Availability of cookies during an academic course session affects evaluation of teaching

Michael Hessler,† Daniel M Pöpping,† Hanna Hollstein, Hendrik Ohlenburg, Philip H Arnemann, Christina Massoth, Laura M Seidel, Alexander Zarbock & Manuel Wenk

OBJECTIVES Results from end-of-course student evaluations of teaching (SETs) are taken seriously by faculties and form part of a decision base for the recruitment of academic staff, the distribution of funds and changes to curricula. However, there is some doubt as to whether these evaluation instruments accurately measure the quality of course content, teaching and knowledge transfer. We investigated whether the provision of chocolate cookies as a content-unrelated intervention influences SET results.

METHODS We performed a randomised controlled trial in the setting of a curricular emergency medicine course. Participants were 118 third-year medical students. Participants were randomly allocated into 20 groups, 10 of which had free access to 500 g of chocolate cookies during an emergency medicine course session (cookie group) and 10 of which did not (control group). All groups were taught by the same teachers. Educational content and course material were the same for both groups. After the course, all students were asked to complete a 38-question evaluation form.

RESULTS A total of 112 students completed the evaluation form. The cookie group evaluated teachers significantly better than the control group (113.4 ± 4.9 versus 109.2 ± 7.3; p = 0.001, effect size 0.68). Course material was considered better (10.1 ± 2.3 versus 8.4 ± 2.8; p = 0.001, effect size 0.66) and summation scores evaluating the course overall were significantly higher (224.5 ± 12.5 versus 217.2 ± 16.1; p = 0.008, effect size 0.51) in the cookie group.

CONCLUSIONS The provision of chocolate cookies had a significant effect on course evaluation. These findings question the validity of SETs and their use in making widespread decisions within a faculty.

Medical Education 2018: 52: 1064–1072
doi: 10.1111/medu.13627

Department of Anaesthesiology, Intensive Care and Pain Medicine, University Hospital of Münster, Münster, Germany

Correspondence: Manuel Wenk, Department of Anaesthesiology, Intensive Care and Pain Medicine, University Hospital of Münster, Albert Schweitzer Campus 1, A1, 48149 Münster, Germany. Tel: 00 49 251 83 47 255;
E-mail: manuelwenk@uni-muenster.de
†These authors contributed equally to this work.

INTRODUCTION

End-of-course feedback in the form of student evaluations of teaching (SETs) has become a standard tool for measuring the 'quality' of curricular high-grade education courses. The results of these evaluations often form the basis for far-reaching decisions by academic faculty staff, such as those involving changes to the curriculum, the promotion of teachers, the tenure of academic appointments, the distribution of funds and merit pay, and the choice of staff.[1,2]

Despite their widespread use and the influential role played by SETs in academia today, the meaningfulness of SETs is frequently questioned. It is not at all clear whether a positive evaluation of a course accurately reflects the quality of teaching.[3] For example, positive correlations between course grades received by students and SET scores were described many years ago.[4–7] These findings led to discussions concerning grading leniency on the part of the instructor as a means of striving for better evaluation results, and the potential for students to misuse their empowerment to reward teachers for apparently easier courses or – even worse – to punish teachers for providing more difficult course content or bad grades.[6,8]

These issues call into question the validity of SETs with reference to the degree to which SETs measure what they claim to measure, which is still subject to scientific discussion.[9] Since Messick's unified conceptualisation of validity,[10] various frameworks have been introduced to assess the validity of SETs.[11,12] Several studies have investigated the validity of student opinions and their relationships to possible bias, hence examining the influence of factors that are not necessarily directly related to teaching quality and belong to the discriminant validity and divergent validity of the SET.[9] For example, several trials have investigated the influence on teaching effectiveness of other factors, including course subject,[13] class size,[14,15] instructor attractiveness[16] and gender.[17] Although some authors consider the impact of these factors to be low, it remains unclear whether other factors besides core content and teaching methods affect SET results and, if so, to what extent.[15,18]

The aim of this study was to prospectively investigate, by means of a randomised controlled trial (RCT), the effect on SETs of providing chocolate cookies during an emergency medicine course session. The primary outcome of the study was the difference between the study groups in summation scores evaluating the course session overall.

METHODS

Ethical approval

The study was approved by the Ethics Committee of the University of Münster (ref. 2017-145-f-S) and registered with the German Clinical Trials Register (Deutsches Register Klinischer Studien [DRKS], registration no. DRKS00012353). Requirements for consent were waived by the Ethics Committee because prior knowledge of the study carried a substantial risk for disturbing the study results. Therefore, students were not aware that they were participating in a study.

Study design, setting and population

The study was designed as a single-centre RCT conducted in the medical faculty of the University of Münster, Germany between 15 and 30 May 2017, during an emergency medicine course for undergraduate medical students in their fifth semester of medical school training. The total cohort of the semester (118 students) was enrolled in the study. Students were taught in small groups with a maximum of six students per group as predetermined by the medical faculty. All students were randomly distributed into these 20 groups.

Over the duration of the course, students met on four occasions to discuss and learn the management of common cases encountered in emergency medicine under the tutelage of an emergency medicine-trained anaesthesiologist.

To minimise the variability in teaching content, this study was carried out during the first session of the course, during which the topic 'acute coronary syndrome' was discussed. The lesson consisted of a case-based discussion. The case and the learning content (causes, acute therapy, prevention) were set out in a course script that teachers were asked to follow. Teaching material was explicitly not made available to the students.

Selection of chocolate cookies

To meet a potentially wide range of tastes among students, a traditional 'drop cookie' type[19] was chosen to deliver chocolate to participants.

Course evaluation

A German version of a course evaluation questionnaire was used. The questionnaire was compiled from questionnaire modules available at the University of Münster for teaching evaluations. It was adapted to the specific course session by, for example, including the session topic into a question. The questionnaire included questions regarding the teacher, the course content, learning environment and material, and several items for a self-assessment by the student. It comprised a total of 38 items, of which 36 used seven-step Likert items. In the two remaining questions, students were asked to rate the teacher and the course session, respectively, using the German school grading system (ranging from 1 [excellent] to 6 [very poor/fail]). In addition, at the end of the survey, students were asked about their age, body weight and height (in order to calculate body mass index [BMI]) and were given the opportunity to make written comments at the end of the form. An English translation of the questionnaire is available in Appendix S1.

All evaluation forms were assessed by two of the authors (HH and MW), who were not involved in the course session and were blinded to group allocation.

Experimental protocol and randomisation

A total of 118 medical students in their fifth semester were randomly allocated to 20 groups. Two experienced lecturers, who had already taught the same course several times, were chosen to participate in the study and groups of students were randomly allocated to these two teachers. Ten groups (five for each teacher) were randomly chosen to receive the intervention (cookie group). The other 10 groups served as controls (control group). All course sessions took place in the afternoons between 14.00 hours and 16.00 hours. Figure 1 shows the experimental protocol and randomisation.

The cookie groups were provided with 500 g of cookies at the beginning of the session; this corresponded to an over-supply of cookies to avoid limiting students to taking one cookie only. The teachers were instructed to comment on the availability of the cookies only by saying: 'I've brought some chocolate biscuits.' Teachers were allowed to eat one chocolate cookie themselves to break any reservations about taking the first cookie. The control groups received no chocolate cookies.

If students brought cookies or other sweets for their group, the teachers were instructed to report this to the study group.

At the end of the session, all students were asked to fill in the anonymous end-of-session evaluation form. While the students answered the questions, the teachers were instructed to leave the room. The forms were collected in a ballot box. After all students had left the room, cookie consumption was determined by subtracting the weight of the remaining cookies from 500 g.

Main outcome measures

The primary outcome of the study was the difference between the study groups in overall summation scores of the course session evaluation. For this purpose, responses to Likert scale-based items on the questionnaire were summed to provide a summation score for all questions. Scores on the two questions that used the German school grading system (1 = excellent, 6 = very poor/fail) were inverted in the calculation of the summation score.

It was hypothesised that students would answer the questionnaire more benevolently if cookies were delivered during the course and therefore that summarising the Likert-type items would show higher sums in the cookie group.

In addition, summation scores for subgroups of questions (about, respectively, the teacher, teaching content, learning environment, course material and self-assessment of students) were calculated and analysed.

Statistical analysis

Statistical analysis was performed using IBM SPSS Statistics for Windows Version 24.0 (IBM Corp., Armonk, NY, USA). All data are presented as the mean ± standard deviation (SD). The quality of the data provided on completed standard course evaluation forms was deemed to be acceptable if responses to no more than two Likert-type items (corresponding to approximately 5% of the Likert scale-based item data per questionnaire) were missing; otherwise the evaluation form was excluded from analysis. Comparisons between groups were made using *t*-tests for independent samples. Asymptotic two-sided p-values of < 0.05 were

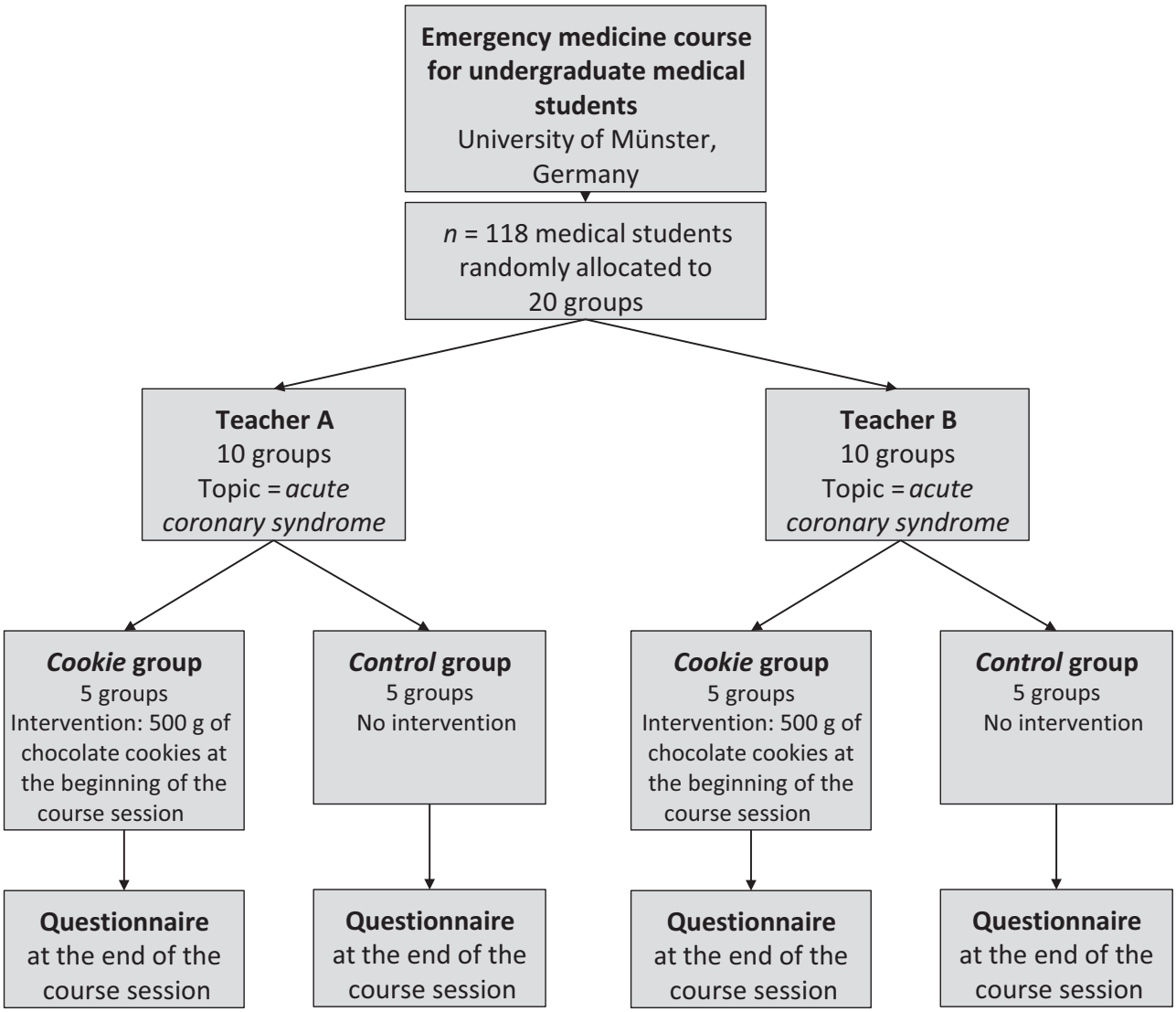

Figure 1 Experimental protocol and randomisation

considered to indicate differences of statistical significance. For comparisons of summation scores on the subgroups of questions, a Bonferroni-corrected threshold of $p < 0.01$ was considered to indicate statistical significance. As a measure of effect size, Cohen's *d* was calculated for the summation scores for all questions and for subgroups of questions (with 0.2 indicating a small effect, 0.5 a medium effect and 0.8 a large effect).[20] The reliability of the standard course evaluation form used was evaluated by analysing internal consistency based on the calculation of the Cronbach's alpha coefficient (Cronbach's α; desirable values of > 0.70 to 0.80).[21] To differentiate between the effects of cookies and teachers and to investigate the possible influences of age, BMI and gender (male or female) on the summation score for all questions, a multiple regression analysis was conducted, using the summation score for all questions as the dependent variable and student age, availability of cookies (yes or no), BMI, student gender and teacher gender (teacher A or teacher B) as independent variables. The normality of the summation scores for all questions was assessed by Q–Q (quantile–quantile) plots and the Shapiro–Wilk test ($p > 0.05$).

RESULTS

Differences between study groups

Of the 118 students enrolled in the study, a total of 112 students submitted the questionnaire (95%).

A mean ± SD of 0.2 ± 0.4 (0.5 ± 1.1%) Likert scale-based item responses were missing per questionnaire. No evaluation form was excluded because of missing data.

There were no differences between the cookie group and the control group with regard to sex (p = 1.000), age (p = 0.173), body weight (p = 0.424) or BMI (p = 0.895). Table 1 shows demographic data for the two groups.

Of the groups taught by teacher A, the control groups were slightly older than the cookie groups (22.0 ± 2.6 years versus 24.2 ± 4.3 years; p = 0.025). There were no further intergroup differences (cookie versus control group) between groups taught by teacher A and teacher B, respectively, in terms of sex (p = 0.629 and p = 0.599), body weight (p = 0.750 and p = 0.384) and BMI (p = 0.787 and p = 0.470) (Table S1).

Cookie consumption in the cookie group and students' comments

A mean ± SD of 365 ± 118 g of cookies were eaten by each of the 10 intervention groups. This corresponds to 68 ± 26 g of cookies consumed by each student or approximately 3.6 ± 1.4 cookies per student. An analysis of comments made by students showed 54% of comments were cookie-related in the cookie group (Table S2 presents all comments made by students in the cookie and control groups).

Comparison of summation scores between the cookie and control groups

Students in the cookie group evaluated the course session significantly better than students in the control group: significant differences were seen between the cookie and control groups in the summation score for all questions (p = 0.008) (Table 2). Comparisons of subgroups revealed significant differences in scores for questions related to teachers or course material (p = 0.001 and p = 0.001, respectively). There were no differences for subgroups relating to teaching content, learning environment and the self-assessment of students. Table 2 shows corresponding data and effect sizes.

Results of the multiple regression analysis

A multiple regression was performed to determine how much of the variation in the summation score for all questions was explained by student age, the availability of cookies, BMI, student gender and teacher gender. The multiple regression model predicted a statistically significant difference in the summation score for all questions ($F_{5,102}$ = 2.474, p < 0.037; overall model, R^2 = 10.7%, adjusted R^2 = 6.3%). Only the additional availability of cookies added statistical significance to the prediction. Regression coefficients, standard errors and 95% confidence intervals for unstandardised coefficients can be found in Table 3.

Table 1 Characteristics of participating students

		Group		
Variable	**All students**	**Cookie group**	**Control group**	**p-value***
n	112	56	56	–
Sex, *n* (%)				
Male	48 (43%)	24 (43%)	24 (43%)	1.000
Female	64 (57%)	32 (57%)	32 (57%)	1.000
Age, years, mean ± SD	23.0 ± 3.8	22.5 ± 3.9	23.5 ± 3.8	0.173
Body weight, kg, mean ± SD	70.6 ± 12.4	69.6 ± 11.4	71.6 ± 13.4	0.424
Body height, m, mean ± SD	1.8 ± 0.1	1.8 ± 0.1	1.8 ± 0.1	0.963
BMI, kg/m^2, mean ± SD	22.2 ± 3.6	22.2 ± 0.5	22.3 ± 4.4	0.895

* Level of significance: p < 0.05.
BMI = body mass index; SD = standard deviation.

Table 2 Comparison of summation scores in the cookie and control groups

	Control group (n = 56)		Cookie group (n = 56)			
	Summation score, mean ± SD	Cronbach's α	Summation score, mean ± SD	Cronbach's α	p-value*	Effect size (Cohen's d)
All questions	217.2 ± 16.1	0.89	224.5 ± 12.5	0.85	0.008	0.51
Subgroups of questions						
The teacher	109.2 ± 7.3	0.83	113.4 ± 4.9	0.71	0.001	0.68
Teaching content	36.2 ± 3.4	0.50	35.7 ± 3.4	0.44	0.387	0.15
Course material	8.4 ± 2.8	0.93	10.1 ± 2.3	0.90	0.001	0.66
Learning environment	11.6 ± 2.2	0.84	11.7 ± 2.3	0.73	0.900	0.04
Self-assessment	52.4 ± 6.0	0.81	53.5 ± 5.2	0.85	0.324	0.20

* Level of significance: $p < 0.05$. For comparisons of summation scores on the subgroups of questions, a Bonferroni-corrected threshold of $p < 0.01$ was considered to indicate statistical significance.
SD = standard deviation.

Table 3 Summary of the multiple regression analysis

Variable	B	SE_B	β	95.0% CI for B	p-Value*
Intercept	257.95	13.29	–	231.59–284.32	< 0.001
Availability of cookies	6.15	2.76	0.210	0.67–11.62	0.028
Teacher A or B	1.80	2.75	0.06	− 3.66 to 7.27	0.514
Gender	0.961	2.94	0.03	− 4.87 to 6.79	0.744
BMI	− 0.670	0.41	− 0.163	− 1.49 to 0.15	0.107
Age	− 0.460	0.37	− 0.121	− 1.19 to 0.27	0.212

* Level of significance: $p < 0.05$.
β = standardised coefficient; B = unstandardised regression coefficient; BMI = body mass index; CI = confidence interval; SE_B = standard error of the coefficient.

DISCUSSION

End-of-course evaluations have become a preferred tool with which to survey medical students for feedback regarding the quality of teaching and education. They allow for the systematic analysis of various parameters, providing a tool for ongoing quality assurance and improvement. The results of these course evaluations are taken very seriously by academic units and form part of a decision base used in the recruitment of academic staff, the distribution of funds from the faculty to units within it and changes to students' curricula.[1,2] However, there is some doubt as to whether these evaluation instruments really measure what they should: the quality of the core teaching and content, and the quality of the transfer of knowledge.

In our study we gave students access to chocolate cookies during a teaching session and compared their evaluations of the course session with the evaluations of a group of students without access to cookies. Our study amply demonstrates that students who had free access to chocolate cookies evaluated both the teacher and the course session significantly more highly than students who did not

have access to cookies, even though both the intellectual content of the session and the teacher were identical. A multiple regression analysis of our data shows that only the addition of cookies added significantly to the prediction of the summation score of evaluation results.

The validity of students' evaluations is always a topic of discussion amongst members of academic faculties and to date various factors that influence student perceptions of teaching content and that are unrelated to course content have been described. Felton et al. found that teacher 'sexiness' was strongly associated with good evaluation scores and were able to reproduce these results in a large sample size.[16,22] Uttl and Smibert described a strong relationship between course subject and evaluation results. They rated the impact of the evaluation as profound but flawed when SET results were linked to teaching performance, distribution of funds and decisions about personnel.[13] Centra and Gaubatz[17] described a same-gender preference, particularly among female students, and interpreted their results as reflecting different teaching styles. Furthermore, an impact of environmental factors on student satisfaction and subsequently on SET has been identified by several authors. Examples are the positive influence of small class sizes,[14,15] classroom attributes[23] and group harmony.[14]

However, some authors consider the impact of these factors, which are part of the discriminant validity and divergent validity of the SET,[9] to be low. For example, in an analysis of 371 131 student ratings on the Universal Student Ratings of Instruction instrument at a major Canadian university, Beran and Violato[18] found that various students and course characteristics explained only 7% of the total variance in SET scores. Spooren[24] reported a significant association between SET scores and the rank of the instructor: associate professors received lower SET scores than did full professors. The same study further showed that students' grades reduced 6.3% of the residual variance in SET scores and only 1.6% was explained by the examination on which the course grade was based (students who were required to retake examinations gave lower SETs).[24]

In the present study, the availability of chocolate cookies during an academic course session explained 6.3% of the variation in summation scores for all questions. In this context, the impact of cookies on SET results seems to be comparable with those of other variables previously found to have influence. Although it might be argued that 'educational meaningfulness' may control the effects of chocolate cookies, SET results are used in major decisions within faculties, such as those involving tenure of academic appointments, distribution of funds, merit pay and choice of staff. Therefore, having a few points less than one's colleagues may be a decisive factor in a teacher's career.

An explanation for the group differences observed in our trial as a result of the provision of cookies may be the cookies' chocolate content. Chocolate consumption has not only somatic effects, such as a positive impact on blood pressure[25] or analgesia,[26] but it also produces changes in the individual's emotional state, particularly by decreasing anxiety.[27] In an RCT, Macht and Mueller described an immediate improvement in mood after the consumption of palatable chocolate.[28] These effects are not just a matter of food intake because a direct comparison of the consumption of chocolate and that of apples showed that chocolate was superior.[29] We therefore assume that the chocolate cookies in our study positively influenced student mood and that resulted in an overall better evaluation of the session.

This study has several limitations. Firstly, the teachers were aware of which groups of students received cookies and had access to cookies themselves. As a result of this, they may have amended their teaching styles to a potentially unnoticed but influential extent. A further limitation is the fact that we have no information on how close the session was to meal and break times for individual students. These factors may also have affected students' and teachers' abilities and their assessments of the sessions. Food intake has even been shown to alter judges' parole and sentencing decisions.[30] Therefore, such 'ego depletion' may potentially change an instructor's performance between sessions.

CONCLUSIONS

The evaluation of teaching illustrates medical students' opinions about the quality and effectiveness of the teaching of curricular courses. However, our results demonstrate that such evaluations should not be used blindly as tools in quality assurance assessments because the opinions expressed within the evaluations depend to an unknown extent on various situational and emotional factors. Obviously, a simple intervention

such as the provision of chocolate cookies resulted in a significant distortion of evaluation results. These are not satisfying findings and call into question the validity of SET results in the context of making far-reaching decisions within a faculty.

Whether this effect is mostly attributable to the cookies themselves or to the influence of the broader social variable of reciprocity cannot be answered.[31] Would we have found similar effects if we had offered the students unpalatable kale and celery, a monogrammed commemorative course T-shirt or a coffee mug? Reciprocity might induce demand effects and enhanced evaluations, but it may also increase motivation and commitment to learning the material. This needs to be investigated in future studies.

Contributors: MH contributed to the conception of the study design, analysis of all samples, data interpretation and the writing of the manuscript. DMP contributed to the conception of the study design and to data interpretation. HH contributed to the conception and design of the study, and to data acquisition and interpretation. HO, PHA and CM contributed to data interpretation. LMS contributed to the study design and acquisition of data. AZ contributed intellectual input to the analysis and interpretation of the data. MW contributed to the conception of the study design, interpretation of data and the writing of the manuscript. All authors contributed to the subsequent drafting and critical revision of the paper and approved the final manuscript for submission.
Acknowledgements: the authors thank Michael Paech, Department of Anaesthesia, King Edward Memorial Hospital for Women (Perth, WA, Australia), for editing the language used in the manuscript.
Funding: this study was supported by intramural funding from the Department of Anaesthesiology, Intensive Care and Pain Medicine, University Hospital of Münster.
Conflicts of interest: none.
Ethical approval: this study was approved by the Ethics Committee of the University of Münster (ref. 2017-145-f-S) and registered with the German Clinical Trials Register (Deutsches Register Klinischer Studien [DRKS], registration no. DRKS00012353).

REFERENCES

1 Benton SL, Cashin WE. Student ratings of instruction in college and university courses. In: Paulsen MB, Smart JC, eds. *Higher Education: Handbook of Theory and Research*. Dordrecht: Springer 2014;279–326.
2 Abrami PC, d'Apollonia S, Rosenfield S. The dimensionality of student ratings of instruction: what we know and what we do not*. In: Abrami PC, d'Apollonia S, Rosenfield S, eds. *The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Perspective*. Dordrecht: Springer 2007;385–456
3 Schiekirka S, Feufel MA, Herrmann-Lingen C, Raupach T. Evaluation in medical education: a topical review of target parameters, data collection tools and confounding factors. *Ger Med Sci* 2015;**13**:Doc15.
4 Howard GS, Maxwell SE. Correlation between student satisfaction and grades: a case of mistaken causation? *J Educ Psychol* 1980;**72** (6):810–20.
5 Krautmann AC, Sander W. Grades and student evaluations of teachers. *Econ Educ Rev* 1999;**18** (1):59–63.
6 Brockx B, Spooren P, Mortelmans D. Taking the grading leniency story to the edge. The influence of student, teacher, and course characteristics on student evaluations of teaching in higher education. *Educ Assess Eval Account* 2011;**23** (4):289–306.
7 Woloschuk W, Coderre S, Wright B, McLaughlin K. What factors affect students' overall ratings of a course? *Acad Med* 2011;**86** (5):640–3.
8 Vaillancourt T. Students aggress against professors in reaction to receiving poor grades: an effect moderated by student narcissism and self-esteem. *Aggress Behav* 2013;**39** (1):71–84.
9 Spooren P, Brockx B, Mortelmans D. On the validity of student evaluation of teaching. *Rev Educ Res* 2013;**83** (4):598–642.
10 Messick S. Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *Am Psychol* 1995;**50** (9):741–9.
11 Onwuegbuzie AJ, Daniel LG, Collins KMT. A meta-validation model for assessing the score-validity of student teaching evaluations. *Qual Quant* 2009;**43** (2):197–209.
12 Ory JC, Ryan K. How do student ratings measure up to a new validity framework? *New Dir Institutional Res* 2001;**2001** (109):27–44.
13 Uttl B, Smibert D. Student evaluations of teaching: teaching quantitative courses can be hazardous to one's career. *PeerJ* 2017;**5**:e3299.
14 Kilgour JM, Grundy L, Monrouxe LV. A rapid review of the factors affecting healthcare students' satisfaction with small-group, active learning methods. *Teach Learn Med* 2016;**28** (1):15–25.
15 Centra JA. Differences in responses to the student instructional report: is it bias? *Listening, Learning, Leading* 2009. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.432.760&rep=rep1&type=pdf. [Accessed 25 May 2018.]
16 Felton J, Mitchell J, Stinson M. Web-based student evaluations of professors: the relations between perceived quality, easiness and sexiness. *Assess Eval High Educ* 2004;**29** (1):91–108.
17 Centra JA, Gaubatz NB. Is there gender bias in student evaluations of teaching? *J Higher Educ* 2000; **71** (1):17.

18 Beran T, Violato C. Ratings of university teacher instruction: how much do student and course characteristics really matter? *Assess Eval High Educ* 2005;**30** (6):593–601.
19 Wikipedia. Cookie. https://en.wikipedia.org/wiki/Cookie. [Accessed 25 May 2018.]
20 Sawilowsky SS. New effect size rules of thumb. *J Mod Appl Stat Methods* 2009;**8** (2):597–9.
21 Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika* 1951;**16** (3):297–334.
22 Felton J, Koper PT, Mitchell J, Stinson M. Attractiveness, easiness and other issues: student evaluations of professors on Ratemyprofessors.com. *Assess Eval High Educ* 2008;**33** (1):45–61.
23 Yang Z, Becerik-Gerber B, Mino L. A study on student perceptions of higher education classrooms: impact of classroom attributes on student satisfaction and performance. *Build Environ* 2013;**70**:171–88.
24 Spooren P. On the credibility of the judge: a cross-classified multilevel analysis on students' evaluation of teaching. *Stud Educ Eval* 2010;**36** (4):121–31.
25 Ried K, Fakler P, Stocks NP. Effect of cocoa on blood pressure. *Cochrane Database Syst Rev* 2017;**8**:CD008893.
26 Foo H, Mason P. Analgesia accompanying food consumption requires ingestion of hedonic foods. *J Neurosci* 2009;**29** (41):13053–62.
27 Martin F-PJ, Antille N, Rezzi S, Kochhar S. Everyday eating experiences of chocolate and non-chocolate snacks impact postprandial anxiety, energy and emotional states. *Nutrients* 2012;**4** (12):554–67.
28 Macht M, Mueller J. Immediate effects of chocolate on experimentally induced mood states. *Appetite* 2007;**49** (3):667–74.
29 Macht M, Dettmer D. Everyday mood and emotions after eating a chocolate bar or an apple. *Appetite* 2006;**46** (3):332–6.
30 Danziger S, Levav J, Avnaim-Pesso L. Extraneous factors in judicial decisions. *Proc Natl Acad Sci U S A* 2011;**108** (17):6889–92.
31 Cialdini RB. *Influence: Science and Practice*. New York, NY: Allyn & Bacon 2001.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article:

Table S1. Summation scores on subgroups of questions on evaluations submitted by groups taught by teachers A and B, respectively.

Table S2. All comments made by students in the cookie and control groups

Appendix S1. English-language translation of the student questionnaire used in this study.

Received 23 November 2017; editorial comments to authors 26 March 2018; accepted for publication 18 April 2018