Research in Higher Education*, 46, 571-591.

Marsh, H.W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and utility. *Journal of Educational Psychology*, 76, 707-754.

Moss, J. & Hendry, G. (2002). Use of electronic surveys in course evaluation. *British Journal of Educational Technology*, 33, 583-592.

Dit artikel van Boom Lemma Tijdschriften is gemaakt voor Vrije Universiteit 180209

Nulty, D.N. (2007). The adequacy of response rates to online and paper surveys: What can be done? *Assessment & Evaluation in Higher Education*, 33, 301-314.
Os, W. van (1999). *Bruikbaarheid en effectiviteit van studentenoordelen over het onderwijs*. Amsterdam: Academisch proefschrift Vrije Universiteit.
Os, W. van (2010). Evaluatie in het hoger onderwijs: de kinderschoenen ruimschoots ontgroeid. In H. van Berkel, A. Bax & H. van Hout (red.), *Kennis delen en inspireren – toen, nu en in de toekomst* (p. 81-87). Groningen/Houten: Noordhoff Uitgevers.
Porter, S.R. & Whitcomb, M.E. (2003). Non-response in student surveys: The role of demographics, engagement and personality. *Research in Higher Education*, 46, 127-152.
Richardson, J.T.E. (2005). Instruments for obtaining student feedback: A review of the literature. *Assessment & Evaluation in Higher Education*, 30, 387-415.
Sax, L.J., Gilmartin, S.K. & Bryant, A.N. (2003). Assessing response rates and nonresponse bias in web and paper surveys. *Research in Higher Education*, 44, 409-432.

Bijlage: tekst vragenlijst

Inhoudelijke aspecten
1. Ik vond het een interessante cursus.
2. Het doel van de cursus was duidelijk.
3. De cursus heeft voldaan aan de verwachtingen.
4. De gebruikte werkvorm (hoorcollege/discussiecollege/werkcollege) was voor deze cursus geschikt.
5. Het voor deze cursus benodigde of voorgeschreven studiemateriaal was duidelijk en begrijpelijk (indien *oneens* graag toelichten op de achterzijde!).
6. Het niveau van de cursus was voor mij adequaat (indien *oneens* graag toelichten op de achterzijde!).
7. Gezien de tijd die ik eraan heb besteed, heb ik van deze cursus voldoende geleerd.
8. Totaaloordeel over de inhoud van de cursus (*van zeer slecht* tot *zeer goed*).

Presentatie van de docent
09. De voorbereiding van de docent(en) was goed.
10. De colleges waren in het algemeen goed opgebouwd.
11. De docent(en) was (waren) goed aanspreekbaar voor de cursisten.
12. Gestelde vragen werden bevredigend beantwoord.
13. Het tempo waarin de stof werd behandeld was voor mij adequaat (indien *oneens* graag toelichten op de achterzijde!).
14. Totaaloordeel over de presentatie van de docent (van *zeer slecht* tot *zeer goed*).

Algemeen
15. De cursus was goed georganiseerd.
16. De voorlichting over deze cursus kwam overeen met de feitelijke inhoud.
17. De voorzieningen (cursusruimte, onderwijsmiddelen, restauratieve voorzieningen en dergelijke) waren adequaat (indien *oneens* graag toelichten op de achterzijde!).
18. Wanneer over dit thema andere cursussen worden aangeboden, zou ik die zeker willen volgen.