Validity of Students’ Evaluations of Teaching: Biasing Effects of Likability and Prior Subject Interest

Daniela Feistauer
University of Kassel

Tobias Richter
University of Würzburg

Accepted for publication in the journal *Studies in Educational Evaluation* (2018)

Corresponding author:
Tobias Richter
University of Würzburg, Department of Psychology IV
Röntgenring 10
97070 Würzburg, Germany
E-mail: tobias.richter@uni-wuerzburg.de
Telephone: +49-931 31 83755
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.

Keywords: cross-classified multilevel analysis, likability, prior subject interest, student evaluations of teaching, variance components
Validity of Students’ Evaluations of Teaching: Biasing Effects of Likability and Prior Subject Interest

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 and Roche 1997; Olivares 2003; Spooren, Brockx, and Mortelmans 2013; Stalnaker and Remmers 1928; Staufenbiel, Seppelfricke, and 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 at 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.

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 and Rosenthal 1993; Clayson

Most of these studies showed such strong relationships between the studied predictor and SETs that some authors named SETs “happy sheets” (Earley and Porritt 2014, p. 112), “likability scales” (Clayson and Haley, 1990, p. 13), or “popularity contests” (Dziuban and Moskal 2011, p. 237; Uranowitz and 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 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.

Prior Subject Interest and Students’ Evaluations of Teaching

Prior subject interest can be understood as the individual student’s initial interest in the subject matter before attending the course. An item assessing prior subject interest is included in most standardized SETs (e.g., Spooren, Mortelmans, and Denekens 2007; Stalnaker and 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 and 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 and 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.

**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
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 and Sheffet 2006; Howard and Schmeck 1979). Howard and Schmeck measured motivation, similar to prior subject interest, and found a significant correlation ($r = .61$) between pre-course motivation and post-course ratings of 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 to 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 lesson 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, and 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.

**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, and 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 and 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 of bias 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 the prior subject interest as predictors 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).

**Method**

**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.

**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 minutes 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 minutes 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.

**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.

**Criterion Variables**

**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: $M = 1.98$, $SD = 0.99$).

**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 $\alpha = .85$) and eight items in seminars ($M = 4.11$, $SD = 0.82$,
Cronbach’s $\alpha = .85$).

**Predictor Variables**

**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 (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 .58 in lectures
and .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 $\alpha$) of .92 in lectures (Table 1) and .89 in seminars (Table 2) and is provided in Appendix A. The single item and the scale correlated to .98 in lectures (Table 1) and .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 his/her courses before.

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 ($r = .87$) in lectures and ($r = .72$) 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 $r = .14$.

**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, and 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.

**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:

$$Y_{sct} = \theta_0 + h_{0st} + i_{0tc} + j_{0ct} + e_{sct}$$ (0)

In Equation 0, $Y_{sct}$ represents the evaluation score provided by student $s$ for courses $c$ given by teacher $t$. The intercept $\theta_0$ represents the grand mean of this score across all students, courses,
and teachers. The random effect \( h_{0s} \) captures the individual deviation of student \( s \) from \( \theta_0 \).
Likewise, the random effect \( i_{0c} \) represents the deviation of course \( c \) from \( \theta_0 \), and the random effect \( j_{0t} \) the deviation of teacher \( t \) from \( \theta_0 \). The variances \( \tau_{0s}, \tau_{0c}, \) and \( \tau_{0t} \) 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 \( \sigma^2 \) (Raudenbush and Bryk 2006).

The model in Equation 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 predictor likability predictor at the beginning of the course \( LT1 \), with its slope \( \beta_1 \) in Model 1a:

\[
Y_{sct} = \theta_0 + \beta_1 LT1_s + h_{0s} + i_{0c} + j_{0t} + e_{sct}
\]  

(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 \( LT2_{act} \) with its slope \( \beta_2 \) was added:

\[
Y_{sct} = \theta_0 + \beta_2 LT2_{act} + h_{0s} + i_{0c} + j_{0t} + e_{sct}
\]  

(1b)

In Model 2a, the prior subject interest predictor at the beginning of the course \( IT1 \), with its slope \( \beta_3 \) was added:
In Model 2b, the prior subject interest predictor at the time of evaluation \(IT_{2ss}\) with its slope \(\beta_4\) was added:

\[
Y_{ssct} = \theta_0 + \beta_4 IT_{2ss} + h_{00s} + i_{00s} + j_{00s} + e_{ssct} \tag{2b}
\]

In Model 3, both predictors were added, likability at the beginning of the course \(LT_1\) with its slope \(\beta_1\) and prior subject interest at the beginning of the course \(IT_1\) with its slope \(\beta_3\).

\[
Y_{ssct} = \theta_0 + \beta_1 LT_1 + \beta_3 IT_1 + h_{00s} + i_{00s} + j_{00s} + e_{ssct} \tag{3}
\]

**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 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 (\(\beta_1 = -0.24, t(245.5) = -4.34, p < .001\)). 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. Figure 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 (\(\beta_5 = -0.12, t(67.7) = -.93, p > .05\); \(\beta_6 = 0.15, t(244.4) = 1.33, p > .05\)).

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 ($\beta_2 = -0.51$, $t(243.8) = -12.21, p < .001$). This predictor explained 36.5% of the total variance and led to an increase in the teacher 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 ($\beta_3 = -0.12$, $t(242) = -2.25, p < .05$). 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 ($\beta_4 = -0.076$, $t(246.9) = -1.45, p > .05$).

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 ($\beta_1 = -0.23$, $t(238.7) = -4.03, p < .001$) but not prior interest ($\beta_3 = -0.081$, $t(238.2) = -1.59$, $p > .05$) 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.

**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.

**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.

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 ($\beta_1 = 0.21$, $t(224.2) = 4.53, p < .001$). 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. Figure 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 ($\beta_5 = 0.15$, $t(36.1) = 1.47, p > .05$; $\beta_6 = -0.06$, $t(195.2) = -0.60, p > .05$).
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 ($\beta_2 = 0.35, t(250.8) = 9.09, p < .001$). 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 ($\beta_3 = 0.16, t(250.8) = 3.62, p < .001$) 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 also 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 ($\beta_4 = 0.10, t(251.1) = 2.26, p < .05$). 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 ($\beta_1 = 0.18, t(230) = 3.92, p < .001$) and more interesting ($\beta_3 = 0.13, t(239.2) = 2.93, p < .01$) 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.

**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.

**Ratings of Teacher Performance in Seminars**

Results for the six models with teacher performance in seminars are shown in Table 5. The overall mean (i.e., the intercept) 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 ($\beta = -0.49$, $t(241.4) = -8.05$, $p < .001$). 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. Figure 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 ($\beta_5 = -0.08$, $t(102.8) = -0.45, p > .05; \beta_6 = 0.08, t(239) = 0.61, p > .05$).

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 ($\beta_2 = -0.68$, $t(255.9) = -17.14, p < .001$). 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 ($\beta_3 = -0.15$, $t(251.3) = -2.46, p < .01$). 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 ($\beta_4 = -0.03$, $t(237.6) = -0.46, p > .05$).

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 ($\beta_1 = -0.48$, $t(242.2) = -7.89, p < .001$) and more interesting ($\beta_3 = -0.11$, $t(249.8) = -1.97, p < .05$) 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.

**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.

**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 ($\beta_1 = 0.30$, $t(234.2) = 6.00$, $p < .001$). 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. Figure 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 ($\beta_5 = -0.06, t(204.4) = -0.39, p > .05$) but the interaction between likability and familiarity was significant ($\beta_6 = -0.19, t(219.7) = -1.87, p < .05$).

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 ($\beta_2 = 0.40, t(240.7) = 10.80, p < .001$). 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 ($\beta_3 = 0.16, t(247) = 3.54, p < .001$). This predictor explained 3.4% 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 ($\beta_4 = 0.05, t(252.6) = 1.00, p > .05$).
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 ($\beta_1 = 0.28$, $t(232.7) = 5.82, p < .001$) and more interesting ($\beta_3 = 0.14$, $t(240.3) = 3.23, p < .001$) 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.

**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 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.

**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 (Delucci 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 (Delucci 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 and 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 and Rosenthal 1993; Willis and 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 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 and 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.

**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 subject (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 et al. 2004). Our solution to this problem was to arrange 10 minutes 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 and Geithner 2013).
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 and Haley 1990).
References


evaluations by in-class and online surveys: Their effects on response rates and

evaluations of their teaching effectiveness: A multilevel test of its effectiveness under
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


variance components approach. *Assessment & 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 and Teaching*.

Friedman, H. S., Riggio, R. E., & Casella, D. F. (1988). Nonverbal skill, personal charisma, and

Frymier, A. B. (1994). The use of affinity-seeking in producing liking and learning in the


### Table 1

*Intercorrelations Between Predictor Variables for Lectures*

<table>
<thead>
<tr>
<th></th>
<th>Likability T1</th>
<th>Likability T2</th>
<th>Likability Scale</th>
<th>Prior subject interest T1</th>
<th>Prior subject interest T2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likability T1</td>
<td>3.97</td>
<td>0.46</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Likability T2</td>
<td>4.10</td>
<td>0.44</td>
<td>0.58</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Likability scale</td>
<td>3.89</td>
<td>0.37</td>
<td>0.98***</td>
<td>0.46</td>
<td>(0.92)</td>
</tr>
<tr>
<td>Prior subject interest T1</td>
<td>3.60</td>
<td>0.49</td>
<td>-0.15</td>
<td>&lt; 0.01</td>
<td>-0.07</td>
</tr>
</tbody>
</table>
| 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 < .001 (two-tailed).**
Table 2

**Intercorrelations Between Predictor Variables for Seminars**

<table>
<thead>
<tr>
<th></th>
<th>Likability T1</th>
<th></th>
<th>Likability T2</th>
<th></th>
<th>Scale Likability</th>
<th>Prior subject interest T1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likability T1</td>
<td>3.93</td>
<td>0.55</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Likability T2</td>
<td>3.93</td>
<td>0.69</td>
<td>0.55***</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Likability scale</td>
<td>3.83</td>
<td>0.38</td>
<td>0.89***</td>
<td>0.45*</td>
<td>(0.89)</td>
<td></td>
</tr>
<tr>
<td>Prior subject interest T1</td>
<td>3.63</td>
<td>0.58</td>
<td>-0.01</td>
<td>0.21</td>
<td>0.10</td>
<td></td>
</tr>
<tr>
<td>Prior subject interest T2</td>
<td>3.72</td>
<td>0.76</td>
<td>0.02</td>
<td>-0.09</td>
<td>0.09</td>
<td>0.72</td>
</tr>
</tbody>
</table>

*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 $\alpha$ 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).

* $p < .05$, *** $p < .001$ (two-tailed).
Table 3

*Estimates for the Cross-Classified Linear Mixed Effect Models for Teacher Performance in Lectures*

<table>
<thead>
<tr>
<th></th>
<th>Model 0 (Null Model)</th>
<th>Model 1a</th>
<th>Model 1b</th>
<th>Model 2a</th>
<th>Model 2b</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fixed effects</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>Estimate (SE)</td>
<td>t</td>
<td>Estimate (SE)</td>
<td>t</td>
<td>Estimate (SE)</td>
<td>t</td>
</tr>
<tr>
<td></td>
<td>1.870 (0.110)</td>
<td>17.05</td>
<td>1.846 (0.099)</td>
<td>18.62</td>
<td>1.799 (0.089)</td>
<td>20.30</td>
</tr>
<tr>
<td></td>
<td>-0.244*** (0.056)</td>
<td>-4.34</td>
<td>-0.244*** (0.056)</td>
<td>-4.34</td>
<td>-0.244*** (0.056)</td>
<td>-4.34</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Likability T1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.228*** (0.057)</td>
<td>-4.03</td>
<td>-0.228*** (0.057)</td>
<td>-4.03</td>
<td>-0.228*** (0.057)</td>
<td>-4.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Likability T2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prior subject</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>interest T1</td>
<td>-0.116* (0.052)</td>
<td>-2.25</td>
<td>-0.116* (0.052)</td>
<td>-2.25</td>
<td>-0.116* (0.052)</td>
<td>-2.25</td>
</tr>
<tr>
<td>Prior subject</td>
<td>-0.076 (0.053)</td>
<td>-1.45</td>
<td>-0.076 (0.053)</td>
<td>-1.45</td>
<td>-0.076 (0.053)</td>
<td>-1.45</td>
</tr>
<tr>
<td>interest T2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Random effects</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Residual</td>
<td>Variance</td>
<td></td>
<td>Variance</td>
<td></td>
<td>Variance</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.419</td>
<td></td>
<td>0.408</td>
<td></td>
<td>0.302</td>
<td></td>
</tr>
<tr>
<td>Course (Intercept)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Student (Intercept)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher (Intercept)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Fit statistics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-2LL</td>
<td>545.0</td>
<td></td>
<td>527.4+++</td>
<td></td>
<td>433.7+++</td>
<td></td>
</tr>
<tr>
<td>AIC</td>
<td>555.0</td>
<td></td>
<td>539.4</td>
<td></td>
<td>445.7</td>
<td></td>
</tr>
<tr>
<td>BIC</td>
<td>572.7</td>
<td></td>
<td>560.6</td>
<td></td>
<td>466.9</td>
<td></td>
</tr>
</tbody>
</table>

*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 < .05, **p < .01, ***p < .001 (one-tailed).

Comparisons of models with the Null model (χ²-difference tests with 1 df based on the deviances): +p < .05, ++p < .01, +++p < .001 (two-tailed).
Table 4

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

<table>
<thead>
<tr>
<th></th>
<th>Model 0 (Null Model)</th>
<th>Model 1a</th>
<th>Model 1b</th>
<th>Model 2a</th>
<th>Model 2b</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fixed effects</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Intercept)</td>
<td>4.111 (0.101)</td>
<td>4.145 (0.078)</td>
<td>4.146 (0.099)</td>
<td>4.094 (0.103)</td>
<td>4.107 (0.106)</td>
<td>4.125 (0.081)</td>
</tr>
<tr>
<td>Likability T1</td>
<td>0.213*** (0.047)</td>
<td></td>
<td>0.350*** (0.039)</td>
<td></td>
<td></td>
<td>0.185*** (0.047)</td>
</tr>
<tr>
<td>Likability T2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prior subject interest T1</td>
<td></td>
<td>0.156*** (0.043)</td>
<td></td>
<td></td>
<td>0.125** (0.043)</td>
<td>2.93</td>
</tr>
<tr>
<td>Prior subject interest T2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.101* (0.044)</td>
<td>2.26</td>
</tr>
<tr>
<td><strong>Random effects</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Residual</td>
<td>0.258</td>
<td>0.257</td>
<td>0.196</td>
<td>0.252</td>
<td>0.257</td>
<td>0.253</td>
</tr>
<tr>
<td>Course (Intercept)</td>
<td>0.019</td>
<td>0.002</td>
<td>0.000</td>
<td>0.031</td>
<td>0.016</td>
<td>0.011</td>
</tr>
<tr>
<td>Student (Intercept)</td>
<td>0.088</td>
<td>0.070</td>
<td>0.066</td>
<td>0.073</td>
<td>0.080</td>
<td>0.059</td>
</tr>
<tr>
<td>Teacher (Intercept)</td>
<td>0.049</td>
<td>0.032</td>
<td>0.065</td>
<td>0.044</td>
<td>0.060</td>
<td>0.029</td>
</tr>
<tr>
<td><strong>Fit statistics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-2LL</td>
<td>458.4</td>
<td>440.5+++</td>
<td>387.9+++</td>
<td>446.0+++</td>
<td>453.5+</td>
<td>432.5+++</td>
</tr>
<tr>
<td>AIC</td>
<td>468.4</td>
<td>452.5</td>
<td>399.9</td>
<td>458.0</td>
<td>465.5</td>
<td>446.5</td>
</tr>
<tr>
<td>BIC</td>
<td>486.1</td>
<td>473.7</td>
<td>421.1</td>
<td>479.2</td>
<td>486.7</td>
<td>471.2</td>
</tr>
</tbody>
</table>

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 < .05, ** p < .01, *** p < .001 (one-tailed).
Comparisons of models with the Null model (χ²-difference tests with 1 df based on the deviances): + p < .05, ++ p < .01, +++ p < .001 (two-tailed).
Table 5

*Estimates for the Cross-Classified Linear Mixed Effect Models for Teacher Performance in Seminars*

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

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: \(* p < .05, ** p < .01, *** p < .001\) (one-tailed). Comparisons of models with the Null model (\(\chi^2\)-difference tests with 1 df based on the deviances): \(+ p < .05, ++ p < .01, +++ p < .001\) (two-tailed).
Table 6

*Estimates for the Cross-Classified Linear Mixed Effect Models for Planning and Presentation in Seminars*

<table>
<thead>
<tr>
<th>Fixed effects</th>
<th>Model 0 (Null Model)</th>
<th>Model 1a</th>
<th>Model 1b</th>
<th>Model 2a</th>
<th>Model 2b</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate (SE)</td>
<td>t</td>
<td>Estimate (SE)</td>
<td>t</td>
<td>Estimate (SE)</td>
<td>t</td>
</tr>
<tr>
<td>(Intercept)</td>
<td>4.062 (0.119)</td>
<td>34.02</td>
<td>4.077 (0.106)</td>
<td>38.59</td>
<td>4.171 (0.091)</td>
<td>45.62</td>
</tr>
<tr>
<td></td>
<td>0.295*** (0.049)</td>
<td>6.00</td>
<td>0.396*** (0.037)</td>
<td>10.80</td>
<td>0.164*** (0.046)</td>
<td>3.54</td>
</tr>
<tr>
<td>Likability T1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Likability T2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prior subject</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>interest T1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prior subject</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>interest T2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Random effects</td>
<td>Variance</td>
<td>Variance</td>
<td>Variance</td>
<td>Variance</td>
<td>Variance</td>
<td>Variance</td>
</tr>
<tr>
<td>Residual</td>
<td>0.312</td>
<td>0.256</td>
<td>0.201</td>
<td>0.307</td>
<td>0.314</td>
<td>0.250</td>
</tr>
<tr>
<td>Course (Intercept)</td>
<td>0.025</td>
<td>0.075</td>
<td>0.071</td>
<td>0.021</td>
<td>0.027</td>
<td>0.068</td>
</tr>
<tr>
<td>Student (Intercept)</td>
<td>0.144</td>
<td>0.147</td>
<td>0.119</td>
<td>0.125</td>
<td>0.139</td>
<td>0.135</td>
</tr>
<tr>
<td>Teacher (Intercept)</td>
<td>0.246</td>
<td>0.146</td>
<td>0.095</td>
<td>0.249</td>
<td>0.243</td>
<td>0.153</td>
</tr>
<tr>
<td>Fit statistics</td>
<td>-2LL</td>
<td>581.3</td>
<td>548.8+++</td>
<td>486.5+++</td>
<td>569.1+++</td>
<td>580.3</td>
</tr>
<tr>
<td></td>
<td>AIC</td>
<td>591.3</td>
<td>560.8</td>
<td>498.5</td>
<td>581.1</td>
<td>592.3</td>
</tr>
<tr>
<td></td>
<td>BIC</td>
<td>609.2</td>
<td>582.3</td>
<td>520.0</td>
<td>602.6</td>
<td>613.8</td>
</tr>
</tbody>
</table>

*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: * $p < .05$, ** $p < .01$, *** $p < .001$ (one-tailed).

Comparisons of models with the Null model ($\chi^2$-difference tests with 1 $df$ based on the deviances): + $p < .05$, ++ $p < .01$, +++ $p < .001$ (two-tailed).
Figures

Figure 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.
Figure 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.
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.