**Over het nut en de waarde van het evalueren door studenten lopen de meningen nogal uiteen. Drie onderwijskundig adviseurs van de Universiteit Utrecht sloegen er negentig jaar wetenschappelijk onderzoek op na en nemen zeven veelbesproken feiten en mythes rond dit thema onder de loep.**

# Laat de beoordeling van de docent maar (niet) over aan de student

## De zin en onzin van cursusevaluaties

*Karin Scager, Johan van Strien & Ineke Lam*

**Universiteit Utrecht**

Hogeronderwijsinstellingen wereldwijd, ook in Nederland, maken al decennialang gebruik van standaardcursusevaluaties. Oorspronkelijk hadden zij vooral een feedbackfunctie, maar in de laatste jaren worden standaardevaluaties steeds meer summatief gebruikt, voor verantwoordingsdoeleinden en als indicator voor de kwaliteit van docenten (Hornstein, 2017). Het belang dat instellingen hechten aan de resultaten van cursusevaluaties is toegenomen, met als resultaat dat de discussie over de validiteit van studentevaluaties opnieuw opvlamt.
Onderscheiden studentevaluaties inderdaad de goede docenten van de slechte? En is het gerechtvaardigd om deze data te gebruiken voor aanstellings- en bevorderingsbesluiten? In discussies over dit soort vragen gaan de partijen vaak voorbij aan het uitvoerige onderzoek dat beschikbaar is over dit thema. De validiteit van cursusevaluaties is al ruim negentig jaar object van onderzoek (de eerste publicatie hierover kwam uit in 1927; Kulik, 2001). De uitkomsten bieden een waardevolle basis voor de discussie. Aan de hand hiervan bespreken we zeven veelgehoorde feiten en mythes over studentevaluaties.

### 1 *'Cursusevaluaties meten studenttevredenheid in plaats van onderwijskwaliteit'*

Onderzoek van Braga, Paccagnella en Pellizzari (2014) liet zien dat studenten hun docenten beoordelen op basis van de mate waarin zij de cursus leuk vinden, en niet zozeer op basis van de kwaliteit van de instructie. Uit een opzienbarende recente studie van Hessler en collega's (2018) bleek dat studenten een cursus significant hoger beoordeelden als zij tijdens de cursus chocoladekoekjes kregen. Braga en collega's (2014) ontdekten daarnaast dat het weer van invloed is op de beoordeling: evaluatie-uitkomsten waren op zonnige dagen hoger dan op regenachtige. Verder vonden zij een verband met het niveau van de studenten: de betere studenten gaven betrouwbaardere evaluaties. Zwakkere studenten vinden het niet fijn om hard te moeten werken, wat zijn weerslag kan hebben op de cursusevaluatie.
Dit alles hangt samen met de vraag wat studenten logischerwijs het best kunnen beoordelen. Studenten zijn in staat een betrouwbaar oordeel te geven over zaken als de

**De partijen gaan vaak voorbij aan het uitvoerige onderzoek over dit thema**

verstaanbaarheid en de beschikbaarheid van de docent. Thema's die verder reiken, zoals didactiek en de deskundigheid van de docent, zijn voor studenten een stuk lastiger te beoordelen (Hornstein, 2017).
Ondanks deze knelpunten en bias is het nog steeds aannemelijk dat de kwaliteit van lesgeven van invloed is op de tevredenheid van studenten: een helder en samenhangend college is waarschijnlijk aanzienlijk aangenamer dan een chaotisch, onduidelijk college (Stroebe, 2016).

2 *'Studenten leren meer van docenten die hoog scoren op cursusevaluaties'*
Hoewel onderzoekers eerder nog bescheiden bewijs vonden voor de stelling dat studenten meer leren van docenten met hoge studentevaluatiescores, laat een recente meta-analyse duidelijk zien dat er geen correlatie is tussen studentevaluaties en leeropbrengsten van studenten (Uttl, White, & Wong Gonzalez, 2017). Uttl en collega's corrigeerden in hun meta-analyse voor de kwaliteit van de beschikbare onderzoeken die in eerdere meta-analyses waren meegenomen en kwamen daardoor tot een nauwkeuriger analyse.
Overigens speelt ook de manier waarop je leeropbrengsten meet een rol: onderzoek laat zien dat hoe objectiever je het leren meet, des te lager de correlaties met studentevaluaties zijn (Clayson, 2009).

3 *'Vrouwelijke docenten krijgen minder positieve evaluaties'*
Onderzoek uit 2016 toont aan dat studenten vrouwelijke docenten over het algemeen lager beoordelen dan mannelijke, ook als er sprake is van vergelijkbaar niveau (Boring, Ottoboni, & Stark, 2016). Onderzoek van Boring (2017) laat zien dat hierbij ook sprake is van een interactie met het geslacht van de student: mannelijke studenten geven hogere evaluatiescores aan mannelijke docenten op aspecten als de kwaliteit van onderwijsmaterialen, voorbereiding en organisatie, en de bruikbaarheid van ontvangen feedback. Opmerkelijk is dat vrouwelijke studenten vrouwelijke docenten ook hoger beoordelen op de twee laatstgenoemde elementen, dus deze bias kan ook in omgekeerde volgorde optreden. Dat neemt niet weg dat studenten vrouwelijke docenten lager beoordelen op aspecten als de wijze van kennisoverdracht en de hoeveelheid kennis die zij bezitten. Bij elementen die samenhangen met mannelijke stereotypen, zoals klassenmanagement, krijgen mannelijke docenten van zowel mannelijke als vrouwelijke studenten een hogere beoordeling dan vrouwelijke docenten.

**Haar bevindingen lieten zien dat achtergrond een belangrijke rol speelt in de waardering**

4 *'Studenten geven lagere scores als ze de inhoud van de cursus niet leuk vinden'*
Docenten van minder populaire cursussen, zoals statistiek, hebben vaak de indruk dat het voor hen moeilijker is hoge scores op cursusevaluaties te halen. Hun hypothese is dat het gebrek aan interesse de waardering van de docent en de cursus beïnvloedt. Onderzoek van Van Os (1999) en Marsh (1987) toonde aan dat studenten in cursussen met impopulaire inhoud vooral de items over de cursusinhoud lager waardeerden. Recenter onderzoek van Feistauer & Richter (2017) resulteerde in een significante maar zwakke relatie tussen de interesse van studenten bij de start van de cursus en de docentkwaliteit.
Concluderend kunnen we stellen dat voorafgaande interesse van studenten in een onderwerp de validiteit van cursusevaluaties aantast, zij het in lichte mate. Een plausibele reden hiervoor is dat het lesgeven makkelijker is als studenten al geïnteresseerd zijn in het onderwerp.

5 *'Kenmerken van studenten beïnvloeden evaluatiescores'*
Verschillen tussen studenten zouden bepalend zijn voor hoe zij de cursus en de docenten evalueren. Clayson (2018) onderzocht de interbeoordelaarsbetrouwbaarheid bij cursusevaluaties en kwam tot de conclusie dat '*the agreement between students was shown to be no better than what would be expected by chance, indicating that students do not agree on what they are being asked to evaluate*'. Anderson (2013) analyseerde de evaluatiescores van meer dan 1600 docenten en kwam tot dezelfde conclusie. De studenten in Claysons studie vulden bovendien de cursusevaluatie in vóór de start van de cursus en in de laatste week (week 16). Merkwaardig genoeg waren de scores van studenten tussen week 0 en week 16 behoorlijk consistent, wat de vraag oproept wat studenten nu eigenlijk evalueerden.
Nasser-Abu Alhija (2017) nam in een andere onderzoekslijn een onlinevragenlijst af waarin ze vroeg naar achtergronden van studenten en hun voorkeuren voor cursuskwaliteit. Haar bevindingen lieten zien dat achtergrond een belangrijke rol speelt in hoe studenten onderwijs waarderen. Dit is vooral problematisch als voorkeuren van studenten niet overeenkomen met de onderwijsmethode van de docent (Kember & Wong, 2000). Als studenten bijvoorbeeld een voorkeur hebben voor passief leren en de docent activerende onderwijsvormen gebruikt, leidt dat tot lagere evaluatiescores.

6 *'Studentevaluaties zijn vooral een populariteitstest'*
Deze stelling verwijst naar het zogenoemde halo-effect: het

**Het idee: een aardige of charismatische docent scoort hoger op álle aspecten**

onvermogen om te onderscheiden tussen de verschillende en potentieel onafhankelijke aspecten in de evaluatie (Spooren, Brockx, & Mortelmans, 2013). Het idee is dat als studenten een docent aardig of charismatisch vinden, deze hogere scores krijgt op alle aspecten van de cursusevaluatie en dat de stijl of persoonlijkheid van de docent dus alle andere dimensies domineert.
De meeste studies naar deze potentiële bias laten sterke relaties zien tussen *likeability* en evaluatiescores (Clayson & Scheffet, 2006; Delucci, 2000; Gurung & Vespia, 2007; Shevlin et al., 2000; Wolbring & Riordan, 2016). Shevlin en collega's vonden bijvoorbeeld een substantieel effect van de 'charismafactor' (69 procent en 37 procent verklaarde variantie voor respectievelijk docentkwaliteit en cursuskwaliteit). Deze studies suggereren dat cursusevaluaties meer een maat zijn voor de populariteit van de docent dan voor de onderwijskwaliteit. De gevonden bias is waarschijnlijk wel wat overschat, want charisma of likeability hangen samen met sommige aspecten van docentkwaliteit, zoals de interpersoonlijke relatie met studenten of het enthousiasme van de docent.

7 *'Alleen bij een hoge respons zijn studentevaluaties informatief'*
Met de opkomst van digitale studentevaluaties neemt ook de discussie over de effecten van het responspercentage op de betrouwbaarheid van evaluaties toe. Bij digitale afname is de respons immers lager dan bij groepsgewijze evaluaties op papier (Stowell, Addison & Smith, 2012). Uit het onderzoek van Stowell en collega's bleek dat de evaluatiescores vergelijkbaar waren met afname op papier, ondanks de lagere respons bij digitale afname. Het enige verschil was dat de geschreven opmerkingen van studenten bij digitale afname gemiddeld drie à vier keer zo lang waren, waarbij de balans tussen positief, negatief en neutraal vergelijkbaar was met die bij afname op papier.
Ook onderzoek van Van Os & Van Beek (2011) liet geen substantieel andere uitkomsten zien tussen afname op papier en digitaal, ondanks de lagere respons. Het benodigde responspercentage neemt af naarmate er meer cursisten zijn: bij 100 cursisten kan een responspercentage van 21 al voldoende zijn om uitspraken te kunnen doen voor de groep als geheel, terwijl dat bij dertig cursisten een percentage van 48 is (Van Os & Van Beek, 2011).
Recent onderzoek van Goos & Salomons (2016) wijst op een selectiebias die een rol kan spelen bij digitale evaluaties: doordat tevreden of betrokken studenten sneller geneigd zijn te reageren, ligt de ware evaluatiescore waarschijnlijk lager dan de feitelijke evaluatiescore. Onderzoek van Adams & Umbach (2012) ondersteunt dit: zij vonden een negatieve samenhang tussen het behaalde cursuscijfer en de respons; studenten met goede cijfers waren vaker geneigd de evaluatie in te vullen.

## Problematisch maar waardevol

Uit bovenstaande onderzoeksresultaten kunnen we concluderen dat de validiteit van standaardcursusevaluaties problematisch is. De evaluaties meten studenttevredenheid, en niet de kwaliteit van de cursus en/of de docent, en al zeker niet als je vindt dat een hoge cursus-/docentkwaliteit moet leiden tot beter leren. Het is dan ook de vraag of het raadzaam is dit soort gegevens te gebruiken voor het beoordelen van docenten.

Ondanks de validiteitsproblemen blijft feedback van studenten waardevol. Studenten zijn tenslotte de doelgroep van het onderwijs; zij zijn het best gekwalificeerd om feedback te geven over aspecten als de moeilijkheidsgraad van een toets, de helderheid van een uitleg, of de mate waarin de docent hen aan het denken heeft gezet.
De manier waarop instellingen nu gebruikmaken van standaardevaluaties is onvoldoende bruikbaar voor feedback. Zoals we eerder al stelden, is de verantwoordingsfunctie in de loop der jaren belangrijker geworden dan de feedbackfunctie (Hornstein, 2017; Darwin, 2017). Verschillende ontwikkelingen gingen hiermee gepaard.
Ten eerste is de verantwoordelijkheid voor zorg voor de cursuskwaliteit meer bij het management komen te liggen dan bij de docenten zelf. Volgens Darwin (2017), die uitgebreid onderzoek deed naar de wijze waarop docenten omgaan met standaardcursusevaluaties, heeft de verschuiving van

**De manier waarop ze nu gebruikmaken van standaardevaluaties is onvoldoende bruikbaar**

**Studenten zijn niet louter consumenten die passief kennis tot zich nemen**

feedback- naar controle-instrument tot gevolg dat docenten de studentevaluaties niet meer serieus nemen, of er zelfs bang voor zijn.
Een tweede ontwikkeling is dat de cursusevaluaties steeds korter en algemener van aard zijn geworden en dus minder contextgebonden, omdat voor managementdoeleinden een snel overzicht in de cursuskwaliteit en vergelijkbaarheid tussen cursussen van belang zijn. Dit leidt tot een verminderde bruikbaarheid van de evaluatieformulieren als feedbackinstrument voor de docent.
Een derde ontwikkeling is de benadering van de student als 'klant' (Darwin, 2017; Sproule, 2000). De positie van student als klant doet geen recht aan het onderwijs, want voor goed onderwijs is juist de inzet van de student belangrijk (Von der Dunk, 2005). Studenten zijn niet louter consumenten die passief de kennis tot zich nemen die de docent hun voorschotelt. Ze zijn voor een groot deel verantwoordelijk voor hun eigen leren en dragen ook bij aan het leren en de motivatie van hun medestudenten, aan de groepsdynamiek en aan het plezier en enthousiasme van de docent. Onderwijs is een samenspel tussen docent en studenten.

## Verbeter standaardevaluaties

Ons advies is om standaardcursusevaluaties – als ze nog worden gebruikt – te verbeteren en uitsluitend te richten op zaken waarin studenten nuttige inzichten kunnen verschaffen. Het gaat dan niet om oordelen over de kwaliteit. De veel gebruikte overallvraag waarin de student een cijfer geeft voor de hele cursus, zorgt volgens docenten die wij spraken voor onveiligheid en stigmatisering. Nu we hebben geconstateerd dat de waarde van die oordelen heel gering is, zou het weglaten ervan geen groot verlies aan informatie geven. Daarnaast zou de docent de verantwoordelijkheid voor de kwaliteitszorg van het onderwijs kunnen terugnemen door studenten naar feedback te vragen en met hen in dialoog te gaan. Vragen zouden kunnen gaan over de uitwerking van (nieuwe) methoden, over wat de student geleerd heeft en welke aspecten het leren bevorderden of juist in de weg stonden.

Tot slot enkele aanbevelingen:

- Gebruik de uitkomsten van studentevaluaties niet voor verantwoordingsdoeleinden, maar voor feedbackdoeleinden, bij voorkeur aangevuld met databronnen als observaties en focusgroepgesprekken.
- Stimuleer docenten om tussentijds te evalueren of feedback te vragen, zodat ze de feedback van studenten kunnen meenemen in de rest van de cursus.
- Evalueer niet elk jaar alles, maar vooral nieuwe cursussen, cursussen waarin grote wijzigingen zijn doorgevoerd en cursussen die in een voorgaand jaar laag zijn beoordeeld.
- Benader de student niet als 'klant' maar als betrokkene of *stakeholder* met wie je als docent in gesprek gaat over het verloop van de cursus en over mogelijke verbeteringen.
- Het streven naar een voldoende en evenwichtige respons blijft een aandachtspunt bij digitaal evalueren. Klassikale afname en het benadrukken van wat er met de resultaten wordt gedaan, kunnen de respons verhogen.

Dit alles vraagt om een andere kijk op evalueren: niet de controlefunctie komt centraal te staan, maar het verkrijgen van feedback op de eigen onderwijspraktijk. Die feedback kunnen docenten ook op andere wijzen verkrijgen, bijvoorbeeld door tussentijdse, formatieve evaluaties met studenten of door collegiale consultatie.

***Karin Scager, Johan van Strien & Ineke Lam***
*zijn als onderwijskundig adviseur verbonden aan de afdeling Onderwijsadvies & Training van de Universiteit Utrecht*

## Literatuur

Adams, M.J. & Umbach, P.D. (2012). Nonresponse and online student evaluations of teaching: Understanding the influence of salience, fatigue, and academic environments. *Research in Higher Education, 53*, 576-591.
Alhija, F.N.-A. (2016). Teaching in higher education: Good teaching through students' lens, *Studies in Educational Evaluation, 54*, 4-12.
Anderson, J.A. 2013. 'Student Feedback Measures: Meta-analysis. A Report to the Academic Senate.' University of Utah. http://fyi.utah.edu/wp-content/uploads/2013/03/Student-Feedback-Measures-final.pdf.
Boring, A. (2017). Gender biases in student evaluations of teaching. *Journal of Public Economics, 145*, 27-41.
Boring, A., Ottoboni, K., & Stark, P.B. (2016). Student evaluations of teaching (mostly) do not measure teaching effectiveness. Science Open Research.
Braga, M., Paccagnella, M., & Pellizzari, M. (2014). Evaluating students evaluations of professors. *Economics of Education Review, 41*, 71-88.
Clayson, D.E. (2018). Student evaluation of teaching and matters of reliability. *Assessment & Evaluation in Higher Education, 43,* 666-68.
Clayson, D.E., & Sheffet, M.J. (2006). Personality and the student evaluation of teaching. *Journal of Marketing Education, 28,* 149-160. doi:10.1177/ 0273475306288402.
Darwin, S. (2017). What contemporary work are student ratings actually doing in higher education? *Studies in Educational Evaluation, 54*, 13-21.
Von der Dunk, T. (2005). Student is klant in opleiding. *de Volkskrant*, 28 april 2005.
Feistauer, D., & Richter, T. (2017) How reliable are students' evaluations of teaching quality? A variance components approach. *Assessment & Evaluation in Higher Education,*

*42*(8), 1263-1279.

Goos, M. & Salomons, A. (2017). Measuring teaching quality in higher education: Assessing selection bias in course evaluations. *Research in Higher Education, 58*(4), pp 341-364.

Hessler, M., Pöpping, D.M., Hollstein, H., Ohlenburg, H., Arnemann, P.H., Massoth, C., Seidel, L.M., Zarbock, A., & Wenk, M. (2018). Availability of cookies during an academic course session affects evaluation of teaching. *Medical Education, 52*, 1064-1072.

Hornstein, H.A. (2017). Student evaluations of teaching are an inadequate assessment tool for evaluating faculty performance. *Cogent Education, 4*, 1304016.

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

Os, W. van (1999). Bruikbaarheid en effectiviteit van studentenoordelen over het onderwijs, Academisch Proefschrift, Vrije Universiteit van Amsterdam.

Os, W. van, & Beek, M. van (2011). De lage respons bij digitale onderwijsevaluaties: een overschat probleem? *Tijdschrift voor Hoger Onderwijs, 29*(2), 98-107.

Shevlin, M., Banyard, P., Davies, M., & Griffiths, M. (2000). The validity of student evaluation of teaching in higher education: Love me, love my lectures? *Assessment & Evaluation in Higher Education, 25(4)*, 397-405, DOI: 10.1080/713611436.

Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching: The state of the art, *Review of Educational Research, 20*, 1-45.

Sproule, R. (2000). Student evaluation of teaching: A methodological critique of conventional practices. *Education Policy Analysis Archives*, 8(2).

Stowell, J.R., Addison, W.E., & Smith, J.L. (2012). Comparison of online and classroom-based student evaluations of instruction. *Assessment & Evaluation in Higher Education, 37*(4), 465-473.

Uttl, B., White, C.A., & Wong-Gonzalez, D. (2017). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation, 54*, 22-42.

Wolbring, T. & Riordan, P. (2000). How beauty works. Theoretical mechanisms and two empirical applications on students' evaluation of teaching, *Social Science Research 5*, 253-272.