Studies in Educational Evaluation 59 (2018) 168–178

Contents lists available at ScienceDirect

Studies in Educational Evaluation

journal homepage: www.elsevier.com/locate/stueduc

Validity of students' evaluations of teaching: Biasing effects of likability and prior subject interest

Daniela Feistauer[a], Tobias Richter[b,*]

[a] University of Kassel, Department of Psychology, Holländische Str. 36–38, 34127 Kassel, Germany
[b] University of Würzburg, Department of Psychology IV, Röntgenring 10, 97070 Würzburg, Germany

ARTICLE INFO

Keywords:
Cross-classified multilevel analysis
Likability
Prior subject interest
Student evaluations of teaching
Variance components

ABSTRACT

This study examined the validity of students' evaluations of teaching as an instrument for measuring teaching quality by examining the effects of likability and prior subject interest as potential biasing effects, measured at the beginning of the course and at the time of evaluation. University students ($N = 260$) evaluated psychology courses in one semester at a German university with a standardized questionnaire, yielding 517 data points. Cross-classified multilevel analyses revealed fixed effects of likability at both times of measurement and fixed effects of prior subject interest measured at the beginning of the course. Likability seems to exert a substantial bias on student evaluations of teaching, albeit one that is overestimated when measured at the time of evaluation. In contrast, prior subject interest seems to introduce a weak bias. Considering that likability bears no conceptual relationship to teaching quality, these findings point to a compromised validity of students' evaluations of teaching.

Every administrator working with student evaluations has probably met at least one university teacher who doubted the validity of students' evaluations of teaching (SETs) as an instrument for measuring teaching quality (e.g., Greenwald, 1997). Validity in this context refers to the standard psychometric definition of validity according to which a test is valid to the extent that it measures what it claims to measure (Kelley, 1927). These teachers' doubts are as old as SETs, and there is extensive research on this topic (Barr, 1943; Marsh & Roche, 1997; Olivares, 2003; Spooren, Brockx, & Mortelmans, 2013; Stalnaker & Remmers, 1928; Staufenbiel, Seppelfricke, & Rickers, 2016). In this study, we examined two potential threats to the validity of SETs, namely the extent that students perceive their teachers as likeable and the extent of their subject interest prior to taking the course. Perceived likability and prior subject interest are conceptually unrelated to teaching quality and can thus be considered threats to the validity of measurements of this construct. Extending earlier research on the role of likability and prior subject interest in SETs, we used a design with two measuring times (at the beginning of each course and concurrent with the course evaluation) to disentangle the causality underlying these constructs. In particular, we were able to distinguish between effects of likability and prior subject interest on SETs and a potential effect of the course on likability and (self-reported) prior subject interest. Moreover, in contrast to previous studies, our design allowed to determine the unique contributions of likability and prior subject interest.

1. Likability and students' evaluations of teaching

Likability or similar constructs, such as physical attractiveness, rapport, and personality of a teacher have already been investigated with SETs (Ambady & Rosenthal, 1993; Clayson & Haley, 1990; Clayson & Sheffet, 2006; Delucchi, 2000; Faranda & Clarke, 2004; Frymier, 1994; Gruber et al., 2012; Gurung & Vespia, 2007; Marks, 2000; Wolbring & Riordan, 2016). Most of these studies showed such strong relationships between the studied predictor and SETs that some authors named SETs "happy sheets" (Earley & Porritt, 2014, p. 112), "likability scales" (Clayson & Haley, 1990, p. 13), or "popularity contests" (Dziuban & Moskal, 2011, p. 237; Uranowitz & Doyle, 1978, p. 16). These and other authors expressed their doubts of whether SETs are a valid indicator of teaching quality and have therefore advised administrators and teachers against their use.

In this study, we construe teacher's likability as a general positive attitude that students hold towards the teacher. The construct includes the facets of perceived similarity, credibility, attraction, compliments, and association (Frymier, 1994; Reysen, 2005). In general, empirical relationships between likability and SETs may be interpreted in two ways (Delucchi, 2000). One interpretation views likability as a bias variable. Delucchi (2000) reported a particularly strong effect on global ratings of teaching quality. Out of 10 predictors that explained 78% of the total variance, likability was the third strongest predictor after

[*] Corresponding author.
E-mail addresses: feistauer@uni-kassel.de (D. Feistauer), tobias.richter@uni-wuerzburg.de (T. Richter).

https://doi.org/10.1016/j.stueduc.2018.07.009
Received 22 January 2018; Received in revised form 20 June 2018; Accepted 25 July 2018
Available online 07 August 2018

teaching behaviour and the stated goals of the course. Likewise, Clayson and Sheffet (2006) demonstrated that 73% of variance in SETs was explained by personality and likability, leaving little room for variance that could possibly be explained by teaching quality. Considering the strong relationship between likability and SETs, Clayson and Haley (1990) proposed that SETs should be regarded as likability scales. In the same vein, Clayson (1999) argued that the long-term stability of teachers' evaluation results found by Marsh and Hocevar (1991) could be explained by the influence of likability, which is based presumably on stable personally traits rather than factors related to teaching quality.

A second interpretation (also suggested by Delucchi, 2000) views likability as a component of teaching quality. Based on this interpretation, the large proportion of shared variance between likability and SETs found, for example by Clayson and Sheffet (2006), would not be considered as a threat to the validity of SETs. The operationalization of likability used by Marks (2000) illustrates this point. Marks combined three items to form the factor liking/concern: (1) "I like the instructor as a person", (2) "The instructor seems to have equal concern for all students", and (3) "The instructor was actively helpful when students had difficulty." The latter two items may be regarded as indicators of teaching quality as part of the social dimension of SETs, because they depict actions of a teacher that arguably represent good teaching. For example, teachers whose instructions are experienced as motivating by the students (Frymier, 1994) might also be perceived as likable.

2. Prior subject interest and students' evaluations of teaching

Prior subject interest can be understood as the individual student's initial interest in the subject before attending the course. An item assessing prior subject interest is included in most standardized SETs (e.g., Spooren, Mortelmans, & Denekens, 2007; Stalnaker & Remmers, 1928; Staufenbiel, 2000), because researchers have shared the assumption that students who are initially more interested in the subject of a course are probably more motivated (Marsh, 1982) and therefore easier to teach (Skinner & Belmont, 1993) than students who are uninterested in the subject. The easier teaching probably results in a more fluent and engaging teaching experience that is rewarded with higher ratings in SETs. For these reasons, a consensus exists that prior subject interest needs to be assessed to allow for a proper interpretation of SET results.

Previous findings concerning the relationship between prior subject interest and SETs have been inconsistent. Some studies showed positive effects of prior subject interest on SETs (Barth, 2008; Dresel & Rindermann, 2011; Marsh, 1981, 1982, 2007; Staufenbiel et al., 2016), whereas other studies have found no support for a relationship between prior subject interest and SETs (e.g., Olivares, 2001). This inconsistency might be due to the different aspects of teaching quality that were assessed. For example, Marsh (1980) found a strong relationship with the general course rating but only a weak relationship with course organization. In contrast to these results, Feistauer and Richter (2018) reported a weak relationship with two similar dimensions, teacher performance and planning and presentation.

According to Marsh (1984, 2007), the interpretation of prior subject interest as a bias variable depends on the dimension of SETs that is affected by prior subject interest. For example, prior subject interest may facilitate effective teaching and, therefore, be related to some dimensions of teaching quality (e.g. Marsh mentions the dimension learning/value of the SEEQ as an example) but not to others. For the present study, this argument implies as personal disposition of individual students influences rightfully some parts of teaching quality (e.g. learning and value). Therefore, it should only have an effect on teacher performance but not on planning and presentation. Influencing planning and presentation can be considered a threat to the validity of SETs as a measure of teaching quality.

3. Measurement time of likability and prior subject interest

Likability and prior subject interest have typically been measured concurrently with SETs in the same questionnaire (e.g., Marsh, 1982; Staufenbiel et al., 2016). Thus, the measurement may have been affected by the teacher performance, implying that the causality underlying the relationships with SETs is unclear (Kenny, 1979; Marsh, 1984; Staufenbiel et al., 2016). When prior subject interest is assessed at the same time as SETs, the responses are retrospective. The problem with retrospective assessments, in general, is that they are vulnerable to biases such as the hindsight bias (Hawkins & Hastie, 1990) or recall biases (Ross, 1989).

To disentangle the causality underlying the relationships of likability and prior subject interest with SETs, we measured both variables twice, at the beginning of the course before it had started and towards the end of the course at the same time when the SETs were assessed. A handful of previous studies on SETs and potential bias variables have already followed a similar design (Clayson & Sheffet, 2006; Howard & Schmeck, 1979). Howard and Schmeck measured motivation, similar to prior subject interest, and found a significant correlation ($r = 0.61$) between pre-course motivation and retrospectively assessed pre-course motivation of single courses. In addition, Clayson and Sheffet assessed likability of the teacher several times, at the beginning of the course (Week 0), after one week (Week 1), after ten weeks (Week 10), and finally at the end of the course (Week 16). They found significant effects of likability (Week 1–16) on SETs. Unfortunately, the likability scores at Week 0 were not reported. Evidently, the likability ratings after week one might already be affected by teacher behaviour at Week 0 (i.e. the first session of the course). Moreover, even the likability of teachers measured at the beginning of a course could be influenced by students' familiarity with the teacher, especially when students have already attended courses taught by the same teacher. Consider, for example, four students who have known a teacher for ten minutes, for three hours, for six month, or for two years. Clearly, the information that their likability rating is based on will differ between these four students. The first student's rating of the teacher's likability will be based on a first impression (Friedman, Riggio, & Casella, 1988) that cannot be related to teaching quality. However, the other three students have a broader stock of experiences (including experience with courses given by the teacher) for judging the teacher's likability. Thus, familiarity is an important covariate that needs to be considered to determine the biasing effect of likability.

4. Rationale of the present study

In the present study, we used a standardized and multidimensional questionnaire utilized in German-speaking countries for SETs in higher education (FEVOR, Staufenbiel, 2000; Staufenbiel et al., 2016) and a likability questionnaire (Reysen, 2005) that we adapted to the teaching context. The FEVOR questionnaire is composed of two global ratings: (a) quality of the entire course and (b) teacher performance; and four different dimensions of teaching quality: (a) planning and presentation, (b) interaction with students, (c) interestingness and relevance, and (d) difficulty and complexity. We focused our analyses on the teacher performance item and the planning and presentation dimension.

Global ratings of teacher performance are a broad indicator of teaching quality found in most SETs, which might be particularly prone to biasing effects, such as likability and prior subject interest, because of its unclear definition and intuitive accessibility. In contrast, planning and presentation consists of several items that reflect single aspects of the organizational part of teaching quality (e.g., "The lecture is clearly structured"). The items clearly describe aspects of teaching quality that, in principle, fall into the teacher's sphere of influence. Therefore, the evaluations based on this scale should be less prone to biasing effects.

Likability was measured once at the beginning of the course and again toward the end of the course as an additional item to the FEVOR questionnaire. Prior subject interest was also assessed at the beginning of the course and in the FEVOR questionnaire. Our study is the first to investigate the unique contributions of each predictor at both times of measurement to disentangle the causality underlying their relationships with SETs.

Each course was evaluated by several students, each student took several courses, teachers usually taught several courses, and some courses were taught by several teachers. Thus, the data have an imperfect or crossed hierarchy. For this data structure, cross-classified multilevel analysis (i.e., mixed models with crossed random effects, Baayen, Davidson, & Bates, 2008) was the method of choice. We included random effects (random intercepts) of all three possible sources of variance: teacher, course, and student (Feistauer & Richter, 2017). Additionally, we ran separate analyses for lectures and seminars because of the didactical and organizational differences between the two course formats (Staufenbiel et al., 2016).

Our analyses focused on four research questions. First, we examined the association between our two dimensions of SETs, teacher performance and planning and presentation, and the likability that individual students attribute to a teacher (Research Question 1). If a relationship were to occur only between teacher performance and likability but not between planning and presentation and likability, this pattern would support the argument that likability conceptually overlaps with certain aspects of teaching quality. However, if a relationship between planning and presentation and likability were also to occur, the result would provide evidence for a biasing effect of likability. The interpretation as biasing effect would receive additional support by a decrease in the teacher variance component compared to a null model after inclusion of likability into the model. Likability should not lead to a decrease in the teacher variance component of planning and presentation, because it is conceptually unrelated to this aspect of teaching quality and beyond the teacher's sphere of influence.

Second, we were interested in the strength of the prior subject interest effect on teacher performance and planning and presentation (Research Question 2). Significant effects were interpreted by examining changes in the variance component teacher, course, and student caused by including prior subject interest as predictor in the model. Again, strong relationships of prior subject interest with the global rating of teacher performance and the planning and presentation ratings would indicate a biasing effect of prior subject interest.

Third, we looked at the measurement time of likability and prior subject interest as a possible biasing effect (Research Question 3). A possible outcome is that likability and prior subject interest measured at the time of evaluation show significant effects on SETs but no effect when measured at the beginning of the course. In this scenario, likability and prior subject interest could not be classified clearly as biasing effects, because they could be influenced by events during the course. Another possible outcome is that likability and prior subject interest measured at the beginning of the course show significant effects on SETs. This outcome would be strong evidence for a biasing effect of these variables, which could be interpreted as a threat to the validity of SETs.

Fourth, considering that likability and prior subject interest measured at the beginning of the course might compete for explained variance in SETs, we investigated the unique contribution of one predictor in the context of the other predictor (Research Question 4).

5. Method

5.1. Sample

This study analysed a dataset of 517 student evaluations (questionnaire data) of all seminars and lectures in psychology held in the summer semester of 2017 at the University of Kassel, Germany. From a total of 26 teachers (14 females), 8 taught 11 lectures and 23 taught 36 seminars (5 teachers taught lectures as well as seminars). The sample of teachers included 11 doctoral students holding a position as researcher and lecturer (43%), 6 assistant professors or post-doctoral lecturers (23%), and 9 professors (34%). The evaluations were rated by 260 students (81% female) who participated in the psychology courses. Participation in the study was voluntary. Although the evaluations were anonymous, students who completed evaluations of multiple courses were coded with the same ID. Of these students, 52 evaluated two or more lectures (*Range* = 1–5) and 53 students evaluated two or more seminars (*Range* = 1–7). The sample included courses such as statistics, educational, cognitive, and clinical psychology.

5.2. Procedure

The evaluations were completed by the students in the last third of the semester (in the second half of June). Only students present in the course participated, which renders the sample a convenience sample. They were given 5–10 min of the course time to complete the online questionnaires. In addition to providing evaluations, students rated at the beginning of the course (within the first 10 min of the first session in the semester) their prior subject interest, how much they liked their teachers, and the familiarity with their teachers from previous courses. All data were collected with the online survey program Unipark, and the first author controlled the accuracy of the data.

5.3. Measures

The study analysed data from a standardized questionnaire used in Germany for the evaluation of university courses (FEVOR, Staufenbiel, 2000; Staufenbiel et al., 2016). Different versions of the questionnaire exist, depending on the course type. The questionnaire has 31 items for lectures and 34 items for seminars. Responses were provided on a Likert scale ranging from 1 (*strongly disagree*) to 5 (*strongly agree*) and "not applicable" as an additional response option. The two versions contain 26 identical items. Eight additional items in the seminar questionnaire refer to the quality of presentations held by students, and four items in the questionnaire for lectures refer to the teacher's presentation style. Students provided an individual alphanumeric code for relating multiple questionnaires completed by the same student, which could not be linked to the students, thus protecting their anonymity. The questionnaire items comprise four psychometrically distinct scales. In this study, we focused on the teacher performance and the planning and presentation scores.

5.4. Criterion variables

5.4.1. Teacher performance

Students rated the teacher's overall performance. Ratings were provided according to the German grading system that ranges from 1 (*very good*) to 5 (*poor*; lectures: $M = 1.87$, $SD = 0.76$; seminars: $M = 1.98$, $SD = 0.99$).

5.4.2. Planning and presentation

The scale assesses the extent to which students perceive a course to be well prepared and structured and the extent to which the contents are presented in a meaningful way. It contains items such as "The seminar provides a good overview of the subject area" and "The lecture is clearly structured." The scale consists of five items in lectures ($M = 4.16$, $SD = 0.63$, Cronbach's $\alpha = 0.85$) and eight items in seminars ($M = 4.11$, $SD = 0.82$, Cronbach's $\alpha = 0.85$).

5.5. Predictor variables

5.5.1. Likability

Students rated the teacher's likability with the item "How likable do you find the teacher?" Ratings ranged from 1 (*not likable at all*) to 5 (*very likable*). The variable was measured at the beginning of the course

Table 1
Intercorrelations Between Predictor Variables for Lectures.

	M	*SD*	Likability T1	Likability T2	Scale Likability	Prior subject interest T1
Likability T1	3.97	0.46				
Likability T2	4.10	0.44	+0.58			
Likability scale	3.89	0.37	+0.98***	0.46	(0.92)	
Prior subject interest T1	3.60	0.49	−0.15	< 0.01	−0.07	
Prior subject interest T2	3.41	0.51	−0.26	0.10	−0.23	0.87***

Note. Correlations based on group means of 11 lectures. Likability T1/T2: one-item measure of likability at the beginning of the course (T1) or at the time of the evaluation (T2). Likability Scale: Scale by Reysen (2005), assessed at T1 (Cronbach's α shown in brackets). Prior subject interest T1/T2: Prior subject interest assessed at the beginning of the course (T1) or at the time of evaluation (T2).

*** $p < 0.001$ (two-tailed).

Table 2
Intercorrelations Between Predictor Variables for Seminars.

	M	*SD*	Likability T1	Likability T2	Scale Likability	Prior subject interest T1
Likability T1	3.93	0.55				
Likability T2	3.93	0.69	+0.55***			
Likability scale	3.83	0.38	+0.89***	+0.45*	(0.89)	
Prior subject interest T1	3.63	0.58	−0.01	+0.21	0.10	
Prior subject interest T2	3.72	0.76	+0.02	−0.09	0.09	0.72

Note. Correlations based on group means of 36 seminars. Likability T1/T2: one-item measure of likability at the beginning of the course (T1) or at the time of the evaluation (T2). Likability scale: Scale by Reysen (2005), assessed at T1 (Cronbach's α shown in brackets). Prior subject interest T1/T2: Prior subject interest assessed at the beginning of the course (T1) or at the time of evaluation (T2).

* $p < 0.05$.

*** $p < 0.001$ (two-tailed).

(Likability T1) and at the time of evaluation (Likability T2). Descriptive statistics and intercorrelations can be found in Table 1 for lectures and in Table 2 for seminars. On the level of courses, Likability T1 and Likability T2 correlated 0.58 in lectures and 0.55 in seminars. In 270 of all 517 questionnaires (52.2%), students' likability ratings did not change from T1 to T2. In 90 questionnaires (17.4%), students rated the teacher at T2 by one point more likable, and in 108 questionnaires (20.9%), they rated the teacher by one point less likable than at T1. Only in 32 questionnaires (6.2%), likability decreased by more than one point and in 17 questionnaires (3.3%), likability increased by more than one point. To obtain an estimate of the reliability of the single likability item, we asked students at the beginning of the course to complete the likability scale by Reysen (2005), which we adapted to the teaching context. The scale reached internal consistencies (Cronbach's α) of 0.92 in lectures (Table 1) and 0.89 in seminars (Table 2) and is provided in Appendix. The single item and the scale correlated to 0.98 in lectures (Table 1) and 0.89 in seminars (Table 2). Because of these high correlations, we used the single likability item in all analyses.

Familiarity with the teacher prior to the course: To account for a possible influence of the students' familiarity with teachers on likability we assessed familiarity as covariate with the item: "Did you know the teacher already before this course?" Possible answers were *Yes – I already attended one of his/her courses*, *Yes - I have another course with him/her this semester*, *Yes – I know him/her from another context outside of courses*, or *No*. As the focus of this item is on previous courses, responses were dichotomized between the first answer and the latter three answers. In 157 questionnaires (30.3%) students stated that they already attended one of the teacher's courses before.

5.5.2. *Prior subject interest*

Students rated their prior subject interest with the item "What is (was) your level of interest in the course subject (before the course began)?" Ratings ranged from 1 (*very low*) to 5 (*very high*). This item was measured at the beginning of the course (Prior subject interest T1) and at the time of evaluation (Prior subject interest T2). Descriptive statistics and intercorrelations can be found in Table 1 for lectures and in Table 2 for seminars. Both items correlated (r = 0.87) in lectures and (r = 0.72) in seminars. In 271 of all 517 questionnaires (52.4%), students' subject interest did not change over time. In 79 questionnaires (15.3%), subject interest increased by one point, and in 136 questionnaires (26.3%) subject interest decreased by one point from T1 to T2. Only in 16 questionnaires (3.1%) the subject interest decreased by more than one point, and in 15 questionnaires (2.9%) subject interest increased by more than one point. Likability and prior subject interest at the beginning of the course showed a significant but weak correlation of r = 0.14.

6. Results

Analyses were performed with cross-classified multilevel models (Baayen et al., 2008) that allowed separating the teacher, course, and student variance components, which were included as random effects (random intercepts) in the analysis. Separate models were estimated for the two outcome variables teacher performance and the scale planning and presentation of the evaluation questionnaire by Staufenbiel (2000). The models were estimated with the statistical software R version 3.4.1 (R Core Team, 2017) and the full Maximum Likelihood estimation procedure included in the lmer function of the R-package lme4 (Bates, Mächler, Bolker, & Walker, 2015). The significance of each fixed effect was tested with the anova function of the R-package stats (R Core Team, 2017), which compares the fit of nested models. Data were analysed separately for lectures and seminars.

6.1. *Estimated models*

We estimated a sequence of models for both criterion variables. In the first step, we estimated a null model with no fixed effects but the student, teacher, and course variance components:

$$Y_{sct} = \theta_0 + h_{00s} + i_{00c} + j_{00t} + e_{sct} \quad (0)$$

Table 3
Estimates for the Cross-Classified Linear Mixed Effect Models for Teacher Performance in Lectures.

	Model 0 (Null Model)		Model 1a		Model 1b		Model 2a		Model 2b		Model 3	
Fixed effects	*Estimate* *(SE)*	*t*	*Estimate* *(SE)*	*t*	*Estimate* *(SE)*	*t*	*Estimate* *(SE)*	*t*	*Estimate* *(SE)*	*t*	*Estimate* *(SE)*	*t*
(Intercept)	1.870 (0.110)	17.05	1.846 (0.099)	18.62	1.799 (0.089)	20.30	1.885 (0.103)	18.36	1.871 (0.105)	17.78	1.859 (0.094)	19.69
Likability T1			−0.244*** (0.056)	−4.34							−0.228*** (0.057)	−4.03
Likability T2					−0.512*** (0.042)	−12.21						
Prior subject interest T1							−0.116* (0.052)	−2.25			−0.081 (0.051)	−1.59
Prior subject interest T2									−0.076 (0.053)	−1.45		
Random effects	*Variance*		*Variance*		*Variance*		*Variance*		*Variance*		*Variance*	
Residual	0.419		0.408		0.302		0.416		0.414		0.407	
Course (Intercept)	0.063		0.030		0.000		0.069		0.062		0.033	
Student (Intercept)	0.056		0.038		0.006		0.050		0.059		0.035	
Teacher (Intercept)	0.026		0.035		0.050		0.011		0.020		0.026	
Fit statistics												
-2LL	545.0		527.4+ + +		433.7+ + +		540.1+		543.0		525.0+ + +	
AIC	555.0		539.4		445.7		552.1		555.0		539.0	
BIC	572.7		560.6		466.9		573.3		576.2		563.7	

Note. Likability and prior subject interest were grand-mean centered before entered as predictors into the model. The number of observations that the variance components are based on: Residual: $N = 253$, Course: $n = 11$, Student: $n = 160$, Teacher: $n = 8$.
-2LL: -2 log-likelihood (deviance), AIC: Akaike information criterion, BIC: Bayesian information criterion.
Tests of fixed effects: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ (one-tailed).
Comparisons of models with the Null model (χ^2-difference tests with 1 *df* based on the deviances): + $p < 0.05$, + + $p < 0.01$, + + + $p < 0.001$ (two-tailed).

In Eq. (0), Y_{sct} represents the evaluation score provided by student *s* for course *c* given by teacher *t*. The intercept θ_0 represents the grand mean of this score across all students, courses, and teachers. The random effect h_{00s} captures the individual deviation of student *s* from θ_0. Likewise, the random effect i_{00c} represents the deviation of course *c* from θ_0, and the random effect j_{00t} the deviation of teacher *t* from θ_0. The variances τ_{s00}, τ_{c00}, and τ_{t00} of these deviations are assumed to be normally distributed with a mean of 0. Finally, the model includes the error term e_{sct}, which captures unsystematic error (such as measurement error) in the evaluation scores that remain after the students, courses, and teachers random effects have been taken into account. These unsystematic errors are also assumed to be normally distributed with mean 0 and variance σ^2 (Raudenbush & Bryk, 2006).

The model in Eq. (0) allowed estimating the student, teacher, and course variance components. Moreover, it served as the background for testing the effects of student background characteristics, which we entered as fixed effects. All predictors were centered at the grand mean. Models 1 and 2 were analysed to check for an impact of each bias variable in general, and their a and b variants let us compare the impact of each predictor's measurement time. Model 3 included both predictors at the beginning of the course.

We added the likability predictor at the beginning of the course $LT1_s$ with its slope β_1 in Model 1a:

$$Y_{sct} = \theta_0 + \beta_1 LT1_s + h_{00s} + i_{00c} + j_{00t} + e_{sct} \tag{1a}$$

For control purposes, we additionally estimated a model that included the familiarity covariate and its interaction with likability as fixed effects. In Model 1b, the likability predictor at the time of evaluation $LT2_{\ sct}$ with its slope β_2 was added:

$$Y_{sct} = \theta_0 + \beta_2 LT2_{sct} + h_{00s} + i_{00c} + j_{00t} + e_{sct} \tag{1b}$$

In Model 2a, the prior subject interest predictor at the beginning of the course $IT1_s$ with its slope β_3 was added:

$$Y_{sct} = \theta_0 + \beta_3 IT1_s + h_{00s} + i_{00c} + j_{00t} + e_{sct} \tag{2a}$$

In Model 2b, the prior subject interest predictor at the time of evaluation $IT2_{sct}$ with its slope β_4 was added:

$$Y_{sct} = \theta_0 + \beta_4 IT2_{sct} + h_{00s} + i_{00c} + j_{00t} + e_{sct} \tag{2b}$$

In Model 3, both predictors were added, likability at the beginning of the course $LT1_s$ with its slope β_1 and prior subject interest at the beginning of the course $IT1_s$ with its slope β_3.

$$Y_{sct} = \theta_0 + \beta_1 LT1_s + \beta_3 IT1_s + h_{00s} + i_{00c} + j_{00t} + e_{sct} \tag{3}$$

6.2. *Ratings of teacher performance in lectures*

Results for the six models with teacher performance in lectures as outcome variable are shown in Table 3. The overall mean (the intercept) of 1.87 estimated in Model 0 indicates that teacher performance in lectures was generally rated as good (in the German grading system, 1 represents "very good" and 2 "good").

Inclusion of the likability predictor at the beginning of the course in Model 1a led to a significantly improved model fit. The more likable that students rated the teacher at the beginning of the course the higher they evaluated teacher performance in lectures ($\beta_1 = -0.24$, $t(245.5) = -4.34$, $p < 0.001$). The addition of this predictor led to an explanation of 9.4% of the total variance and an increase of the teacher variance component by 34.6% compared to the null model. Fig. 1 shows the total variance and the differences in the variance components of the null model compared to Model 1a and 1b. Familiarity and the interaction between likability and familiarity had no effects ($\beta_5 = -0.12$, $t(67.7) = -0.93$, $p > 0.05$; $\beta_6 = 0.15$, $t(244.4) = 1.33$, $p > 0.05$).

The likability predictor assessed at the time of evaluation in Model 1b improved the model fit compared to the null model even more than the same predictor did when assessed at the beginning of the course. The more likable that students rated the teacher at the time of evaluation, the higher they evaluated teacher performance in lectures ($\beta_2 = -0.51$, $t(243.8) = -12.21$, $p < 0.001$). This predictor explained 36.5% of the total variance and led to an increase in the teacher

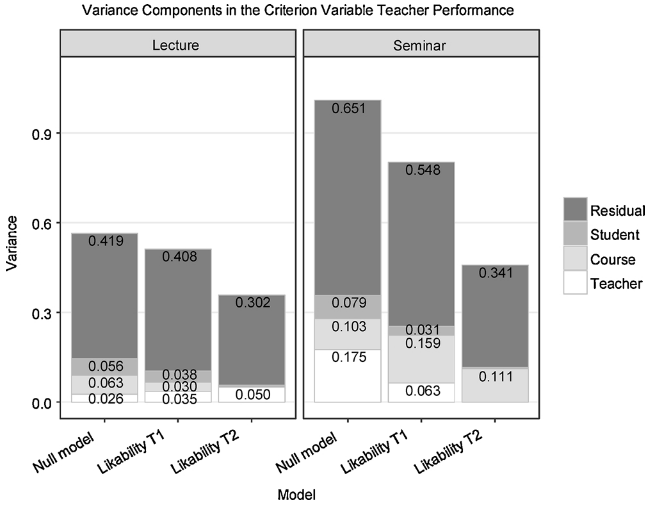

Fig. 1. Comparison of the different likability model variance components in stapled bar plots for the teacher performance criterion variable. Null model: Model without predictors, Likability T1: Model with the likability predictor at the beginning of the course, Likability T2: Model with the likability predictor at the time of evaluation.

variance component by 92.3% compared to the null model.

Inclusion of the prior subject interest predictor at the beginning of the course in Model 2a also led to a significant improvement of model fit compared to the null model. The more interesting that students rated the course at the beginning, the higher they evaluated teacher performance in lectures ($\beta_3 = -0.12$, $t(242) = -2.25$, $p < 0.05$). The addition of this predictor led to an explanation of 3.2% of the total variance, a decrease in the teacher variance component by 57.7%, a decrease in the student variance component by 10.7%, and an increase in the course variance component by 9.5% compared to the null model.

The prior subject interest predictor at the time of evaluation in Model 2b did not improve model fit compared to the null model. There was no significant relationship between prior subject interest at the time of evaluation and teacher performance in lectures ($\beta_4 = -0.076$, t $(246.9) = -1.45$, $p > 0.05$).

Including both likability and prior subject interest at the beginning of the course in Model 3 also led to an improved model fit compared to the null model. However, only likability ($\beta_1 = -0.23$, $t(238.7) = -4.03$, $p < 0.001$) but not prior interest ($\beta_3 = -0.081$, $t(238.2) = -1.59$, $p > 0.05$) had a significant effect on the global rating of teaching quality. The addition of both predictors led to an explanation of 11.2% of the total variance but no change in the teacher variance component compared to the null model.

6.2.1. *Summary and implications for the research questions*

The results provide support for a strong association between likability and the global rating of teaching quality whereas the association between prior subject interest and the global rating of teaching quality was rather weak (Research Questions 1 and 2). The association between prior subject interest and the rating of teaching quality was no longer significant when both likability and prior subject interest were included in the model (Research Question 4). The association of likability and the global rating of teaching quality was strong even when it was measured at the beginning of the course (Research Question 3). This pattern of effects suggests a potentially biasing effect of likability but not prior subject interest for SETs in lectures. However, stronger and convergent support for this conclusion would be provided by a similar pattern of results for the scale planning and presentation lectures. The results for this dimension of SETs are presented next.

6.3. *Planning and presentation in lectures*

Results for the six models with planning and presentation in lectures are shown in Table 4. The overall mean of 4.11 (maximum 5) estimated in Model 0 indicates that planning and presentation in lectures was rated as well prepared, structured, and presented in a meaningful way.

Table 4
Estimates for the Cross-Classified Linear Mixed Effect Models for Planning and Presentation in Lectures.

	Model 0 (Null Model)		Model 1a		Model 1b		Model 2a		Model 2b		Model 3	
Fixed effects	*Estimate* *(SE)*	*t*	*Estimate* *(SE)*	*t*	*Estimate* *(SE)*	*t*	*Estimate* *(SE)*	*t*	*Estimate* *(SE)*	*t*	*Estimate* *(SE)*	*t*
(Intercept)	4.111 (0.101)	40.76	4.145 (0.078)	53.35	4.146 (0.099)	41.92	4.094 (0.103)	39.57	4.107 (0.106)	38.76	4.125 (0.081)	50.91
Likability T1			0.213*** (0.047)	4.53							0.185*** (0.047)	3.92
Likability T2					0.350*** (0.039)	9.09						
Prior subject interest T1							0.156*** (0.043)	3.62			0.125** (0.043)	2.93
Prior subject interest T2									0.101* (0.044)	2.26		
Random effects	*Variance*		*Variance*		*Variance*		*Variance*		*Variance*		*Variance*	
Residual	0.258		0.257		0.196		0.252		0.257		0.253	
Course (Intercept)	0.019		0.002		0.000		0.031		0.016		0.011	
Student (Intercept)	0.088		0.070		0.066		0.073		0.080		0.059	
Teacher (Intercept)	0.049		0.032		0.065		0.044		0.060		0.029	
Fit statistics												
-2LL	458.4		440.5+ + +		387.9+ + +		446.0+ + +		453.5+		432.5+ + +	
AIC	468.4		452.5		399.9		458.0		465.5		446.5	
BIC	486.1		473.7		421.1		479.2		486.7		471.2	

Note. Likability and prior subject interest were grand-mean centered before entered as predictors into the model. The number of observations that the variance components are based on: Residual: $N = 253$, Course: $n = 11$, Student: $n = 160$, Teacher: $n = 8$.
-2LL: -2 log-likelihood (deviance), AIC: Akaike information criterion, BIC: Bayesian information criterion.
Tests of fixed effects: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ (one-tailed).
Comparisons of models with the Null model (χ^2-difference tests with 1 *df* based on the deviances): + $p < 0.05$, + + $p < 0.01$, + + + $p < 0.001$ (two-tailed).

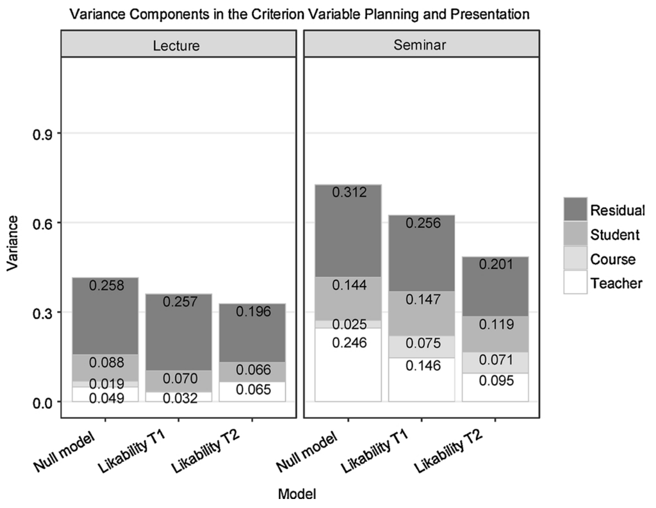

Fig. 2. Comparison of the different likability model variance components in stapled bar plots for the planning and presentation criterion variable. Null model: Model without predictors, Likability T1: Model with the likability predictor at the beginning of the course, Likability T2: Model with the likability predictor at the time of evaluation.

The likability predictor at the beginning of the course in Model 1a had a significant effect on planning and presentation. The more likable that students rated the teacher at the beginning of the course, the higher they evaluated planning and presentation in lectures ($\beta_1 = 0.21$, $t(224.2) = 4.53$, $p < 0.001$). This predictor explained 12.8% of the total variance and led to a decrease in the teacher variance component by 34.7% compared to the null model. Fig. 2 shows the total variance and the differences in the variance components of the null model compared to Model 1a and 1b. Familiarity and the interaction between likability and familiarity had no effects on planning and presentation in lectures ($\beta_5 = 0.15$, $t(36.1) = 1.47$, $p > 0.05$; $\beta_6 = -0.06$, $t(195.2) = -0.60$, $p > 0.05$).

The likability predictor at the time of evaluation in Model 1b led to a higher improvement of model fit compared to the null model than in Model 1a. Again, the more likable that students rated the teacher at the time of evaluation, the higher they evaluated planning and presentation in lectures ($\beta_2 = 0.35$, $t(250.8) = 9.09$, $p < 0.001$). This predictor explained 21% of the total variance and increased the teacher variance component by 32.7% compared to the null model.

The prior subject interest predictor assessed at the beginning of the course (Model 2a) also exerted a significant positive effect ($\beta_3 = 0.16$, $t(250.8) = 3.62$, $p < 0.001$) on the scale planning and presentation. This predictor explained 3.4% of the total variance and led to a decrease in the teacher variance components by 10.2%, a decrease in the student variance components by 17%, and an increase in the course variance components by 63.2% compared to the null model.

Including the prior subject interest predictor at the time of evaluation in Model 2b led to a significantly improved model fit compared to the null model. The more interesting that students rated the course at the time of evaluation, the higher they evaluated planning and presentation in lectures ($\beta_4 = 0.10$, $t(251.1) = 2.26$, $p < 0.05$). This predictor explained 0.2% of the total variance and led to an increase in the teacher variance components by 22.4%, a decrease in the student variance components by 9.1%, and a decrease in the course variance components by 15.8% compared to the null model.

Including both predictors assessed at the beginning of the course in Model 3 led to a significantly improved model fit compared to the null model. The more likable ($\beta_1 = 0.18$, $t(230) = 3.92$, $p < 0.001$) and more interesting ($\beta_3 = 0.13$, $t(239.2) = 2.93$, $p < 0.01$) that students rated the teacher and the course at the beginning of the course, the higher they evaluated planning and presentation in lectures. The addition of both predictors led to an explanation of 15% of the total variance and a decrease in the teacher variance component by 40.8% compared to the null model.

6.3.1. *Summary and implications for the research questions*

Similar to the results for the global rating for teaching quality, there was a strong association of likability and a weak (but this time significant) association with planning and presentation scale in lectures (Research Questions 1 and 2). Both effects remained significant when both predictors were included in the model, suggesting unique biasing effects of likability and prior subject interest (Research Question 4). These associations were again found also when likability and prior subject interest were measured at the beginning of the course (Research Question 3). In the next steps, we investigated whether similar effects occurred in seminars, again with global ratings of teacher performance and planning and presentation as criterion variables.

6.4. *Ratings of teacher performance in seminars*

Results for the six models with teacher performance in seminars are shown in Table 5. The overall mean of 2 estimated in Model 0 indicates that, on average, teacher performance in seminars was rated as good.

The likability predictor assessed at the beginning of the course in Model 1a had a significant effect on the global rating of teacher performance in seminars. The more likable teachers were rated at the beginning of the course, the more positive was the rating of their performance ($\beta_1 = -0.49$, $t(241.4) = -8.05$, $p < 0.001$). This predictor explained 20.5% of the total variance and led to a decrease in the teacher variance component by 64% compared to the null model. Fig. 1 shows the total variance and the differences in the variance components of the null model compared to Model 1a and 1b. Familiarity and the interaction of likability and familiarity had no effects on ratings of teacher performance in seminars ($\beta_5 = -0.08$, $t(102.8) = -0.45$, $p > 0.05$; $\beta_6 = 0.08$, $t(239) = 0.61$, $p > 0.05$).

Adding the likability predictor at the time of evaluation in Model 1b led to a higher improvement of model fit compared to the null model than in Model 1a. The more likable that students rated the teacher at the time of evaluation, the higher they evaluated teacher performance in seminars ($\beta_2 = -0.68$, $t(255.9) = -17.14$, $p < 0.001$). This predictor explained 54.7% of the total variance and led to a decrease in the teacher variance component by 100% compared to the null model.

Including the prior subject interest predictor at the beginning of the course in Model 2a led to a significantly improved model fit compared to the null model. The more interesting that students rated the course at the beginning, the higher they evaluated teacher performance in seminars ($\beta_3 = -0.15$, $t(251.3) = -2.46$, $p < 0.01$). This predictor explained 3.1% of the total variance and led to an increase in the teacher variance component by 1.7%, a decrease in the student variance component by 8.9%, and a decrease in the course variance component by 21.4% compared to the null model.

Adding the prior subject interest predictor at the time of the evaluation in Model 2b led to no improvement of model fit compared to the null model. Accordingly, there was no significant relationship between prior subject interest at the time of evaluation and teacher performance in seminars ($\beta_4 = -0.03$, $t(237.6) = -0.46$, $p > 0.05$).

Including both predictors at the beginning of the course in Model 3 caused an improvement in model fit compared to the null model. The more likable ($\beta_1 = -0.48$, $t(242.2) = -7.89$, $p < 0.001$) and more interesting ($\beta_3 = -0.11$, $t(249.8) = -1.97$, $p < 0.05$) that students rated the teacher and the course at the beginning of the course, the higher they evaluated teacher performance in seminars. Both predictors together explained 22.3% of the total variance and led to a decrease in the teacher variance component by 62.3% compared to the null model.

Table 5
Estimates for the Cross-Classified Linear Mixed Effect Models for Teacher Performance in Seminars.

	Model 0 (Null Model)		Model 1a		Model 1b		Model 2a		Model 2b		Model 3	
Fixed effects	*Estimate* *(SE)*	*t*	*Estimate* *(SE)*	*t*	*Estimate* *(SE)*	*t*	*Estimate* *(SE)*	*t*	*Estimate* *(SE)*	*t*	*Estimate* *(SE)*	*t*
(Intercept)	2.005 (0.122)	16.48	1.973 (0.103)	19.22	1.828 (0.070)	25.98	2.035 (0.120)	16.97	2.014 (0.123)	16.43	1.995 (0.102)	19.60
Likability T1			−0.489*** (0.061)	−8.05							−0.479*** (0.061)	−7.89
Likability T2					−0.676*** (0.039)	−17.14						
Prior subject interest T1							−0.148** (0.060)	−2.46			−0.107* (0.054)	−1.97
Prior subject interest T2									−0.028 (0.060)	−0.46		
Random effects	*Variance*		*Variance*		*Variance*		*Variance*		*Variance*		*Variance*	
Residual	0.651		0.548		0.341		0.646		0.652		0.541	
Course (Intercept)	0.103		0.159		0.111		0.081		0.103		0.145	
Student (Intercept)	0.079		0.031		0.005		0.072		0.078		0.031	
Teacher (Intercept)	0.175		0.063		0.000		0.178		0.171		0.066	
Fit statistics												
-2LL	691.4		634.6+ + +		498.2+ + +		685.5+		691.2		630.7+ + +	
AIC	701.4		646.6		510.2		697.5		703.2		644.7	
BIC	719.2		667.9		531.5		718.8		724.5		669.6	

Note. Likability and prior subject interest were grand-mean centered before entered as predictors into the model. The number of observations that the variance components are based on: Residual: $N = 258$, Course: $n = 36$, Student: $n = 184$, Teacher: $n = 23$.
-2LL: -2 log-likelihood (deviance), AIC: Akaike information criterion, BIC: Bayesian information criterion.
Tests of fixed effects: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ (one-tailed).
Comparisons of models with the Null model (χ^2-difference tests with 1 *df* based on the deviances): + $p < 0.05$, + + $p < 0.01$, + + + $p < 0.001$ (two-tailed).

6.4.1. Summary and implications for the research questions

Similar to the results for lectures, likability was strongly associated with the planning and presentation scale. The association between prior subject interest and the global rating of teaching quality was significant, too, but again relatively weak (Research Questions 1 and 2). However, it remained significant when both likability and prior subject interest were included in the model (Research Question 4). These associations were found also when likability and prior subject interest were measured at the beginning of the course (Research Question 3). This pattern of effects corroborates the conclusions that may be drawn from the results obtained for lectures: There seems to be a strong biasing effect of likability and a much weaker but independent biasing effect of prior subject interest. In the next and final step of the analysis, we examined whether the same pattern of effects holds for planning and presentation in seminars, which would be even stronger evidence for a biasing effect of the two predictors.

6.5. Planning and presentation in seminars

Results for the six models with planning and presentation in seminars are shown in Table 6. The overall mean of 4.06 estimated in Model 0 indicates that planning and presentation in seminars was also rated as good.

Including the likability predictor at the beginning of the course in Model 1a led to a significantly improved model fit compared to the null model. The more likable that students rated the teacher at the beginning of the course, the higher they evaluated planning and presentation in seminars ($\beta_1 = 0.30$, $t(234.2) = 6.00$, $p < 0.001$). The predictor explained 14.2% of the total variance and led to an increase in the teacher variance component by 40.7% compared to the null model. Fig. 2 shows the total variance and the differences in the variance components of the null model compared to Model 1a and 1b. Inclusion of familiarity had no influence on planning and presentation in seminars (β_5 = -0.06, t (204.4) = -0.39, $p > 0.05$) but the interaction between likability and familiarity was significant ($\beta_6 = -0.19$, $t(219.7) = -1.87$, $p < 0.05$).

Including the predictor likability at the time of evaluation in Model 1b led to a higher improvement of model fit compared to the null model than in Model 1a. The more likable that students rated the teacher at the time of evaluation, the higher they evaluated planning and presentation in seminars ($\beta_2 = 0.40$, $t(240.7) = 10.80$, $p < 0.001$). This predictor explained 33.1% of the total variance and led to a decrease in the teacher variance component by 61.4% compared to the null model.

Including the prior subject interest predictor at the beginning of the course in Model 2a led to a significantly improved model fit compared to the null model. The more interesting that students rated the course at the beginning, the higher they evaluated planning and presentation in seminars ($\beta_3 = 0.16$, $t(247) = 3.54$, $p < 0.001$). This predictor explained 1.6% of the total variance and increased the teacher variance component by 1.2%, decreased the student variance component by 13.2%, and decreased the course variance component by 16% compared to the null model.

Adding the prior subject interest predictor at the time of evaluation in Model 2b led to no improvement of model fit compared to the null model. No significant relationship was found between prior subject interest at the time of evaluation and planning and presentation in seminars ($\beta_4 = 0.05$, $t(252.6) = 1.00$, $p > 0.05$).

Inclusion of both predictors at the beginning of the course in Model 3 caused an improvement in model fit compared to the null model. The more likable ($\beta_1 = 0.28$, $t(232.7) = 5.82$, $p < 0.001$) and more interesting ($\beta_3 = 0.14$, $t(240.3) = 3.23$, $p < 0.001$) that students rated the teacher and the course at the beginning of the course, the higher they evaluated planning and presentation in seminars. Both predictors together explained 16.6% of the total variance and led to a decrease in the teacher variance component by 37.8% compared to the null model.

6.5.1. Summary and implications for the research questions

The pattern of effects for planning and presentation in seminars exactly mirrors the effects obtained for the global rating of teaching quality. Again, strong associations were found for likability and weak associations for prior subject interest (Research Questions 1 and 2), when these variables were measured at the beginning of the course (Research Question 3) and even when both likability and prior subject

Table 6
Estimates for the Cross-Classified Linear Mixed Effect Models for Planning and Presentation in Seminars.

	Model 0 (Null Model)		Model 1a		Model 1b		Model 2a		Model 2b		Model 3	
Fixed effects	*Estimate* *(SE)*	*t*	*Estimate* *(SE)*	*t*	*Estimate* *(SE)*	*t*	*Estimate* *(SE)*	*t*	*Estimate* *(SE)*	*t*	*Estimate* *(SE)*	*t*
(Intercept)	4.062 (0.119)	34.02	4.077 (0.106)	38.59	4.171 (0.091)	45.62	4.023 (0.119)	33.74	4.046 (0.120)	33.68	4.044 (0.106)	38.08
Likability T1			0.295*** (0.049)	6.00							0.282*** (0.048)	5.82
Likability T2					0.396*** (0.037)	10.80						
Prior subject interest T1							0.164*** (0.046)	3.54			0.140*** (0.043)	3.23
Prior subject interest T2									0.047 (0.047)	1.00		
Random effects	*Variance*		*Variance*		*Variance*		*Variance*		*Variance*		*Variance*	
Residual	0.312		0.256		0.201		0.307		0.314		0.250	
Course (Intercept)	0.025		0.075		0.071		0.021		0.027		0.068	
Student (Intercept)	0.144		0.147		0.119		0.125		0.139		0.135	
Teacher (Intercept)	0.246		0.146		0.095		0.249		0.243		0.153	
Fit statistics												
-2LL	581.3		548.8+ + +		486.5+ + +		569.1+ + +		580.3		538.7+ + +	
AIC	591.3		560.8		498.5		581.1		592.3		552.7	
BIC	609.2		582.3		520.0		602.6		613.8		577.7	

Note. Likability and prior subject interest were grand-mean centered before entered as predictors into the model. The number of observations that the variance components are based on: Residual: $N = 264$, Course: $n = 36$, Student: $n = 189$, Teacher: $n = 23$.
-2LL: -2 log-likelihood (deviance), AIC: Akaike information criterion, BIC: Bayesian information criterion.
Tests of fixed effects: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ (one-tailed).
Comparisons of models with the Null model (χ^2-difference tests with 1 *df* based on the deviances): + $p < 0.05$, + + $p < 0.01$, + + + $p < 0.001$ (two-tailed).

interest were included in the model (Research Question 4). Thus, these results provide converging evidence for a strong biasing effect of likability and a much weaker but independent biasing effect of prior subject interest.

7. Discussion

Our study examined the validity of SETs by analysing the effects of teachers' likability perceived by students and the students' prior subject interest in the course. The results revealed that likability has a stronger effect on SETs than prior subject interest. These effects occurred with the global ratings of teacher performance but also with the more clearly defined measure of planning and presentation in lectures and in seminars. Most importantly, likability had consistent effects on both SET dimensions also when it was assessed at the beginning of the course even though these effects were smaller than the effects of likability assessed at the time of evaluation. Thus, likability seems to be affected by teacher behaviour to some degree, which is consistent with the assumption that likability overlaps to some extent with certain aspects of teaching quality (Delucchi, 2000). However, its robust relationship with planning and presentation and the large effects of likability on SETs assessed at the beginning of the course (even if students had never taken a class taught by the teacher) clearly attests to the classification of likability as a bias variable (Delucchi, 2000). The substantial decrease in the teacher variance component (> 30%) provides further evidence for this interpretation. One possible psychological mechanism behind this bias is the halo effect, i.e. the (unconscious) colouring or even distortion of judgments concerning specific attributes of a person due to global evaluations (e.g., Nisbett & Wilson, 1977).

The finding that the effect of likability assessed at the beginning of the course was smaller (9–20% explained variance) than the effect of likability assessed at the time of evaluation (21–55% explained variance) suggests that the biasing effect of likability is overestimated when it is assessed retrospectively after the course has started. Likability assessed at this point might be affected by events occurring during the course, some of which might be related to teaching quality.

Students' familiarity with the teacher had no influence on the effects of likability on SETs. This result is noteworthy and consistent with previous findings that impressions of people are formed fast and remain stable even after short exposure times (e.g., Ambady & Rosenthal, 1993; Willis & Todorov, 2006). Judgments of likability apparently change little after the first impression of a teacher has been formed. Another possible explanation might be that likability judgments by students who did not know the teacher before the course might have been based at least in part on the reputation of the teacher among the students. This reputation might have created expectations in students that might have had an influence on their ratings of likability of the teacher and also on their SETs (Griffin, 2001). Further research might shed light on the mechanism behind these surprisingly stable likability ratings.

The second potential bias variable, prior subject interest, was consistently related to both the global rating of teacher performance and the scale planning and presentation when it was measured at the beginning of the course, whereas there was only a significant effect of prior subject interest measured at the time of evaluation on planning and presentation in lectures. However, with only 1–3% explained variance (compared to the null model), the bias introduced by prior subject interest seems to be relatively weak. At first glance, this result seems to be at odds with prior research that has identified prior subject interest as one of the strongest background variables related to SETs (for a review, see Marsh, 2007; Marsh & Cooper, 1981). However, a closer look at previous studies provides a different picture. Marsh and Cooper (1981) reported a proportion of variance of only 5% explained by prior subject interest. Wolbring and Treischl (2016) found 5% variance in SETs explained by four variables that included prior subject interest, and Marsh (1982) found in his study that prior subject interest explained only 4% of the variance of a global rating of teacher performance and less than 1% variance explained for a SET dimension called organisation. Similarly, Olivares (2001) found that only 4% of variance was explained by cognitive ability and prior subject interest measured at the beginning of the course. In sum, the majority of previous studies found rather weak relationships of prior subject interest and SETs, which suggests that prior subject interest exerts a consistent but relatively harmless bias that only slightly compromises the validity of SETs. Moreover, it is important to note that only the association of prior subject interest measured at the beginning of the lecture or seminar with planning and presentation subscale can clearly be interpreted as a biasing effect. Following Marsh (1984, 2007), the association with the global rating of teaching quality does necessarily not speak against the validity of this SET dimension. The reason is that higher prior subject interest in the learners might enable teachers to provide better instruction, which may very well account for the (weak) association between the two variables.

7.1. Limitations of the present study

The results of the present study are informative but need to be interpreted with certain limitations in mind. First, the results are based on a sample of SETs from only one course progam (psychology) and on only one semester measured at one university in Germany. We included students, teachers, and courses as random effects in our models to account for the fact that they were drawn from larger populations, but at this point the exact definition of these populations remains unclear. The problem of unclear generalisability is aggravated by the fact that the lecture sample was based on only eight lectures and that students voluntarily took part in the SETs, yielding a convenience sample (a shortcoming shared by most other studies in this area).

Another potential limitation is that we based our analyses on online SETs. This survey mode might lead to different results compared to the research based on paper-pencil SETs due to a lower response rate (e.g., Dommeyer, Baum, Hanna, & Chapman, 2004). Our solution to this problem was to arrange 10 min in each course to fill out the online questionnaire. During this time, students were asked to provide evaluations with their smartphones or laptops. The survey platform used for this study (Unipark) supports surveys designed for both types of devices. It should also be noted that Dommeyer et al. (2004) and Treischl and Wolbring (2017) found no or only small differences between the two modes of administering SETs. The differences that are found seem to be caused more by the time and place of evaluation (in-class vs. after-class) than by the survey mode (paper-pencil vs. online, Kordts-Freudinger & Geithner, 2013).

8. Conclusion

Our study provides evidence that SETs are affected by strong biasing effects of how likable students find a teacher and by weak biasing effects of how strongly they are interested in the course subject. Given that both constructs were measured at the beginning of the course and were thus outside the influence the teacher's behaviour, our results (especially for likability) cast some doubt on the validity of SETs. Results from SETs should be used and interpreted with caution. They seem to reflect likability but not teaching quality to a considerable degree (Clayson & Haley, 1990).

Appendix

Adapted Likability Scale (Reysen, 2005; German/English Translation)

Likert-scale with response options ranging from 1 (*strongly disagree*) to 5 (*strongly agree*).

1 Der Dozent/Die Dozentin ist freundlich.
The teacher is friendly.
2 Der Dozent/Die Dozentin ist sympathisch.
The teacher is likeable.
3 Der Dozent/Die Dozentin ist warmherzig.
The teacher is warm.
4 Der Dozent/Die Dozentin ist zugänglich.
The teacher is approachable.
5 Ich würde den Dozenten/die Dozentin um Rat bitten.
I would ask the teacher for advice.
6 Der Dozent/Die Dozentin ist attraktiv.
The teacher is physically attractive.
7 Der Dozent/Die Dozentin ist mir ähnlich.
The teacher is similar to me.
8 Der Dozent/Die Dozentin ist kenntnisreich.
The teacher is knowledgeable.

References

Ambady, N., & Rosenthal, R. (1993). Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *Journal of Personality and Social Psychology, 64*, 431–441.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390–412.

Barr, A. S. (1943). Chapter II: The measurement and prediction of teaching efficiency. *Review of Educational Research, 13*, 218–223.

Barth, M. M. (2008). Deciphering student evaluations of teaching: A factor analysis approach. *Journal of Education for Business, 84*, 40–46.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*, 1–48.

Clayson, D. E. (1999). Students' evaluation of teaching effectiveness: Some implications of stability. *Journal of Marketing Education, 21*, 68–75.

Clayson, D. E., & Haley, D. A. (1990). Student evaluations in marketing: What is actually being measured? *Journal of Marketing Education, 12*, 9–17.

Clayson, D. E., & Sheffet, M. J. (2006). Personality and the student evaluation of teaching. *Journal of Marketing Education, 28*, 149–160.

Delucchi, M. (2000). Don't worry, be happy: Instructor likability, student perceptions of learning, and teacher ratings in upper-level sociology courses. *Teaching Sociology, 28*, 220–231.

Dommeyer, C. J., Baum, P., Hanna, R. W., & Chapman, K. S. (2004). Gathering faculty teaching evaluations by in-class and online surveys: Their effects on response rates and evaluations. *Assessment and Evaluation in Higher Education, 29*, 611–623.

Dresel, M., & Rindermann, H. (2011). Counseling university instructors based on student evaluations of their teaching effectiveness: A multilevel test of its effectiveness under consideration of bias and unfairness variables. *Research in Higher Education, 52*, 717–737.

Dziuban, C., & Moskal, P. (2011). A course is a course is a course: Factor invariance in student evaluation of online, blended and face-to-face learning environments. *The Internet and Higher Education, 14*, 236–241.

Earley, P., & Porritt, V. (2014). Evaluating the impact of professional development: The need for a student-focused approach. *Professional Development in Education, 40*, 112–129.

Faranda, W. T., & Clarke, I. (2004). Student observations of outstanding teaching: Implications for marketing educators. *Journal of Marketing Education, 26*, 271–281.

Feistauer, D., & Richter, T. (2017). How reliable are students' evaluations of teaching quality? A variance components approach. *Assessment and Evaluation in Higher Education, 42*, 1263–1279.

Feistauer, D., & Richter, T. (2018). The role of clarity about study programme contents and interest in student evaluations of teaching. *Psychology Learning & Teaching*.

Friedman, H. S., Riggio, R. E., & Casella, D. F. (1988). Nonverbal skill, personal charisma, and initial attraction. *Personality and Social Psychology Bulletin, 14*, 203–221.

Frymier, A. B. (1994). The use of affinity-seeking in producing liking and learning in the classroom. *Journal of Applied Communication Research, 22*, 87–105.

Greenwald, A. G. (1997). Validity concerns and usefulness of student ratings of instruction. *American Psychologist, 52*, 1182–1186.

Griffin, B. W. (2001). Instructor reputation and student ratings of instruction. *Contemporary Educational Psychology, 26*, 534–552.

Gruber, T., Lowrie, A., Brodowsky, G. H., Reppel, A. E., Voss, R., & Chowdhury, I. N. (2012). Investigating the influence of professor characteristics on student satisfaction and dissatisfaction: A comparative study. *Journal of Marketing Education, 34*, 165–178.

Gurung, R. A. R., & Vespia, K. M. (2007). Looking good, teaching well? Linking liking, looks, and learning. *Teaching of Psychology, 34*, 5–10.

Hawkins, S. A., & Hastie, R. (1990). Hindsight: Biased judgments of past events after the outcomes are known. *Psychological Bulletin, 107*, 311–327.

Howard, G. S., & Schmeck, R. R. (1979). Relationship of changes in student motivation to student evaluations of instruction. *Research in Higher Education, 10*, 305–315.

Kelley, T. L. (1927). *Interpretation of educational measurements.* New York: Macmillan.

Kenny, D. A. (1979). *Correlation and causality.* New York: Wiley.

Kordts-Freudinger, R., & Geithner, E. (2013). When mode does not matter: Evaluation in class versus out of class. *Educational Research and Evaluation, 19*, 605–614.

Marks, R. B. (2000). Determinants of student evaluations of global measures of instructor and course value. *Journal of Marketing Education, 22*, 108–119.

Marsh, H. W. (1980). The influence of student, course and instructor characteristics on evaluations of university teaching. *American Educational Research Journal, 17*, 219–237.

Marsh, H. W. (1981). Students' evaluations of tertiary instruction: Testing the applicability of American surveys in an Australian setting. *Australian Journal of Education, 25*, 177–192.

Marsh, H. W. (1982). SEEQ: A reliable, valid, and useful instrument for collecting students' evaluations of university teaching. *British Journal of Educational Psychology, 52*, 77–95.

Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology, 76*(5), 707–754.

Marsh, H. W. (2007). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. In R. P. Perry, & J. C. Smart (Eds.). *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 319–383). Dordrecht: Springer Netherlands.

Marsh, H. W., & Cooper, T. L. (1981). Prior subject interest, students' evaluations, and instructional effectiveness. *Multivariate Behavioral Research, 16*, 83–104.

Marsh, H. W., & Hocevar, D. (1991). Students' evaluations of teaching effectiveness: The stability of mean ratings of the same teachers over a 13-year period. *Teaching and Teacher Education, 7*, 303–314.

Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist, 52*, 1187–1197.

Nisbett, R. E., & Wilson, T. D. (1977). The halo effect: Evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology, 35*, 250–256.

Olivares, O. J. (2001). Student interest, grading leniency, and teacher ratings: A conceptual analysis. *Contemporary Educational Psychology, 26*, 382–399.

Olivares, O. J. (2003). A conceptual and analytic critique of student ratings of teachers in the USA with implications for teacher effectiveness and student learning. *Teaching in Higher Education, 8*, 233–245.

R Core Team (2017). *R: A language and environment for statistical computing [Computer program]*. URLVienna, Austria: R foundation for statistical computing. https://www.R-project.org/.

Raudenbush, S. W., & Bryk, A. S. (2006). *Hierarchical linear models applications and data analysis methods: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Reysen, S. (2005). Construction of a new scale: The Reysen Likability Scale. *Social Behavior and Personality an International Journal, 33*, 201–208.

Ross, M. (1989). Relation of implicit theories to the construction of personal histories. *Psychological Review, 96*, 341–357.

Skinner, E. A., & Belmont, M. J. (1993). Motivation in the classroom: Reciprocal effects of teacher behavior and student engagement across the school year. *Journal of Educational Psychology, 85*, 571–581.

Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching: The state of the art. *Review of Educational Research, 83*, 598–642.

Spooren, P., Mortelmans, D., & Denekens, J. (2007). Student evaluation of teaching quality in higher education. Development of an instrument based on 10 Likert scales. *Assessment and Evaluation in Higher Education, 32*, 667–679.

Stalnaker, J. M., & Remmers, H. H. (1928). Can students discriminate traits associated with success in teaching? *Journal of Applied Psychology, 12*, 602–610.

Staufenbiel, T. (2000). Fragebogen zur Evaluation von universitären Lehrveranstaltungen durch Studierende und Lehrende [Students course assessment questionnaire for evaluation of university courses]. *Diagnostica, 46*, 169–181.

Staufenbiel, T., Seppelfricke, T., & Rickers, J. (2016). Prädiktoren studentischer Lehrveranstaltungsevaluationen: Eine Mehrebenenanalyse [Predictors of student evaluations of teaching: A multilevel analysis]. *Diagnostica, 62*, 44–59.

Treischl, E., & Wolbring, T. (2017). The causal effect of survey mode on students' evaluations of teaching: Empirical evidence from three field experiments. *Research in Higher Education, 58*, 904–921.

Uranowitz, S. W., & Doyle, K. O. (1978). Being liked and teaching: The effects and bases of personal likability in college instruction. *Research in Higher Education, 9*, 15–41.

Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological Science, 17*, 592–598.

Wolbring, T., & Riordan, P. (2016). How beauty works: Theoretical mechanisms and two empirical applications on students' evaluation of teaching. *Social Science Research, 57*, 253–272.

Wolbring, T., & Treischl, E. (2016). Selection bias in students' evaluation of teaching. *Research in Higher Education, 57*, 51–71.