Teaching Quality in Higher Education: Do Student Evaluation of Teaching Questionnaires Allow a Reliable and Valid Assessment of Teaching Quality?

Dissertation
Zur Erlangung des akademischen Grades
Doktor der Philosophie (Dr. phil.)

Vorgelegt im Fachbereich 01 Humanwissenschaften
der Universität Kassel

von Dipl.-Psych. Daniela Feistauer

Kassel, Januar 2018

Erstgutachter:	Prof. Dr. Tobias Richter
Zweitgutachter:	Prof. Dr. Martin Hänze
Drittgutachter:	Prof. Dr. Holger Horz
Tag der Disputation:	16.04.2018

Table of Content

Summary

Teaching quality is an important and relevant topic for many stakeholders in society (e.g., the government, taxpayers, students, parents, and teachers) who view education as an imperative to nurture the innate potential of all students and to meet the challenges in the 21st century. Typically, education is achieved through systematic teaching that can be delivered, for example, by professionally trained teachers. Many strategies and tools are used to promote discussion and continuous improvement in teaching quality. One often applied tool for assessing teaching quality in higher education around the world is the student evaluation of teaching (SET). SETs are also used as a criterion for making important decisions in higher education, for example, when employing teachers, distributing funds, and making changes in the curriculum. Consequently, the reliability and validity of these instruments are important topics in applied research, especially given the widespread use of SETs as a tool for making important decisions in education that affect the future of teachers and indirectly the future of their students. Since their first implementation in 1905 (Sears, 1921), SETs have been the focus of extensive research. However, the extent that SETs are a valid and reliable assessment of teaching quality is still relatively unclear. Thus, more research is needed to inform educators as to whether SETs provide a valid and reliable measure of teaching quality and whether they should be used to make important decisions that affect the future of all relevant stakeholders.

The research in the current dissertation focuses on this topic and was guided by the two main concerns of validity and reliability of SETs as a measure of teaching quality. An important aspect of the investigation was to examine the influence of student characteristics that are not conceptually related to teaching quality. The influence of student characteristics on SETs was analyzed in two ways. First, the variance of students was estimated through cross-classified multilevel models. This type of analyses allows for the direct estimation of variance attributed to the students while separating it from the residual variance. Second, some student characteristics were added as predictors to the models to estimate their effect on SETs. The purpose of this analysis was to enhance the interpretation of the validity of SETs. In all studies, the German questionnaire FEVOR (Fragebogen zur Evaluation von universitären Lehrveranstaltungen durch Studierende [Students Course Assessment Questionnaire for Evaluation of University Courses]), developed by Staufenbiel (2000), was applied as the SET instrument.

The first study investigated the reliability of SETs as measurement of teaching quality. In this study, reliability was operationalized as interrater reliability calculated from intra-class correlations. The reliability can be assessed by comparing the variance components of teachers, courses, and students, and the instrument can be considered reliable only when a high proportion of variance is explained by teachers and not by students. The variance components of all possible variance sources (teacher, course, and student) were estimated with cross-classified multilevel models. The analysis of psychology student data revealed that teachers and courses were large sources of variance in different dimension of the FEVOR. This result suggests that SETs are reliable instruments if a sufficient number of students (at least 24) evaluate a teacher and course. However, the reliability of the FEVOR was higher in lectures than in seminars. Moreover, the data also revealed students and the interaction of students and teachers as relevant sources of variance. This finding implies that student characteristics and the individual fit between students and teachers can affect SETs and should be considered in the ongoing discussion of the instrument's validity.

The latter two studies examined the validity of SETs. The second study focused on the student's characteristics, *content expectations* and *prior subject interest*. Both variables are theoretically unrelated to teaching quality and are outside of the teacher's sphere of influence. Consequently, they should not affect the SET questionnaire. Therefore, the inclusion of both variables as predictors into cross-classified multilevel models should not lead to significant fixed effects. However, analyses of the data provided by students revealed significant but weak effects of both variables on the SET and thus did not pose a strong threat to the validity of SETs.

The third study also addressed the validity of SETs, extending the research of *prior subject interest* by measuring it twice, once at the beginning of the course and a second time retrospectively. Additionally, *likability of the teacher perceived by students* was measured once at the beginning of the course and a second time at the typical time of evaluation toward the end of the semester (in the 10th-12th session of a course). Similar to the results from Study 2, a weak effect was found of prior subject interest on the validity of SETs. However, the likability of teachers had a strong effect on the validity of SETs at both times of measurement. Likability measured at the time of evaluation explained nearly 50% of the variance compared to 20% of the variance at the beginning of the course. This result was interpreted as a strong effect of likability, which is theoretically unconnected to teaching quality, and thus undermines the validity of SETs.

The results of the three studies suggest that SETs are a reliable instrument (when a sufficient number of at least 24 students complete the evaluation). The student characteristics, content expectations and prior subject interest, showed weak effects on SETs, whereas the likability of the teacher perceived by students strongly affected the SETs and hence was a solid threat to its validity as a measurement of teaching quality.

CHAPTER 1

Introduction and Aims

Introduction and Aims

Quality is an important goal in a performance-orientated society. Hence, the transfer from its original economic context to the context of higher education is not surprising. Educating citizens is in the best interest of society to achieve the highest possible performance from its citizens in solving social problems and improving the standard of living. Such a high performance can be achieved when citizens are educated in a way that they participate positively in a society (Colby & Witt, 2000). This goal can be achieved best by obtaining quality measures that help distinguish between poor and excellent education. A society is created by different stakeholders such as government, students, parents, teachers and taxpayers. These stakeholders are interested in the quality of education despite their different perspectives on the topic. For example, taxpayers want their taxes to be allocated in the most appropriate way for the educational system (Volante, 2007) unencumbered by special interest. In contrast, students and parents are concerned about quality in education in a more direct way: An education of high quality can result in a personal advantage, because students are likely to find better jobs than students who attend schools of less quality. The term better means in this context more prestige, more money, or even free choice in the job market. Quality provides students and their parents with “comfort” (Power, 1997, p. 96) that they have made the right choice out of many educational institutions. Finally, teachers are interested in the quality of education as they gain feedback on their efforts.

Teaching quality is one of the main topics in education, and the evaluation instruments used to assess teaching quality in higher education is the focus of this dissertation. Several evaluation systems and procedures exist in higher education, but the most commonly used method is student evaluations of teaching (SETs). SETs allow students to evaluate their teachers and courses (Pruitt, Dicks, & Tilley, 2010). Results from these SETs are then regularly used for decisions such as hiring or terminating faculty members, distributing funds, and curriculum development.

The aim of this dissertation was to test the appropriateness of SETs for measuring teaching quality by methodically analyzing their reliability and validity. A reliable and valid instrument is essential for measuring teaching quality. If the instrument reliably measures teaching quality, it can foster quality improvement by providing feedback that can inform teachers about aspects of their teaching and the course that might need improvement

(Ramsden, 1996). First, reliability was assessed as a measure of interrater reliability with intra-class correlations calculated from the students, teachers, and courses variance components. Reliability is an essential requirement for the second step for analyzing the validity of SETs (Moosbrugger & Kelava, 2008). Evidence for (construct) validity was derived from several predictors that are outside of the theoretical definition of teaching quality and therefore should not predict SETs nor significantly change the variance accounted for by other components.

Research Questions

This dissertation reports results from three empirical studies that tested the appropriateness of the application of SETs for measuring teaching quality. The central research question that addressed the reliability and validity in assessing teaching quality through SETs was split into three questions:

1) Are SETs a reliable instrument for measuring teaching quality?
2) To what extent does student characteristics, such as content expectations, prior subject interest, and the perceived likability of a teacher, influence SET scores?
3) Does the time of measurement of prior subject interest and likability influence SET scores?

Structure of the Dissertation

The following sections in Chapter 1 provide a theoretical overview of quality in higher education and different concepts of teaching quality. Based on this theoretical background, the aims and scopes of this dissertation will be elaborated. Chapters 2-4 report results of three empirical studies, which have already been published in or submitted to psychological peer-reviewed journals. In Chapter 5, findings and practical implications of the three empirical studies reported in this dissertation are summarized and discussed.

Teaching, Learning, and Quality

Quality is a construct that is essentially about the compliance of requirements (Crosby, 1979). Its importance generalized from an economic context to education mainly because of the need for human capital in the industrialized world (Woodhall, 1987). One way to increase the quality of human capital is to employ educated people. The main constructs of education are teaching and learning. They are defined in the next paragraph followed by possible options how quality can be applied to education.

Teaching and learning. The operational definition of teaching for this dissertation is "(1) There is a person, T, who possesses some (2) content, C, and who (3) intends to convey or impart C to (4) a person, S, who initially lacks C, such that (5) T and S engage in a relationship for the purpose of S's acquiring C" (Fenstermacher & Richardson, 2005, p. 189). Moreover, teaching is a process (Bligh, 1993; Vervecken, Ulrich, Braun, & Hannover, 2010) or an activity (Goodyear, 2015) that helps somebody learn. In more technical terms, teaching is a process that comprise several subtasks such as planning, imaging, explaining, instructing, advising, creating learning situations, marking assessments, and giving feedback. The teacher needs to arrange these subtasks in such a way that they lead to successful learning outcomes (Goodyear, 2015).

In addition to the definition of teaching that mainly focuses on the teacher, the term learning places the focus on the student. Without learning, teaching would be a directionless act. In this dissertation, learning is defined as understanding some content (Ramsden, 1996) mostly through the actions or explanations of teachers (Bligh, 1993).

Quality in the educational context. Quality, independently seen from an educational context, is a rather vague concept (e.g., Harvey & Green, 1993; de Weert, 1990) that was transferred from the economy to other domains (Green, 1995). Connecting quality to any activity is equal to validating or justifying this activity (Harvey & Green, 1993). In the educational context, for example, the Hessian Universities Act (Hessisches Hochschulgesetz [HHG], 2004, § 27 Abs. 4 and § 92 Abs. 2) requires universities to frequently write reports about the success of teaching. Normally, success equates to quality, but quality is an unspecific term that is context dependent and relative to the stakeholder. It can convey different meanings to different stakeholders (Harvey & Green, 1993) depending on which aspects of quality are most important (de Weert, 1990). Even more confusing is the influence of context: A person can adopt different conceptualizations of quality at different times (Harvey & Green, 1993).

To broaden the discussion above, quality transferred to the educational context can mean teaching quality, and teaching quality can convey different quality concepts (see Ellis, 1995; Green, 1995; Harvey & Green, 1993). For example, quality can be understood as excellence that distinguishes the highest performance, or quality can be equated to standards that are based on predefined requirements. In the context of higher education, good teaching is often associated with teaching excellence (Harvey & Green, 1993). However, this dissertation uses the concept of standards. The reason to not use the term excellence is that teaching excellence is rarely attainable (Green, 1995; Harvey & Green, 1993), and it could imply that improvement is not necessary. This dissertation focuses more on teaching quality as a construct that can be measured and used to facilitate improvement of teaching. As defined by current standards, teaching quality can be measured in various ways. If the instrument for measuring teaching quality is developed in a way that it captures the full range of quality (i.e., from poor to excellent), it can still be used to assess excellent teaching, if the situation warrants it.

To measure teaching quality across all types of teaching situations, achievable standards (Harvey & Green, 1993) or goals (de Weert, 1990) need to be applied to be able to detect necessary improvements in teaching. An important issue in this approach is to determine appropriate standards for assessing teachers. Two possible methods have been discussed in the literature. First, common standards can be established for all institutions of higher education, which allows for comparisons across different universities (Harvey & Green, 1993). A disadvantage of this approach is the high ambiguity in finding a universal set of standards for good teaching (Ory & Ryan, 2001; Penny, 2003). The second approach favors standards that are tailored to the institution, which follows the American way of universities finding their own niche of excellence. According to Crawford (cited in Harvey & Green, 1993), each university needs to internally negotiate their standards, and teaching should be compared to these standards. This approach would require every university to develop an organizational entity to define their standards and then develop a reliable and valid measurement of these standards. This approach prevents comparisons of teaching from different universities, because some items will reflect different criteria of quality standards (Harvey & Green, 1993). Differences in standards could even occur across different departments at one university. Globally defined standards were used in the current dissertation, because these standards are legally required (HHG, 2004), and this approach allows the generalization of research results on teaching quality to other educational institutions, and it facilitates the application of a validated instrument for the measurement of teaching quality.

Teaching Quality

After clarification of how to assess teaching quality, this section generally focuses on the concept of teaching quality. Research can also be found under equivalent terms such as teaching effectiveness (e.g., Marsh & Roche, 1997; Uttl, White, & Gonzalez, 2017), effective instruction (e.g., Feldman, 1989), instructional quality (e.g., Marsh, Fleiner, & Thomas, 1975), and teaching assurance (e.g., Carbone, Evans, & Ye, 2016). Teaching quality is commonly assumed as a multidimensional construct (e.g., Cohen, 1981; Feldman, 1989; Rindermann, 2009; Marsh & Roche, 1997; Ory & Ryan, 2001) in which different factors interact with each other (e.g., Fenstermacher & Richardson, 2005; Helmke, 2007; Rindermann & Schofield, 2001). The primary factors are the actors of a course such as the teacher and the students, followed by context factors such as time of the course and special features of the room. For example, optimal teaching conditions, such as a quiet environment are difficult to achieve. Although context factors are easy to measure, they mostly cannot be changed by a teacher, and even though experienced teachers might incorporate such context factors into their teaching after detection, this dissertation does not focus on this topic.

Even after excluding context factors, teaching quality is still a complex and difficult construct to assess. Following Fenstermacher and Richardson (2005) in their detailed distinction between different kinds of teaching goals, teaching quality can be interpreted in two ways, either based on student learning outcomes (i.e., measuring successful teaching) or the chosen teaching method and content (i.e., measuring good teaching). Given that teaching is a two-way process, the willingness and effort (Fenstermacher & Richardson, 2005), the use of the learning opportunity (Helmke, 2007), the motivation (Ramsden, 1996), the expectations (McKeachie, 1997), and the prior knowledge of the students (Helmke, 2007; Rantanen, 2013) are all potential factors that could be measured. However, the assessment of teaching quality should only be measuring factors relevant to the efforts of the teacher, not factors outside the teacher’s sphere of influence (Fenstermacher & Richardson, 2005; Helmke, 2007; Schacter & Thum, 2004). In other words, the measurement of teaching quality should concentrate on the definition of good teaching instead of successful teaching. Such efforts may be, for example, the use of different teaching methods (Helmke, 2007; Ramsden, 1996; Rantanen, 2013), setting clear goals for a course (Ramsden, 1996) and the clarity and the structure of teaching content (Rindermann, 2003).

To date, no clear, straightforward approach exists for measuring teaching quality (Marsh, 1987). Construct validity (e.g., Campbell, 1960) has been used to make decisions about the measures of teaching quality. It has also been used to assess for biasing effects. For example, Marsh and Overall (1980) examined some parameters of teaching quality, such as organization and breadth of coverage, as criteria that indicate good teaching. In contrast, criteria have also been excluded that do not correlate with ratings of teaching quality, such as teacher attractiveness (Wolbring & Riordan, 2016).

Measurement of teaching quality. Teaching quality can be assessed from different perspectives, for example, by (1) self-evaluation of the teacher (Feldman, 1989), (2) SETs (Marsh, 1982a), (3) expert or neutral evaluation (de Weert, 1990), (4) peer evaluation (Rindermann, 2003), or (5) evaluation completed by administrative staff (Feldman, 1989). For an extensive summary about the effects of these different perspectives see Marsh and Roche (1997). The evaluation of teaching in higher education and the involvement of students as evaluators is legally required (HHG, 2004). Thus, factors such as costs and administrative effort might be considered in choosing which perspective to use. SETs are the most common and practical approach in universities for regular course evaluations. They allow students to give feedback and to participate in improving teaching in higher education.

A major criticism of SETs is that they merely measure student satisfaction and therefore are invalid assessments of teaching quality (Pötschke, 2010; Prosser, 2011; Spooren, Brockx, & Mortelmans, 2013; Uttl et al., 2017). According to Helmke’s (2007) offer-use model (Angebots-Nutzungs-Modell), teaching is an offer that students can take or refuse. Following this assumption, measuring satisfaction is quite useful, because the satisfaction level might indicate whether the teaching offer was used by the students. Presumably, satisfied students who make the effort to follow the teaching will learn more than an unsatisfied student. However, research has found that students taught by highly rated teachers, and therefore assumingly good teachers, sometimes achieved lesser grades in final exams than students taught by less highly rated teachers (Weinberg, Hashimoto, & Fleisher, 2009; Yunker & Yunker, 2003). In contrast, other researchers found that students from lower rated teachers, and therefore probably more unsatisfied students, take the effort to learn from other sources (Perry, 1995) to compensate for the lack of expected input. These findings suggest that SETs might not measure teaching quality, because high quality teaching is assumed to lead to more success in learning than low quality teaching (e.g., Overall & Marsh, 1979).

Although different methods can be used to measure teaching quality, this dissertation focused on the students' perspective and more precisely the application of SETs for several reasons: (1) SETs are easily administered, and they are widely used throughout the world, including in Germany, (2) The method allows students to give feedback on teaching, and (3) Results of this research can generate knowledge for improving SETs.

Research on student evaluations of teaching. Extensive research can be found on SETs that either support or oppose their application for measuring teaching quality (Clayson & Haley, 1990; Dresel & Rindermann, 2011; Dziuban & Moskal, 2011; Earley & Porritt, 2014; Greenwald, 1997; Marsh, 2007). The majority of research has focused on analyzing the reliability and validity of SETs. The reliability research addresses the consistency of teacher and course evaluations. It is a requirement for the validity of SETs. The validity research analyses the extent that a SET questionnaire measures teaching quality, which is the only construct the SET should be measuring (Ory & Ryan, 2001).

Research on reliability focuses on the consistency with which SETs measure teaching quality (Bühner, 2004). A well-constructed SET questionnaire should reliably measure teaching quality in different teaching situations (Moosbrugger & Kelava, 2008) assuming most of the variance is due to teachers and not due to other possible variance sources such as students. In the past, different approaches were used to derive the reliability of SETs. For example, correlative designs, such as different combinations of teachers and courses, showed that the teacher is a main variance source in SETs (Marsh & Roche, 1997; Rindermann & Schofield, 2001). In contrast, Gillmore, Kane and Naccarato (1978) calculated generalizability coefficients from an Analysis of Variance with random factors to deal with different sources of measurement error (e.g., intraindividual variation of students and teachers or working environment, Feldt & Brennan, 1989) in SETs. They reported a large proportion of variance due to the teacher and the interaction of teacher and course, but only a small part of the variance due to the course. The most recent approach in reliability research is the use of cross-classified multilevel analyses (Baayen, Davidson, & Bates, 2008) that produces a straightforward partitioning of total variance in its different variance components. This kind of analysis allows to account for the multilevel structure of evaluation data and additionally to estimate the student variance component. For example, Rantanen (2013) and Spooren (2010) found in Finnish and in Belgian samples nearly 25% of the variance attributed to the teacher, whereas another ca. 50% of the variance could not be explained by the student, teacher or course variance components. Such high unexplained variance is quite alarming considering that SET results could impact the future employment of a teacher.

Reliability is a necessary but not sufficient condition for validity. Even with evidence of acceptable reliability, an instrument’s validity still needs to be demonstrated. Following Messick (1989), validity is a cumulative support of empirical evidence on theoretical assumptions that leads to logical conclusions (Olivares, 2003). In the context of SETs, validity is frequently based on “bias studies” (Ory & Ryan, 2001, p. 31) in which a biasing effect is a positive or negative influence of student, teacher, or course characteristics of SETs, even if the characteristic are theoretically unrelated to teaching quality (Centra & Gaubatz, 2000). Some examples of biasing effects are listed in Table 1.1. For an extensive overview of biasing effects, see Greenwald (1997), Marsh and Roche (1997), Pounder (2007), Spiel and Gössler (2000), or Spooren et al. (2013).

Table 1.1

Examples of biasing effects reported in the literature

Biasing effect	Level of observation	Authors
Course difficulty	Course	Remedios and Lieberman (2008)
Course size	Course	Wood, Linsky, and Straus (1974)
Course format	Course	Freeman (1994)
Expectations	Student	Pruitt et al. (2010)
Expected grade	Student	Marsh and Overall (1980)
Non-response	Student	Rantanen (2013)
Pre-course motivation	Student	Howard and Schmeck (1979)
Prior Subject Interest	Student	Olivares (2001)
Attraction	Teacher	Wolbring and Riordan (2016)
Grade leniency	Teacher	Greenwald and Gillmore (1997)
Likability	Teacher	Delucchi (2000)
Personality	Teacher	Clayson and Sheffet (2006)

However, validity is a complex construct. Single findings provide insufficient evidence to draw conclusions about such a complex construct. The sum of several empirical studies, however, allows for conclusions about the validity of SETs. As Ory and Ryan (2001) stated, “validity is not an all-or-nothing issue” (p. 28), because it comprises several components such as content-, criterion- and construct-related validity. Given its complexity, structuring validity and the empirical evidence in a theoretical validity framework (e.g., Olivares, 2003; Onwuegbuzie, Daniel & Collins, 2009; Ory & Ryan, 2001) helps to organize the different findings and draw conclusions about the validity of SETs in total. Most researchers conclude that SETs are a valid measurement (e.g., Dresel & Rindermann, 2011; Marsh & Roche, 1997) of teaching quality, whereas others speak (vehemently) against their use (e.g., Clayson, 2017; Olivares, 2003; Spooren et al., 2013).

Aim of Research

In the previous sections, the literature about teaching quality and research on the reliability and the validity of SETs was discussed. Although, research on this topic has been extensive, it has not yet addressed all possible biasing effects. Some research questions remain. For example, are the results from other countries also valid in Germany? Germany has a slightly different educational system compared to some other countries, because lectures are attended by numerous students, seminars can be held in a presentation format, and students evaluate their courses before writing their exams in the course. Such different teaching settings might influence the results of SET research. Research on German SETs, as an instrument to assess teaching quality, has already focused on biasing effects such as prior subject interest (Dresel & Rindermann, 2011; Fondel, Lischetzke, Weis, & Gollwitzer, 2015; Staufenbiel, Seppelfricke, & Rickers, 2016), prior knowledge (Dresel & Rindermann, 2011), student’s effort (Dresel & Rindermann, 2011; Rindermann, Kohler, & Meisenberg, 2007), the teacher’s attractiveness (Wolbring, 2010; Wolbring & Riordan, 2016), the gender of students and teachers (Dresel & Rindermann, 2011; Staufenbiel et al., 2016; Wolbring & Riordan, 2016), the size of the course (Staufenbiel et al., 2016), student’s class attendance (Staufenbiel et al., 2016; Wolbring, 2012; Wolbring & Treischl, 2016), student’s age (Dresel & Rindermann, 2011), difficulty of the course (Fondel et al, 2015), and likability of the teacher (Fondel et al., 2015). Additionally, the difference of course formats have been typically considered (e.g., Dresel & Rindermann, 2011; Fondel et al., 2015; Staufenbiel et al., 2016).

Course type is an important factor, because Spiel and Gössler (2000) found that biasing effects were twice as strong in lectures as in seminars.

This dissertation advances the ongoing discussion about the reliability and validity of SETs as a measure of teaching quality by using a cross-classified multi-level-approach (mixed models with crossed random effects, Baayen et al., 2008). This approach differs from prior German research that has already used a hierarchical multi-level-approach (e.g., Dresel & Rindermann, 2011; Staufenbiel et al., 2016) by incorporating the student variance component. If this variance component is not accounted for, the students' variance is confounded with the residual variance (e.g., measurement error) of the single questionnaire. Thus, individual difference variables should be included in the analysis, because they are an important source of variance, for example, analyzing leniency effects while rating (Bernardin, 1978; Wolfe, 2004). This inclusion allows to check for the influence of bias variables on the variance that is solely attributed to the student. Analyzing the student variance component requires coding multiple SET questionnaires to the same student.

The following empirical studies of this dissertation were based on the standardized German evaluation questionnaire for university courses FEVOR (Fragebogen zur Evaluation von universitären Lehrveranstaltungen durch Studierende [Students Course Assessment Questionnaire for Evaluation of University Courses], Staufenbiel, 2000; Staufenbiel et al., 2016). The FEVOR questionnaire is a scientifically validated questionnaire that allows evaluations in different course formats such as lectures and seminars. An individual alphanumeric code was added to the top of the questionnaire for linking multiple questionnaires completed by the same student. The resulting ID could not be linked to the students, thus protecting their anonymity. The FEVOR questionnaire was constructed with multiple dimensions like many other questionnaires (e.g., HILVE II, Rindermann & Schofield, 2001; SEEQ, Marsh 1982b). The instrument includes four scales (planning and presentation, interactions with students, interest and relevance, and difficulty and complexity) and two global ratings (ratings of the teacher performance and the course). Interestingly, most of the previous studies based their conclusions about the reliability or validity of teaching quality on only one dimension (e.g., Rantanen, 2013; Rindermann & Schofield, 2001; Spooren, 2010). Only a few researchers used the widely accepted multidimensionality (e.g., Centra, 1993; Frazer, 1995; Marsh 1983; Marsh & Roche, 1997; Rindermann, 2009;

Rindermann & Schofield, 2001) when analyzing the relationships between potential biasing effects and the dimensions of teaching quality (e.g., Marsh, 1982a, Staufenbiel et al., 2016). The following studies focused either on all dimensions when drawing conclusions about the reliability or at least on two dimensions (planning and presentation and the global rating of the teacher performance) when concluding about the validity.

The first study of this dissertation, presented in Chapter 2, focused on the reliability of the questionnaire as a first step. The interrater reliability was calculated based on the variance components of cross-classified multilevel models (Baayen, et al., 2008). This type of analysis enables to split the total variance into teacher, student and course components and can be used as the basis for calculating the intra-class correlation (Shrout & Fleiss, 1979).

As stated in the introduction, reliability is an important requirement to be able to conclude SETs as a valid measure of teaching quality. To classify a SET instrument as reliable this dissertation followed the suggestion of Marsh and Roche (1997) who set a normative cut-off of an intra-class correlation (Shrout & Fleiss, 1979) of .20 between two randomly chosen students. This limit is justified by the following three assumptions: First, normally more than two students evaluate a teacher and a course. Second, the Spearman-Brown-Formula (Brown, 1910; Spearman, 1910) allows to calculate the necessary number of student to reach a widely accepted interrater reliability of .90, and third, the number of students necessary for evaluating one course must be smaller than the actual number of students attending this course to allow a randomized sample of students. Application of the Spearman-Brown-Formula on an interrater reliability of .20 led to 25 students who need to be randomly chosen to evaluate a course (Marsh & Roche, 1997). Considering that some students in German courses leave the course before the time of evaluation starts, only lectures and seminars (usually 40 students attend a seminar) can fulfil the necessary course size. Lectures and seminars are also the most frequent types of courses offered. Therefore, the focus of this dissertation was on these two types of courses.

The latter two studies of this dissertation assessed the validity of the FEVOR questionnaire by introducing potential biasing effects in the analyses to gain an insight on their influence on SETs. The second study in Chapter 3 focused on the effects of students' expectations of course content, operationalized via self-reported general and specific preconceptions of psychology, and prior subject interest on SETs. Given the high interest in studying psychology in the bachelor program (e.g., Fisch, Orlik, & Saterdag, 1970), a large student sample is available for investigating individual differences. Some students start their studies directly after finishing high school, whereas others already have children or have served in the military. Some of these students are interested only in special modules (e.g., clinical psychology) of psychology as a step to further education while others are interested in the whole spectrum taught in the science psychology. These students are likely to start their studies with different expectations and interests about psychology. Their expectations and prior subject interest are outside of the teacher's sphere of influence. Thus, they should theoretically have no relationship with SETs, thereby establishing SETs as a valid measure of teaching quality. Although, research exists on different expectations, such as expected grades (e.g., Holmes, 1972; Marsh, 1980; Feldman, 1997) or expectations regarding teachers (Bejar & Doyle, 1976; Pruitt et al., 2010), the field of content expectations has received little attention in SET research. Spiel and Gössler (2000) suggested that content expectations might have a possible confounding effect on SETs. In contrast, prior subject interest is a widely researched bias, and for this reason it is an item in the FEVOR. Its importance as a biasing effect on SETs notwithstanding, results from prior studies have been inconsistent (e.g., Feldman, 1976; Kromrey, 1994a; Marsh, 1980; Olivares, 2001; Spiel & Gössler, 2000; Staufenbiel et al., 2016; Uttl et al., 2017; Wolbring & Treischl, 2016). To augment the findings in this research area and to support an informed conclusion about the validity of SETs, the second study focused on content expectations and prior subject interest as possible biasing effects on the FEVOR.

In the third study (Chapter 4), the focus was on two topics to strengthen the evidence for the validity of SETs. The two potential biasing effects, likability of the teacher perceived by students and prior subject interest, were included, and the measurement time was varied. Likability was chosen in this study, because some researchers see SETs as a popularity or a likability contest (Clayson & Haley, 1990; Dziuban & Moskal, 2011; Uranowitz & Doyle, 1978) instead of a valid measurement of teaching quality. Former research has supported this concern by showing an alarmingly high influence of likability on SETs (Clayson & Sheffet, 2006; Fondel et al., 2015; Delucchi, 2000). Given that retrospective measurements are commonly used to assess biasing effects (e.g., Marsh, 1982b; Staufenbiel et al., 2016), the third study examined the assumption that retrospective measurements of likability and prior subject interest is influenced by unknown factors in the time between the beginning of the course and the time of evaluation. Consequently, this influence could result in invalid scores from retrospective measurements (Döring & Bortz, 2016). Study 3 compared the two biasing effects likability and prior subject interest once measured at the beginning of the course and at the time of evaluation near the end of the course. This design allowed to check whether an effect was actually a biasing effect even when the construct was measured at the correct time, and it allowed to assess whether results of the retrospective measurement of the construct were over- or underestimated and therefore influenced by other factors.

CHAPTER 2 Study 1 -

How reliable are students’ evaluations of teaching quality? A variance components approach

How reliable are students’ evaluations of teaching quality?
A variance components approach

Daniela Feistauer & Tobias Richter

Published in Assessment & Evaluation in Higher Education, 42, 1263-1279.

Abstract. The interrater reliability of university students’ evaluations of teaching quality was examined with cross-classified multilevel models. Students (*N* = 480) evaluated lectures and seminars over three years with a standardized evaluation questionnaire, yielding 4,224 data points. The total variance of these student evaluations was separated into the variance components of courses, teachers, students, and the Students x Teachers interaction. The substantial variance components of teachers and courses suggest reliability. However, a similar proportion of variance was due to students, and the interaction of students and teachers was the strongest source of variance. Students’ individual perceptions of teaching and the fit of these perceptions with the particular teacher greatly influence their evaluations. This casts some doubt on the validity of student evaluations as indicators of teaching quality and suggests that aggregated evaluation scores should be used with caution.

Keywords: cross-classified multilevel analysis, interrater reliability, student evaluations of teaching, variance components

Teaching quality is an important criterion to assess in higher education, for example, to identify improvement needs or to justify tuition costs. Students and their parents demand high-quality teaching, and teachers and department heads need good measures of teaching quality. Its importance begs the question of whether current teaching assessments provide reliable data of teaching effectiveness. This study examines the ratings of students on the quality of single courses, also referred to as teaching effectiveness (e.g., Gillmore et al., 1978; Marsh, 1984, 2007). The construct of teaching effectiveness comprises several facets of good teaching, including how teachers perform in communicating with students, organizing a course and its contents, stimulating interest, and behaving friendly and tolerantly (Hattie & Marsh, 1996).

Teaching quality can be measured in different ways and by tapping different sources. The most common way to measure teaching quality in higher education is through student evaluations of teachers (e.g., Marsh, 1984; Spooren, 2010; Rantanen, 2013). In most cases, student evaluations of teaching quality are obtained via evaluation questionnaires (e.g., Students' Evaluations of Educational Quality (SEEQ), Marsh, 1982b, Marsh et al., 2009;

Student Course Experience Questionaire (SCEQ); Ginns, Prosser, & Barrie, 2007; FEVOR/FESEM (Fragebogen zur Evaluation von universitären Lehrveranstaltungen durch Studierende [Students Course Assessment Questionnaire for Evaluation of University Courses]), Staufenbiel, 2000). All of these questionnaires mirror the above-mentioned multidimensionality of teaching quality by differentiated scales.

This study focuses on the interrater reliability of student's evaluations of teaching quality (e.g., Marsh, 1984, 2007; Rantanen, 2013). Reliability is a fundamental criterion of student evaluations of teaching and a necessary (though not sufficient) precondition for their validity as indicators of teaching quality. The most common measure of interrater reliability for interval-scaled ratings is the intra-class correlation (ICC; Shrout & Fleiss, 1979). The ICC for teaching evaluations is defined as the correlation between assessments of randomly determined pairs of students evaluating the same course or teacher. It may also be regarded as the proportion of the total variance of student evaluations that can be explained by courses and teachers. In terms of Kenny's (1994) social relations modeling, the variance components of courses and teachers can be seen as target variance, that is, the degree to which different perceivers (students) rate the target (the course or the teacher) in the same way. Reliability is maximized by maximizing the proportion of target variance.

Marsh (1982a) and Rindermann and Schofield (2001) addressed the issue of interrater reliability by comparing different combinations of teachers and courses. For example, they compared one teacher giving several courses with one course taught by different teachers. An evaluation questionnaire was deemed as more reliable (and also more valid) when there was a higher effect of teachers in parallel courses with the same content compared to courses covering different content. They inferred from correlational analyses that teachers have a high impact on teaching evaluations. Gillmore et al. (1978) estimated variance components via an Analysis of Variance with random factors. They found that the variance explained by courses (6%) was much smaller than that explained by teachers (40%). Moreover, they found a substantial effect of the interaction of teachers and courses. In most studies, the ICC for student evaluations of courses or teachers is about .20 (e.g., Marsh, 1984; Solomon, Speer, Rosebraugh, & DiPette, 1997). An ICC in this order of magnitude indicates a modest reliability of individual evaluations but corresponds to an acceptable or even high reliability when means of evaluations from a considerable number of students are considered (e.g., .90

for 25 students, Marsh & Roche, 1997).

The reviewed studies did not examine the influence of students on course evaluations and compare this influence with those of teachers or courses. The modern analytical approach of multilevel analysis (linear mixed models) can address this question in a straightforward way (Raudenbush & Bryk, 2006; Richter, 2006). Multilevel analyses can account for the multilevel structure of evaluation data. Typically, students evaluate several courses and teachers, teachers teach several courses, and the same courses are taught by different teachers, yielding an imperfect hierarchy or *cross-classified* data structure. Such data structures call for models with crossed random effects (Baayen et al., 2008), also known as cross-classified or non-hierarchical multilevel-models (Rasbash & Browne, 2008). These models estimate the effect sizes of teachers, courses, students, and also their interactions, provided that a sufficient number of different combinations of units from different levels are obtained. From the estimates of the variance components, the ICCs (variance proportions) can be calculated for student evaluations of courses or teachers and for all other sources of variance included in the model (such as students or interactions of students and teachers).

Several authors (Spooren, 2010; Rantanen, 2013; Staufenbiel et al., 2016) have already adopted the approach of estimating hierarchical or cross-classified multilevel analyses. Despite methodological differences (such as different questionnaires used), these studies yielded estimates of variance components for students, courses, and teachers of similar magnitude. They found nearly one quarter at the student level and one quarter at the teacher and course level, whereas about 50% of the variance remained unexplained. More details are available in Figure 2.1.

All of the previous studies focused on only one dependent variable. Staufenbiel et al. (2016) used an average score of three scales capturing three different facets of teaching effectiveness (planning and presentation, interaction with students, and interestingness and relevance). Rantanen (2013) used a mean score based on five items (teacher’s expertise in the subject, teaching skills, visual aids, interaction with the students, and learning assignments). Spooren (2010) calculated a rescaled global factor of seven factors (e.g., clarity of course objectives and value of subject-matter). Yet, to our knowledge, no study decomposing the total variance of student evaluations in its different components (teachers, courses, and students) has taken multiple dimensions of teaching effectiveness into account.

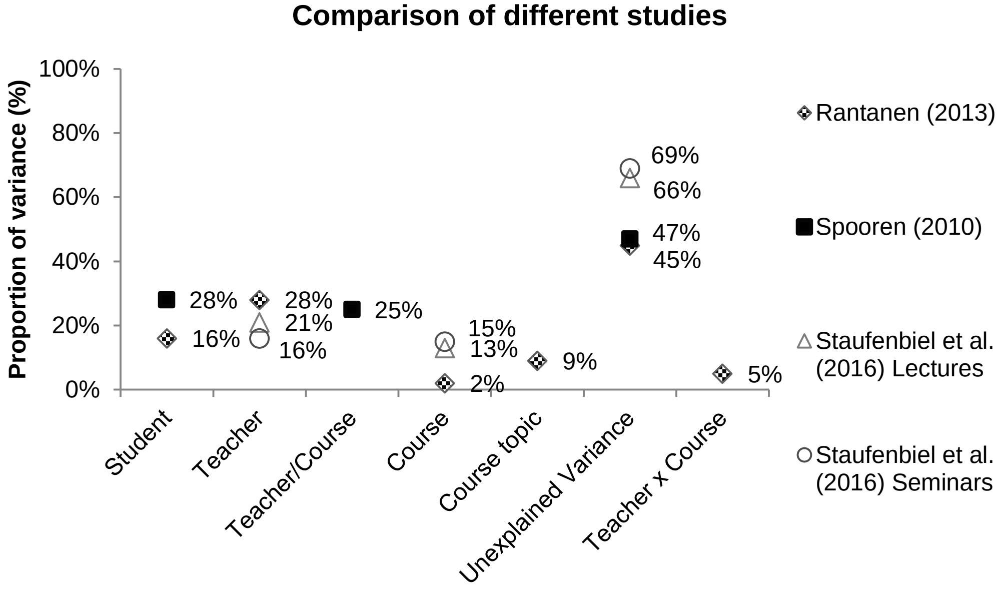

Figure 2.1. Estimates of variance components estimated in previous studies. The study by Rantanen (2013) included the interaction of teachers and courses. The study by Spooren (2010) did not distinguish between teachers and courses as sources of variance.

Rationale of the Present Study

The present study was based on a psychometrically sound questionnaire that is used in German-speaking countries for student evaluations of teaching in higher education (Staufenbiel, 2000; Staufenbiel et al., 2016). The questionnaire assesses four different aspects of teaching quality (planning and presentation, interaction with students, interestingness and relevance, difficulty and complexity) and contains two global ratings of the quality of the entire course and the teacher. We analyzed the data using a cross-classified multilevel analysis (mixed models with crossed random effects, Baayen et al., 2008), including random effects of all three possible sources of variance: teachers, courses, and students. Moreover, we ran separate analyses for lectures and seminars because of the didactical and organizational differences of the two course formats (cf. Staufenbiel et al., 2016). Our aim was to provide an in-depth analysis of the interrater reliability of the different scales and the global ratings included in the questionnaire by Staufenbiel (2000) and to compare the variance components that contribute to a high reliability (teachers and courses) with variance components that are due to students and the interaction of students and teachers.

This analysis tackled three novel research questions. First, we examined whether the variance components of teachers, courses, and students would differ systematically between different aspects of teaching quality. Conceptually, the evaluations regarding planning and presentation of the course contents might be influenced more strongly by teacher behaviour than evaluations regarding the aspects of interaction with students, interestingness and relevance, and difficulty and complexity. For the assessment of the latter three aspects, students' behaviour, interests, and abilities are also likely to play a major role. For example, a student with strong prior knowledge might find the same course easier than a student with weak prior knowledge. Likewise, students might find courses interesting and relevant that match their personal interests, which can differ between students. Second, we investigated potential differences in the variance components in models for lectures and seminars. Lectures are a much more teacher-centered course format than seminars, which include active contributions from students in the form of class-room presentations and discussions, implying that teachers should be a stronger source of variance in lectures. Third, we extended previous multilevel studies by including the interaction of students and teachers as an additional random effect in the model. By including this interaction, we were able to account for the possibility that student evaluations of teaching might depend systematically on whether student characteristics (such as their expectations, abilities and interests) match characteristic of the teacher giving the course (such as their expectations, teaching styles, or level of difficulty). Our expectation was that including this interaction, as another source of systematic variance in the model, would considerably reduce the large amount of unexplained variance found in previous studies (typically around 50%).

Method

Sample

The present study used a data set of 4,224 evaluations (questionnaire data) of psychology courses held between the winter semester 2011 and the summer semester 2014 at the University of Kassel, Germany. During this period, courses were taught by 53 different teachers (30 women). Sixty lectures were given by 18 different teachers and 115 seminars were held by 49 different teachers. Of the 53 teachers, 19 teachers (36%) were professors on the associate/full professor level (tenured or visiting), 13 (24%) were assistant professors or post-doctoral lecturers, and 21 (40%) were doctoral students holding a position as researcher or lecturer. The evaluations were provided by 480 students (73.2% women) who attended the courses. The students remained anonymous but evaluations provided by the same student could be identified by an individual code. On average, each student evaluated seven courses (lectures: *Md* = 4, *Range* = 1-18, seminars: *Md* = 4, *Range* = 1-17). The sample of courses comprised courses on research methods such as statistics and content courses like educational, cognitive, social or clinical psychology. Of all courses, 145 (83%) were bachelor-level courses (BSc program in Psychology) and 30 (17%) were master level courses (MSc programs in Psychology).

Procedure

The questionnaires were administered in the last third of each semester (in the second half of January in winter semesters and the second half of June in summer semesters). The teachers handed out the questionnaires. The students were given 5-10 minutes to complete the questionnaires. Data were scanned with the program Remark Office OMR 8. Accuracy of data scanning was controlled by a student research assistant. Each teacher received elaborated reports for their courses to discuss with the students in the last session of the semester.

Measures

The study was based on a standardized questionnaire that is widely used in Germany for the evaluation of university courses (FEVOR/FESEM, Staufenbiel, 2000; Staufenbiel et al., 2016). There are different versions of the questionnaire for different course types. The version for lectures contains 31 items and the version for seminars 34 items. Apart from 26 parallel items, eight seminar items on discussions and presentations held by students and four lecture items on the presentation style were included. In the header of the questionnaire, students provided an individual alphanumerical code that allowed assigning different questionnaires to one student while protecting students' anonymity. The questionnaire assesses students' evaluations of university courses on three psychometrically distinct unipolar scales that are based on Likert-scaled items (ranging from 1 = *strongly disagree* to 5 = *strongly agree,* and "*not applicable*" as an additional response option). On the bipolar difficulty and complexity scale the response options range from 1 = *much too low* to 5 = *much too high*.

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 eight items in the version for seminars (*M* = 4.13, *SD* = 0.63, Cronbach’s α = .78) and five items in the version for lectures (*M* = 4.13, *SD* = 0.70, Cronbach’s α = .90).

Interaction with students. This scale assesses the perceived respect and concern that teachers show to their students. It contains items such as “There is a good working climate in the seminar” and “The lecturer seems to care about the students’ learning success”. The scale consists of four items in the version for seminars (*M* = 4.35, *SD* = 0.63, Cronbach’s α = .79) and three items in the version for lectures (*M* = 4.38, *SD* = 0.66, Cronbach’s α = .77).

Interestingness and relevance. This scale measures how useful students perceive the contents of the course for other contexts and how interesting they find the content. The scale consists of items such as “The lecturer encourages my interest in the subject area” and “The lecturer makes the lecture interesting”. The scale consists of four items in the version for seminars (*M* = 3.94, *SD* = 0.76, Cronbach's α = .85) and lectures (*M* = 4.03, *SD* = 0.73, Cronbach's α = .80).

Difficulty and complexity. This scale measures the perceived difficulty, scope, and pace of the course. The scale consists of items such as “The pace of the seminar is:”; *M* = 3.17, *SD* = 0.42, Cronbach’s α = .74 for seminars and *M* = 3.29, *SD* = 0.51, Cronbach’s α = .77 for lectures.

Ratings of teacher and course. Respondents also rated the teacher’s overall performance and the quality of the course on a general level. Ratings were provided according to the German grading system that ranges from 1 = *very good* to 5 = *poor* (teachers: *M* = 1.88, *SD* = 0.82 for seminars, *M* = 1.81, *SD* = 0.80 for lectures; course: *M* = 2.11, *SD* = 0.79 for seminars; *M* = 2.06, *SD* = 0.76 for lectures).

Results

Analyses were performed with cross-classified linear mixed-effects models (Baayen et al., 2008) that allowed separating the variance components of teachers, courses, and students. These three sources of variances were included as random effects in the analysis. Separate models were estimated for the four scales and the two overall ratings of the evaluation questionnaire by Staufenbiel (2000). The models were estimated with the statistical software R Version 3.2.2 (R Core Team, 2015) and the full Maximum Likelihood estimation procedure built-in the function lmer of the R-package lme4 (Bates, Mächler, Bolker, & Walker, 2015). The significance of each random effect was tested with the function called anova of the R-package stats that compares the fit of models that differ in their random effects structure (R Core Team, 2015). Data were analysed separately for lectures and seminars because of the heterogeneous course format and the differences in course sizes.

Based on the random effects models, ICCs were computed that reflect the proportions of variance due to students, teachers and courses. The ICCs for teachers and courses may be interpreted as measures of interrater reliability (absolute agreement between students) of the evaluations of teachers and courses. For example, the proportion of variance due to teachers reflects how reliably students were able to assess teachers with regard to a certain criterion (e.g., an overall rating or score on one of the four evaluation scales):

$$ICC_{teachers} = \frac{\sigma^2_{teachers}}{\sigma^2_{total}} \quad (1)$$

High values of ICC (i.e., values close to 1) indicate that students strongly agreed in their assessments of the teachers and that different courses given by the same teacher were judged similarly. Low values of this ICC (i.e., values close to 0), indicate that students, classes, or both differed in their evaluations, implying that the evaluations are not a reliable assessments of teachers.

Figure 2.2 illustrates the structure of the cross-classified multilevel models. The first model (Model 1) included for all four scales and two ratings random effects (random intercepts) of students and courses:

$$Y_{sc} = \theta_0 + h_{0s} + i_{0c} + e_{sc} \quad (2)$$

In this model, Y_{sc} represents the evaluation score of student *s* in courses *c*. The intercept θ_0 represents the grand mean of this score across all students and courses. The term h_{0s} captures the individual deviation of student *s* from θ_0, that is, the contribution of this student across all courses. Likewise, i_{0c} represents the deviation of course *c*, that is, its contribution across all students. The individual deviations of students and courses are conceptualized as random effects that follow a normal distribution with a mean of 0 and the variances τ_{s0} and τ_{c0}. Finally, the model includes the error term e_{sc} that captures unsystematic errors in the evaluation scores that remain after the random effects of students and courses 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).

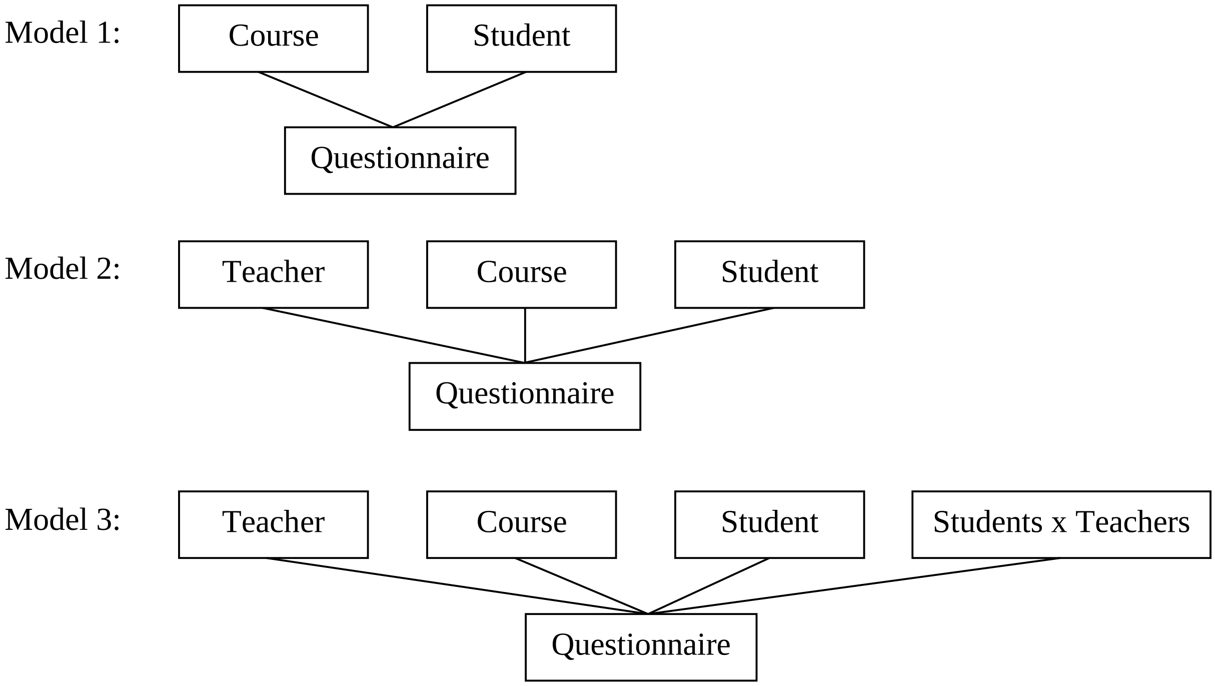

Figure 2.2. Random effects included in the three models estimated for student evaluations of teaching.

Model 1 allows estimating the interrater reliability for course evaluations by separating the systematic variance due to courses from the systematic variance due to students (e.g., Spooren, 2010) plus a component of unsystematic error variance. However, courses are confounded with teachers, who may be regarded as a separate source of variance. Thus, we estimated a second model (Model 2) that disentangles the contributions of teachers and courses to the total variance. This model included random effects (random intercepts) of students, teachers, and courses.

$$Y_{sct} = \theta_0 + h_{00s} + i_{00c} + j_{00t} + e_{sct} \qquad (3)$$

In addition to the effects contained in Model 1, this model includes the term j_{00t} that represents the deviation of teacher *t* from the grand mean, that is, his or her contribution to the evaluations across all courses and students taught by this teacher. Again, the effects of teachers are assumed to be random effects that are normally distributed with mean 0 and variances τ_{s00}, τ_{c00}, and τ_{t00}.

Finally, differences in course evaluations might depend systematically on who evaluates whom, that is, on an interaction of students and teachers. Therefore, we estimated a third model (Model 3) that included another random effect reflecting the interaction of students and teachers:

$$Y_{sct} = \theta_0 + h_{00s} + i_{00c} + j_{00t} + k_{0st} + e_{sct} \qquad (4)$$

In this model, k_{00st} represents the random deviations of specific student-teacher combinations *st* from the grand mean; all other terms are the same as in Model 2 (Equation (3)). The variance of these deviations, assumed to be normally distributed with a mean 0 and variances τ_{s00}, τ_{c00} τ_{t00}, and τ_{st0}, captures the variance due to the two-way interaction effect of students and teachers.

In principle, including interactions of students and courses or interactions of courses and teachers would also be possible. However, such interactions could not be estimated with the present data set, because the frequencies of different student-course combinations and teacher-course combinations were too small. For example, there was mostly only one evaluation per course and student.

Table 2.1, 2.2 and 2.3 display the variance components, variance proportions, and model fit for the models estimated for seminars and lectures. In Table 2.1 and 2.2 display the four scales, and Table 2.3 displays the overall ratings of course and teacher.

Planning and Presentation

The estimates for Models 1-3 with the scale Planning and Presentation as criterion variable are displayed in Table 2.1 (upper half). For this scale, the proportion of variance due to courses, 22% for seminars and 33% for lectures, estimated in Model 1 nearly equalled the sum of the variance components of teachers and courses in Model 2. Thus, taking the confounding of teachers and courses into account did not reduce the unexplained variance. It also did not change the proportion of variance due to students, which was 15% in seminars and 21% in lectures. The proportion of variance explained by teachers was 27% in lectures but only 6% in seminars (Model 2), reflecting the different role of teachers in seminars vs. lectures. Model 3, which additionally included the interaction of students and teachers, strongly increased variance explained. The proportion of variance explained by this interaction was 20% in seminars and 29% in lectures (Model 3). Thus, the fit of the individual students with the individual teachers determined the student's evaluations on this scale to a considerable extent. Overall, we obtained a large proportion of variance due to students and a large proportion of variance that remained unexplained in evaluations of seminars. These problems were less aggravated in the evaluation of lectures, with only 23% unexplained variance and a relatively low proportion of variance due to students (15%).

Interaction with Students

The estimates for the models with the scale Interaction with Students as criterion variable are displayed in Table 2.1 (lower half). Generally, the results for this scale are similar to the results for Planning and Presentation. Model 1 revealed a considerable proportion of variance due to students (17% for seminars and 18% for lectures) but a much higher proportion of variance due to courses (35% for seminars and 27% for lectures). Model 2 decomposed the latter variance component further in similarly large proportions due to courses and teachers. Including the interaction of students and teachers in Model 3 provided a large increment in variance explained (26% in seminars and 32% in lectures), reducing the proportion of unexplained variance to only 24% in seminars and 27% in lectures.

Table 2.1

Estimates for the Cross-classified Linear Mixed Effect Models (Random Intercepts) for the Scales Planning and Presentation and Interactions with Students

	Seminars									**Lectures**								
	Model 1: Student and Course			Model 2: Student, Teacher and Course			Model 3: Student, Teacher, Course and Students x Teachers			Model 1: Student and Course			Model 2: Student, Teacher and Course			Model 3: Student, Teacher, Course and Students x Teachers		
Planning and Presentation																		
Fixed effects	Estimate		*SE*	Estimate		*SE*	Estimate		*SE*	Estimate		*SE*	Estimate		*SE*	Estimate		*SE*
Intercept	4.131		0.033	4.133		0.040	4.133		0.040	4.132		0.057	4.070		0.097	4.067		0.097
Random effects	σ²		% σ²	σ²		% σ²	σ²		% σ²	σ²		% σ²	σ²		% σ²	σ²		% σ²
Residual	0.230		58.0%	0.231		58.1%	0.153		38.6%	0.251		50.0%	0.251		49.1%	0.117		22.5%
Student (Intercept)	0.085	***	21.4%	0.085	***	21.4%	0.084	***	21.1%	0.088	***	17.5%	0.089	***	17.4%	0.078	***	15.1%
Teacher (Intercept)				0.027	**	6.9%	0.028	**	7.1%				0.142	***	27.8%	0.142	***	27.4%
Course (Intercept)	0.081	***	20.5%	0.054	***	13.6%	0.052	***	13.1%	0.163	***	32.5%	0.029	***	5.8%	0.029	***	5.5%
Students x Teachers (Intercept)							0.080	**	20.1%							0.152	***	29.4%
Fit statistics																		
Deviance			3499.4			3492.0			3481.2			3623.2			3579.8			3460.8
AIC			3507.4			3502.0			3493.2			3631.2			3589.8			3472.8
BIC			3530.0			3530.3			3527.2			3653.8			3618.1			3506.7
Interaction with Students																		
Fixed effects	Estimate		*SE*	Estimate		*SE*	Estimate		*SE*	Estimate		*SE*	Estimate		*SE*	Estimate		*SE*
Intercept	4.385		0.039	4.399		0.050	4.399		0.050	4.364		0.049	4.421		0.073	4.419		0.072
Random effects	σ²		% σ²	σ²		% σ²	σ²		% σ²	σ²		% σ²	σ²		% σ²	σ²		% σ²
Residual	0.195		48.5%	0.195		49.3%	0.094		23.7%	0.239		55.6%	0.239		55.7%	0.114		27.1%
Student (Intercept)	0.068	***	16.9%	0.068	***	17.1%	0.066	***	16.5%	0.075	***	17.5%	0.076	***	17.7%	0.063	***	14.6%
Teacher (Intercept)				0.058	***	14.7%	0.059	***	14.8%				0.062	***	14.6%	0.069	***	15.3%
Course (Intercept)	0.139	***	34.6%	0.075	***	19.0%	0.074	***	18.6%	0.115	***	26.9%	0.051	***	12.0%	0.047	***	11.0%
Students x Teachers (Intercept)							0.105	***	26.3%							0.139	***	32.1%
Fit statistics																		
Deviance			3202.0			3183.5			3149.9			3482.0			3463.2			3332.7
AIC			3210.0			3193.5			3161.9			3490.0			3473.2			3344.7
BIC			3232.6			3221.8			3195.8			3512.6			3501.5			3378.6

Note. σ² = variance, % σ² = Proportion of variance (percentage)
* p<.05, ** p<.01, *** p<.001.

Interestingness and Relevance

The estimates for the models with the scale Interestingness and Relevance as criterion variable are displayed in Table 2.2 (upper half). For seminars, Model 1 revealed a considerable proportion of variance due to students (17%) and a much higher proportion of variance due to courses (31%). For lectures, there was also a considerable proportion of variance due to students, which was of a similar magnitude to the variance due to courses (20%). Including the interaction of students and teachers in Model 3 led to a considerable increase of variance explained (18% in seminars and 26% in lectures). Nevertheless, relatively large proportions of variance remained unexplained even in Model 3 (35% in seminars and 37% in lectures).

Difficulty and Complexity

The estimates for the models with the scale Difficulty and Complexity as criterion variable are displayed in Table 2.2 (lower half). Similar to the scales Planning and Presentation and Interactions with Students, Model 1 revealed a considerable proportion of variance due to students (15% for seminars and 20% for lectures) but a higher proportion of variance due to courses (26% for seminars and 29% for lectures). However, the decomposition of this variance component in Model 2 suggests that the variance explained by teachers (17% for seminars and 19% for lectures) was higher than the variance explained by courses (8% and 11%). Compared to the other three scales, including the interaction of students and teachers in Model 3 led to a relatively small increase of variance explained (8% in seminars and 13% in lectures). Consequently, relatively large proportions of variance remained unexplained even in Model 3 (54% in seminars and 38% in lectures).

Table 2.2

Estimates for the Cross-Classified Linear Mixed Effect Models (Random Intercepts) for the Scales Interestingness and Relevance and Difficulty and Complexity

	Seminars									Lectures								
	Model 1: Student and Course			Model 2: Student, Teacher and Course			Model 3: Student, Teacher, Course and Students x Teachers			Model 1: Student and Course			Model 2: Student, Teacher and Course			Model 3: Student, Teacher, Course and Students x Teachers		
Interestingness and Relevance																		
Fixed effects	Estimate		*SE*	Estimate		*SE*	Estimate		*SE*	Estimate		*SE*	Estimate		*SE*	Estimate		*SE*
Intercept	3.989		0.045	4.015		0.059	4.016		0.059	4.027		0.049	4.042		0.065	4.040		0.066
Random effects	σ²		% σ²	σ²		% σ²	σ²		% σ²	σ²		% σ²	σ²		% σ²	σ²		% σ²
Residual	0.296		51.8%	0.296		51.8%	0.197		34.4%	0.317		60.1%	0.317		61.1%	0.193		36.9%
Student (Intercept)	0.096	***	16.8%	0.095	***	16.7%	0.093	***	16.2%	0.103	***	19.6%	0.103	***	19.9%	0.091	***	17.4%
Teacher (Intercept)				0.092	***	16.1%	0.094	***	16.5%				0.037	***	7.1%	0.040	***	7.7%
Course (Intercept)	0.179	***	31.4%	0.088	***	15.4%	0.086	***	15.0%	0.107	***	20.3%	0.061	***	11.9%	0.061	***	11.6%
Students x Teachers (Intercept)							0.103	***	18.0%							0.138	***	26.4%
Fit statistics																		
Deviance			4048.5			4027.6			4016.5			4064.6			4052.0			3983.7
AIC			4056.5			4037.6			4028.5			4072.6			4062.0			3995.7
BIC			4079.1			4065.9			4062.4			4095.2			4090.2			4029.6
Difficulty and Complexity																		
Fixed effects	Estimate		*SE*	Estimate		*SE*	Estimate		*SE*	Estimate		*SE*	Estimate		*SE*	Estimate		*SE*
Intercept	3.140		0.023	3.098		0.031	3.098		0.031	3.254		0.039	3.210		0.064	3.209		0.064
Random effects	σ²		% σ²	σ²		% σ²	σ²		% σ²	σ²		% σ²	σ²		% σ²	σ²		% σ²
Residual	0.104		59.8%	0.104		60.5%	0.092		53.4%	0.131		50.6%	0.131		49.0%	0.101		37.7%
Student (Intercept)	0.025	***	14.6%	0.025	***	14.4%	0.024	***	13.9%	0.054	***	20.7%	0.054	***	20.2%	0.050	***	18.8%
Teacher (Intercept)				0.030	***	17.3%	0.030	***	17.3%				0.053	***	20.0%	0.054	***	20.2%
Course (Intercept)	0.044	***	25.5%	0.013	***	7.7%	0.014	***	8.0%	0.074	***	28.6%	0.029	***	10.8%	0.028	***	10.6%
Students x Teachers (Intercept)							0.013		7.5%							0.034	***	12.7%
Fit statistics																		
Deviance			1738.1			1699.1			1696.2			2282.7			2262.4			2237.8
AIC			1746.1			1709.1			1708.2			2290.7			2272.4			2249.8
BIC			1768.7			1737.4			1742.1			2313.4			2300.6			2283.7

Note. σ² = variance, % σ² = Proportion of variance (percentage)
* p<.05, ** p<.01, *** p<.001.

Overall Rating of Course

Table 2.3 (upper half) provides the estimates for the models with the overall rating of course quality as criterion variable. Model 1 showed that the proportion of variance due to students (11% for seminars and 16% for lectures) was substantial but still markedly smaller than the proportion of variance due to courses (33% for seminars and 27% for lectures). Despite the fact that course quality was in the focus of this rating, Model 2 suggests that teachers were a similarly strong source of variance (10% in seminars and 9% in lectures) as the courses (14% in seminars and 7% in lectures). Finally, Model 3 revealed a strong increment of the interaction of students and teachers (17% variance explained in seminars and 25% in lectures). Nevertheless, even in Model 3 large proportions of variance remained unexplained (48% in seminars and 45% in lectures).

Overall Rating of Teacher

Table 2.3 (lower half) provides the estimates for the models with the overall rating of teacher performance as criterion variable. The results resemble those obtained for the overall rating of the course (which is not surprising given the strong correlation of the two ratings, see Table 2.4). In Model 1, students were a significant source of variance (11% for seminars and 12% for lectures). However, their contribution was much smaller than the proportion of variance due to courses (33% for seminars and 27% for lectures). Despite the fact that teacher performance was in the focus of this rating, Model 2 suggests that courses were a similarly strong source of variance (16% in seminars and 14% in lectures) as teachers (17% in seminars and 14% in lectures). Thus, students were influenced by (teacher-independent) characteristics of the courses when rating teacher performance. Finally, Model 3 revealed a particularly strong increment of the interaction of students and teachers (24% variance explained in seminars and 32% in lectures). Thus, similar to most of the other evaluation measures, the fit of students and teachers had a strong and systematic effect on the overall rating of the teacher. In Model 3, about one third of the overall ratings variance remained unexplained (33% in seminars and 32% in lectures).

Table 2.3

Estimates for the Cross-classified Linear Mixed Effect Models for the Global Ratings of Courses and Teachers

	Seminars									**Lectures**								
	Model 1: Student and Course			Model 2: Student, Teacher and Course			Model 3: Student, Teacher, Course and Students x Teachers			Model 1: Student and Course			Model 2: Student, Teacher and Course			Model 3: Student, Teacher, Course and Students x Teachers		
Rating Course																		
Fixed effects	Estimate		*SE*	Estimate		*SE*	Estimate		*SE*	Estimate		*SE*	Estimate		*SE*	Estimate		*SE*
Intercept	3.917		0.043	3.934		0.054	3.935		0.054	3.921		0.046	3.917		0.068	3.913		0.069
Random effects	σ²		% σ²	σ²		% σ²	σ²		% σ²	σ²		% σ²	σ²		% σ²	σ²		% σ²
Residual	0.409		64.5%	0.409		64.7%	0.305		48.2%	0.394		68.2%	0.394		68.2%	0.262		45.2%
Student (Intercept)	0.070	***	11.0%	0.069	***	11.0%	0.068	***	10.7%	0.092	***	15.9%	0.092	***	15.9%	0.079	***	13.7%
Teacher (Intercept)				0.066	***	10.4%	0.066	***	10.5%				0.052	***	8.9%	0.053	***	9.1%
Course (Intercept)	0.155	***	24.5%	0.088	***	13.8%	0.087	***	13.7%	0.091	***	15.8%	0.040	***	6.9%	0.040	***	7.0%
Students x Teachers (Intercept)							0.107	*	16.9%							0.146	***	25.1%
Fit statistics																		
Deviance			4532.6			4519.8			4513.3			4431.0			4414.1			4362.4
AIC			4540.6			4529.8			4525.3			4439.0			4424.1			4374.4
BIC			4563.2			4558.0			4559.2			4461.7			4452.3			4408.4
Rating Teacher																		
Fixed effects	Estimate		*SE*	Estimate		*SE*	Estimate		*SE*	Estimate		*SE*	Estimate		*SE*	Estimate		*SE*
Intercept	1.849		0.049	1.835		0.065	1.835		0.065	1.858		0.059	1.848		0.089	1.852		0.090
Random effects	σ²		% σ²	σ²		% σ²	σ²		% σ²	σ²		% σ²	σ²		% σ²	σ²		% σ²
Residual	0.377		56.5%	0.378		56.2%	0.220		32.7%	0.387		60.8%	0.387		60.8%	0.201		31.2%
Student (Intercept)	0.073	***	11.0%	0.073	***	10.8%	0.072	***	10.6%	0.076	***	11.9%	0.075	***	11.8%	0.059	***	9.1%
Teacher (Intercept)				0.117	***	17.4%	0.118	***	17.5%				0.090	***	14.1%	0.094	***	14.6%
Course (Intercept)	0.217	***	32.5%	0.104	***	15.5%	0.101	***	14.9%	0.174	***	27.4%	0.085	***	13.3%	0.083	***	12.9%
Students x Teachers (Intercept)							0.163	***	24.2%							0.207	***	32.1%
Fit statistics																		
Deviance			4398.8			4379.3			4365.0			4395.4			4379.4			4269.4
AIC			4406.8			4389.3			4377.0			4403.4			4389.4			4281.4
BIC			4429.4			4417.5			4410.9			4426.0			4417.6			4315.3

Note. σ² = variance, rel. σ² = Proportion of variance (percentage)
* p<.05, ** p<.01, *** p<.001.

Table 2.4

Correlations between Criterion Variables for Seminars and Lectures

	Seminars								**Lectures**							
	Mean	*SD*	Rating Teacher	Rating Course	Planning and Presentation	Interaction with Students	Interestingness and Relevance		*Mean*	*SD*	Rating Teacher	Rating Course	Planning and Presentation	Interaction with Students	Interestingness and Relevance	
Rating Teacher	1.88	0.82							1.81	0.80						
Rating Course	2.11	0.79	.69						2.06	0.76	.69					
Planning and Presentation	4.13	0.63	-.67	-.65	(.78)				4.13	0.70	-.69	-.66	(.90)			
Interaction with Students	4.35	0.63	-.74	-.63	.71	(.79)			4.38	0.66	-.67	-.52	.62	(.77)		
Interestingness and Relevance	3.94	0.79	-.71	-.71	.77	.73	(.85)		4.03	0.73	-.67	-.67	.67	.58	(.80)	
Difficulty and Complexity	3.17	0.42	.18	.21	-.15	-.23	-.22	(.74)	3.29	0.51	.02	.11	.04	-.11	-.10	(.77)

Note. Cronbach's α is shown in brackets on the main diagonal for the four scales of the questionnaire by Staufenbiel (2000).

All correlations were significant (*p*<.001)

Discussion

The present study examined the reliability of student evaluations of teaching quality. Our results showed that teachers and courses were essential sources of variance for all four facets of teaching quality examined in this study and also for the overall ratings of courses and teachers. A considerable proportion of variance, mostly around one fifth of the total variance in these measures, was also explained by students. However, in five of the six measures, the proportions of variance explained by teachers and courses together was nearly twice the proportion of variance explained by students. Dissociating the confound of teachers and courses did not reduce the unexplained variance but offered a more detailed picture by showing that both courses and teachers have an impact on how students evaluate teaching quality. Not surprisingly, the effects of teachers were larger in lectures compared to seminars. The most striking result was the finding that the interaction of students and teachers introduced in Model 3 was the strongest source of variance in most of the models. Including this interaction substantially reduced the unexplained variance.

The proportion of variance which is due to courses and teachers may, in principle, be interpreted as interrater reliability. This proportion ranged between 16% and 35% (*median* = 27%) in our study and was in most cases above the normative cut-off value offered by Marsh and Roche (1997) for single assessments. This finding is in line with those of other studies (Spooren, 2010; Rantanen, 2013; Staufenbiel et al., 2016). However, to achieve an acceptable reliability that can serve as the basis for reasonable and fair instructional and administrational decisions, average ratings based on the evaluations of several students must be used instead of individual ratings. These individual ratings differ considerably, are affected by individual student characteristics, and are therefore subject to a huge measurement error. The required number of evaluations can be estimated by the Spearman-Brown-Formula (Brown, 1910; Spearman, 1910). Given the average interrater reliability of 27% found in this study, one would need a sample 24 students to evaluate a course to achieve an interrater reliability of 90%. This number is comparable to the 25 students that Marsh and Roche (1997) suggested for courses with a single-rater reliability of 20%. In sum, teaching quality can be assessed reliably by student evaluations, provided that average evaluations are used that are based on a sufficient sample size.

Our data, however, also suggest four important qualifications and differentiations of this optimistic conclusion. First, the variance component of courses and teachers and the amount of measurement error vary considerably between different scales and also between lectures and seminars (as illustrated in Figure 2.3). For example, student evaluations of the difficulty and complexity but also the planning and presentation in seminars seem to be particularly unreliable, whereas the same evaluations with regard to lectures achieve a better interrater reliability. Overall, student evaluations of teaching quality seem to be more reliable in lectures than in seminars. This makes sense in so far as the instructional design of lectures (one teacher, teacher-centered instruction, and little variation in teaching methods) presumably yields a higher stability of teaching quality throughout the semester than it is the case with seminars in which student presentations of varying quality alternate with other forms of instruction. In contrast to the discrepancies between lectures and seminars, the discrepancies between different scales are more difficult to interpret. At this point, a replication of these differences seems to be required before any attempt at a substantial interpretation is made. Nevertheless, it is important to note that these discrepancies exist and that researchers and practitioners using student evaluations are aware that their reliability might depend on the exact content of the evaluations.

Second, this study aimed at dissociating the variance components of teachers and courses, which were confounded in previous studies. The separation of both sources of variance sources allowed a more detailed picture of student evaluations. For example, the results show that in lectures with their teacher-centered format of instruction, the relative effect of teachers compared to courses was stronger than in seminars. Overall, course characteristics were a major source of variance apart from the teacher. These findings have important implications for practical applications of student evaluations of teaching. Factors affecting the variance component of teachers can often be controlled by the teacher, with the exception of teachers’ personality characteristics (e.g., Clayson & Sheffet, 2006) or possible bias variables such as gender or attractiveness (Campbell, Gerdes, & Steiner, 2005; Basow, Codos, & Martin, 2013; Wolbring & Riordan, 2016). Factors that can be influenced include those central for teaching quality, such as teaching style, teaching method, or the way that discussions in class are organized. In contrast, this is often not the case with factors that contribute to the variance component of courses. For example, teachers often have little

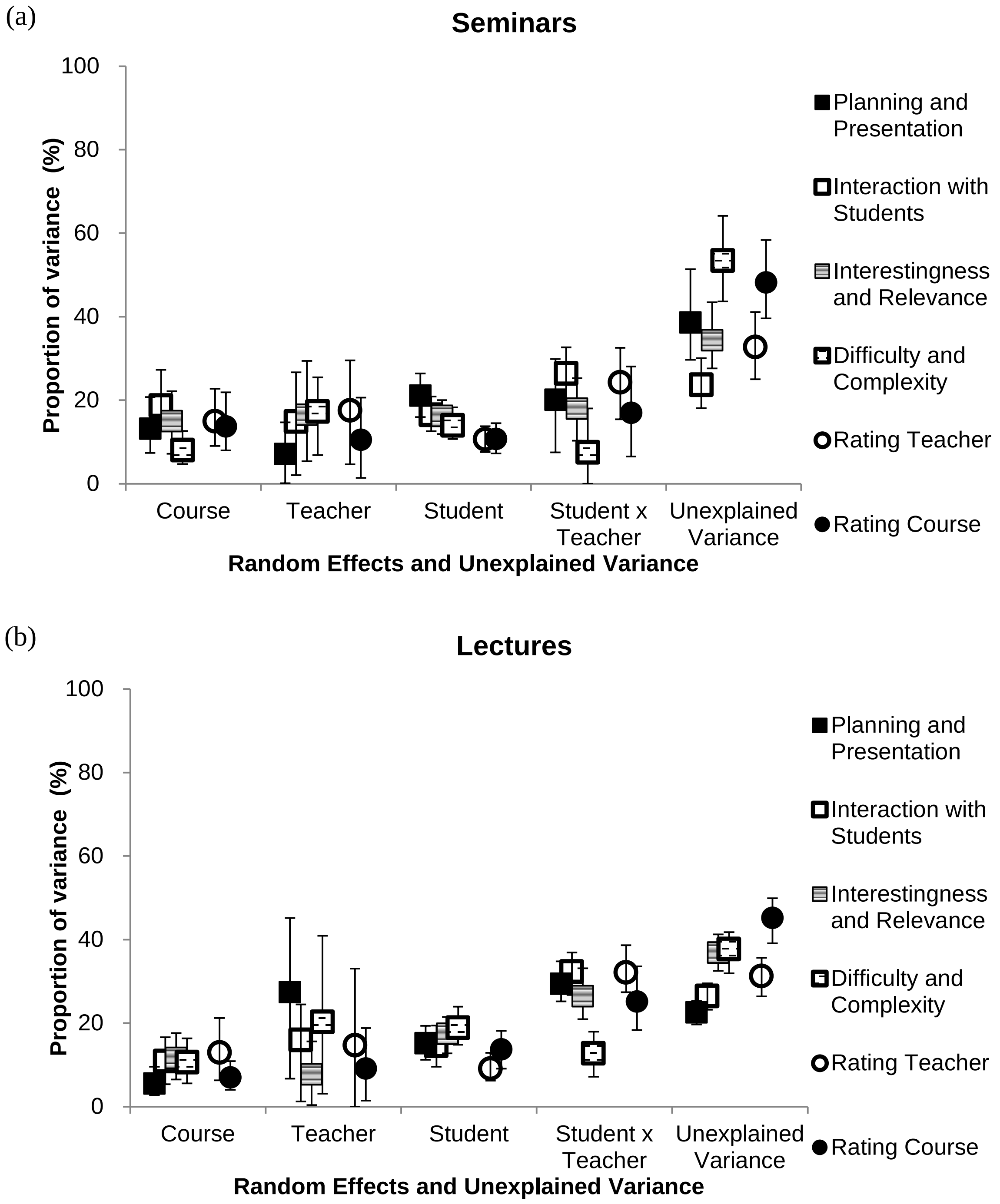

Figure 2.3. Proportions of variance explained by course, teacher, student, the Students x Teachers interaction and unexplained variance for the four scales of the evaluation questionnaire by Staufenbiel (2000) and the two overall ratings for seminars (a) and lectures (b). Error bars represent 95%-confidence intervals.

influence on environmental conditions such as room size, room location, the topic of the course, the required workload, and the number of students and their composition in a course. These factors are likely to affect student evaluations but are not (or only indirectly) related to teaching quality.

Third, the large proportion of systematic variance in the evaluations (between 11% and 21%) due to students was remarkable. In fact, the variance component of students was only slightly lower than the average proportions of variance explained by courses and teachers (27%). This finding has implications that go beyond reliability and touch upon the construct validity of student evaluations. Apparently, the systematic influences of student characteristics were almost as strong as the effects of course or teacher characteristics, suggesting that student evaluations cannot be regarded as pure measures of teaching quality but also capture (hitherto largely unknown) student characteristics to a considerable degree. Possibly, for example, students' personality traits (Patrick, 2011), their competence level (Spooren, 2010), or general response biases (Dommeyer, Baum, & Hanna, 2002; Sax, Gilmartin, & Bryant, 2003) systematically affect their evaluations of teaching quality. Future research should clarify the extent to that the variance component of students can be explained further by these different types of student characteristics.

Fourth, the total variance explained improved markedly by including the interaction of students and teachers in the model. The proportion of variance explained by this interaction ranged between 8% and 32% (for similar results obtained with factor analyses, see Leamon & Fields, 2005). Thus, the interaction was in the same order of magnitude as the variance components (i.e., main effects) of teachers and courses together. This finding has important implications for the interpretation of student evaluations of teaching. Apparently, different students evaluate the same teachers' performance in a (systematically) different way. Some students evaluate particular teachers consistently positive, whereas other students evaluate the same teachers consistently negative. Thus, how a student evaluates teaching quality seems to depend to a large extent on the fit of an individual student with individual teachers. The common practice of averaging across evaluations to derive one score that describes the teacher essentially neglects the student-teacher interaction and can therefore be misleading to some degree.

Our results provide an initial contribution to learning more about what exactly affects the degree of fit between students and teachers that leads to more positive or negative evaluations. Future research could examine this topic in different ways. According to Kenny (1994), heterogeneity in student assessments can be due to three factors: 1) Different information is used for the assessment, 2) students refer the same behaviours in their evaluation but interpret them differently, 3) or different kinds of non-behavioural information are used such as sympathy for their assessment. All three factors could play a role in the student-teacher interaction, but the effects of these factors could not be analysed in our data. Future multilevel studies should include potentially relevant teacher and student characteristics and their interactions as fixed effects to elucidate the factors that determine the fit of students and teachers that leads to more positive student evaluations of teaching. Another approach would be to examine whether the interaction itself is moderated by the expertise of the teacher (e.g., Tigelaar, Dolmans, Wolfhagen, & van der Vleuten, 2004). Expert teachers might be better able to adapt their teaching style to special characteristics of the students such as prior knowledge (e.g., Thompson & Zamboanga, 2003). This could result in a generally more positive evaluation by all students.

Despite its informative results, this study suffers from a number of limitations. One limitation is that the study was based on data from university courses in just one academic subject (Psychology) at one particular university in Germany in a limited number of student cohorts. Although we included students, teachers, and courses as random effects in our models to account for the fact that they are drawn from larger populations, the extent that the results can be generalized is unclear. This problem is complicated by the fact that students voluntarily took part in the evaluations, and only students who were present in the course when the evaluation was conducted could take part in the study. In other words, this study was based on a convenience sample (a weakness, though, that is shared by virtually all studies in the field). Nevertheless, the variance components estimated in this study lie in a range that is comparable to other studies (e.g., Spooren, 2010; Rantanen, 2013; Staufenbiel et al., 2016), which provides some confidence that our results can be generalized to some extent. Cross-validations of the results, in particular of the novel finding of a strong interaction of students and teachers, with different subjects, universities, and university systems would be very useful. Given the important practical implications of the variance components approach, these additional analyses would certainly be worthwhile.

In sum, our results suggest that student evaluations of teaching can be reliable assessments of the course and the teacher when aggregated evaluations based on a sufficient number of students are used (Marsh & Roche, 1997). However, the interrater reliability of student evaluations of teaching varies between different measures and course types (seminar vs. lecture). Moreover, these evaluations depend to a large extent on the students that provide the evaluations, implying that student characteristics can affect the evaluations and biases can occur through selection effects. These effects need to be taken into account when student evaluations of teaching are collected in university courses. It may also be regarded as a general problem for the validity of these evaluations, because data are almost as informative with regard to the students that provide the evaluations as they are with regard to the courses or teachers that are the actual focus of the evaluations. Finally, the large interaction effect of teachers and students that fundamentally reduced the amount of unexplained variance advances the findings from previous studies. This finding suggests that the fit of the individual students with their teachers plays an important role for student evaluations of teaching, a phenomenon of high practical relevance that requires clarification in future research.

CHAPTER 3 Study 2 -

Content expectations and prior subject interest affect psychology students' evaluations of teaching

Content expectations and prior subject interest affect psychology students' evaluations of teaching

Daniela Feistauer & Tobias Richter

Submitted to: Psychology Learning & Teaching

Abstract. Validity is an important issue when measuring teaching quality with student evaluations. This study examined effects of psychology students' general and specific preconceptions of psychology and their prior interest in the subject as variables possibly biasing the evaluations of psychology courses. German psychology students (*N* = 292) evaluated lectures and seminars over five years with a standardized questionnaire, yielding 3,348 data points. In cross-classified multilevel models, we separated the total variance into the variance components of course, teacher, student, and the Teacher x Student interaction and found evidence for biasing effects of general and specific preconceptions of psychology and prior subject interest. These effects were small overall and were stronger in lectures than in seminars. The results suggest that the validity of evaluations of teaching in psychology might be improved by creating realistic expectations of what psychology is about before students choose psychology as a study subject.

Keywords: content expectations, cross-classified multilevel analysis, student evaluations of teaching, prior subject interest, preconceptions of psychology, validity

Teaching quality in university courses is often measured via students' evaluations of teaching (SETs). With standardized questionnaires (e.g., Staufenbiel, 2000; Toland & De Ayala, 2005), such evaluations can be collected quickly and easily. These questionnaires are usually based on the assumption that teaching quality is a multidimensional construct (e.g., Cohen, 1981; Remedios & Lieberman, 2008) with dimensions referring to teaching methods, interaction with students, enthusiasm of the teacher, and feedback (Marsh, 2007; Rantanen, 2013). SETs are conducted at many universities and their outcomes are used for curriculum development and even hiring decisions. Despite their widespread use, SETs have often been criticized for a potential lack of validity (e.g., Kulik, 2001). One major concern has been that students are not capable of evaluating teaching quality, because their judgments are biased by background variables that are external to the criterion that needs to be assessed (for an overview, see Spooren et al., 2013; for a critical discussion, see Marsh, 1984, 2007). In this study, we used a broad range of psychology course SETs to examine two possible yet ubiquitous sources of biases: The clarity of expectations that students hold regarding the subject matter of psychology and their prior subject interest. The effects of these potential sources of bias on SETs will be first discussed to derive the research questions examined in our study.

Content Expectations and Students' Evaluations of Teaching

Psychology is one of the most popular undergraduate study programs. The number of applicants by far exceeds the number of places (e.g., the ratio was 16:1 at the University of Kassel in 2016, M. Keim, personal communication, June 16, 2017). Students who are newly enrolled in a psychology program hold fairly elaborate expectations of what psychology as an academic discipline is about (e.g., Goedeke & Gibson, 2011; Remedios & Lieberman, 2008). Expectations may be construed as predictions about what will happen in a given situation or probability judgments based on previous learning (Gigliotti, 1987). According to this definition, expectations about psychology as an academic discipline are predictions about what will happen in psychology courses, including expectations about the content that will be taught. Such content expectations may be informed by different sources such as common-sense everyday psychology (e.g., Fletcher, 1984), the display of psychology in the media (Holmes & Beins, 2009), input from peers (Pruitt et al., 2010), and knowledge about the professional activities within the discipline (Rowley, Hartley, & Larkin, 2008).

Studies exploring students' expectations about the subject matter of psychology found that they are quite variable, with some students holding rather imprecise or even unrealistic expectations. For example, in a survey of German psychology students (Orlik, Fisch, & Saterdag, 1971), 33% stated that they were studying the wrong subject, and 67% of students stated that they cannot do what they want in their studies. This finding is in line with more recent studies suggesting that psychology students expect practical and skill-focused content in their curriculum (Goedeke & Gibson, 2011; Hertwig & Stoltze, 2001). Many students expect that their undergraduate studies would provide them with knowledge directly relevant for helping people with psychological problems (Gaither & Butler, 2005; Hofmann & Stiksrud, 1993). In Germany, a considerable number of students aim to work later in therapy and counselling (estimates range from 50-73%, Handerer, 2014; Hertwig & Stoltze, 2001; Hofmann & Stiksrud, 1993). The expectations of these students are likely to be at odds with the typical content of undergraduate psychology programs that focus on research methods, statistics, diagnostic techniques and theory-based research on a variety of psychological subjects. One might suspect that the initial expectations would be corrected after students have started taking courses in psychology. However, Gardner and Dalsing (1986) found that students stick to some of their initial views of psychology even after attending several psychology courses.

Expectations may affect SETs, depending on the extent that they are met (yielding more positive evaluations) or violated (yielding more negative evaluations). Previous research focused on the effects of expectations regarding teachers (Bejar & Doyle, 1976; Pruitt et al., 2010) or expected grades (e.g., Holmes, 1972; Marsh, 1980; Feldman, 1997) on SETs. In contrast, the role of expectations concerning course content, or the subject matter of psychology in general, has received little attention in research. In the present study, we were interested in the question of whether students holding clearer expectations of the subject matter of psychology would evaluate such courses more positively than students holding less clear expectations. We assessed the self-reported clarity of first-semester students' general and specific preconceptions of psychology before they took their first psychology course. For general preconceptions of psychology, we subsume expectations concerning the subject matter of psychology in general, knowledge about the professions for which psychologists qualify, and the intended careers of the students. Specific preconceptions of psychology reflect expectations concerning the content of psychological subfields, such as social, developmental, or clinical psychology. These preconceptions translate into expectations of content in a given psychology course. We assume that the expectations of students holding clearer general and specific preconceptions of psychology will be met more often, leading to more positive evaluations compared to the evaluations provided by students with less clear preconceptions.

Prior Subject Interest and Students' Evaluations of Teaching

Prior subject interest, conceived as a personal disposition, is an important precondition of intrinsically motivated learning and has shown to be a strong predictor of learning achievements (e.g., Hidi, 2001). However, interest in the subject of a course has also been found in some studies to positively predict SETs (e.g., Feldman, 1976; Marsh, 1980; Staufenbiel et al., 2016). In addition, findings that elective courses receive moderately higher ratings than compulsory courses (e.g., Feldman, 1978) might be in part due to the greater interest in the elective courses chosen by students. In contrast to these results, other studies yielded no evidence in favour of a relationship between prior subject interest and course ratings (e.g., Olivares, 2001). The inconsistency of results could be influenced by the dimensions of teaching quality that are investigated in a study. Marsh (1980) found that prior subject interest accounted for most of the variance in SETs among 16 background variables (including expected grades and workload). However, the relationship differed between dimensions of teaching quality. He found relatively strong relationships with the perceived learning value of a course and the general course rating and weaker relationships for more objective dimensions of teaching quality such as the course organization. Conceptually, prior subject interest is a characteristic of individual students and not an aspect of teaching quality. Nevertheless, interest might colour how individual students experience a course, which, in turn, might bias their assessment of teaching quality (Marsh & Cooper, 1981; Paget, 1984). Therefore, prior subject interest may be considered a potential threat to the validity of SETs as a measure of teaching quality.

Rationale of the Present Study

The present study used Staufenbiel’s (2000) FEVOR questionnaire, a typical standardized multidimensional instrument, and newly developed questionnaires to examine students’ content expectations (i.e., their general and specific preconceptions of psychology) and their prior subject interest as potential threats to the validity of SETs. The questionnaire is based on a theoretical conception of teaching quality and is psychometrically sound and widely used in higher education in German-speaking countries.

We focused our analyses on the teacher performance item and the planning and presentation scale of the FEVOR. Teacher performance is a variable found in most SET questionnaires, because it is a broad indicator of teaching quality. However, despite its pervasiveness and intuitive accessibility, the measure is difficult to interpret, because it may comprise many (mostly unknown and possibly varying) components, including instinctive ratings (Merrit, 2008). Thus, teacher performance might be prone to the biasing effects of expectations about the course content and prior subject interest. In contrast, the second criterion variable planning and presentation is a central construct of teaching quality that appears with different names in most multidimensional models of teaching quality (e.g., Marsh, 1983) and is assessed with similarly worded items in all of the well-established SETs (e.g., “Course materials were well prepared and carefully explained”). The items used to assess the planning and presentation scale focus on course details (e.g., “The lecture is clearly structured”), possibly triggering rather reflective ratings (Merrit, 2008). Accordingly, planning and presentation might be less prone to biasing effects of content expectations and prior subject interest than teacher performance.

Content expectations were operationalized as general and specific preconceptions of psychology, measured with newly constructed scales and assessed before newly enrolled students of psychology attended their first course. Prior subject interest was assessed with an item from the FEVOR questionnaire. Considering that all three predictors are likely to overlap and compete for explained student variance, we investigated the unique contribution of each predictor in the context of the other predictors.

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 hierarchical (or crossed) structure. To this end, we used cross-classified multilevel analyses (Baayen et al., 2008), which included random effects (random intercepts) of all three possible sources of variance: teacher, course, and student (Feistauer & Richter, 2017a). Separate models were estimated for both criterion variables, teacher performance and planning and presentation.

Within the framework of cross-classified multilevel models, potential biasing effects of content expectations (general and specific preconceptions) and prior subject interest may be evaluated by estimating and testing the fixed effects of these variables on the criterion variables. Significant effects were interpreted by examining changes in the variance components teacher, course, student, and the Teacher x Student interaction caused by including the bias variables as predictors in the model. Finally, we also estimated models in which the biasing effect of content expectations and prior subject interest were included as random effects (i.e., effects varying randomly between students). These models allowed testing the possibility that the magnitude of biasing effects varies between students, with some students exhibiting greater bias than others.

We ran separate analyses for lectures and seminars because of the didactical and organizational differences of the two course formats (cf. Staufenbiel et al., 2016). Lectures have a much more strongly teacher-centered course format in comparison to seminars, which includes more contributions from the students. Therefore, we expected biasing effects on teacher performance to be greater in lectures compared to seminars, because unfulfilled expectations or unsatisfied interest in the course might be attributed more readily to teacher behaviour in lectures.

Method

Sample

This study used a data set of 3,348 student evaluations (questionnaire data) of courses in the Bachelor of Science psychology program held between 2011 and 2015 at the University of Kassel, Germany. From a total of 47 teachers (29 women), 17 taught 60 lectures and 40 taught 102 seminars. The sample of teachers included 18 doctoral students holding a position as researchers and lecturers (46%), 13 assistant professors or post-doctoral lecturers (32%), and 9 professors (22%). The evaluations were completed by 292 bachelor students (77% women) who participated in the courses. Although the evaluations were anonymous, students who completed evaluations of multiple courses were coded with the same ID. On average, each student evaluated 11 courses (lectures: *Md* = 6, *Range* = 1-14, seminars: *Md* = 5, *Range* = 1-14). The sample included courses such as statistics, cognitive, or social psychology (Table 3.1).

Procedure

The questionnaires were distributed in the last third of each semester (in the second half of January or June) by the teachers. The students were given 5-10 minutes to complete the questionnaires. Data were scanned using the software Remark Office OMR 8, and a student research assistant controlled the accuracy of the data. In addition to providing course evaluations, students completed a questionnaire during the first day of orientation, one week before the start of the first semester, asking them for their preconceptions of psychology.

Criterion Variables

The study was based on a standardized questionnaire used in German-speaking countries for the evaluation of university courses (FEVOR, Staufenbiel, 2000). 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 can be combined to four psychometrically distinct scales. In this study, we focused on the teacher performance item and on the planning and presentation scale consisting of five to eight items.

Table 3.1

Numbers of Evaluation Questionnaires Split by Course Type and Course Subject

	Lecture		Seminar	
	n	%	*n*	%
Cognitive Psychology	190	10.6%	187	12.0%
Psychology of Emotion, Motivation, and Learning	127	7.1%	150	9.7%
Counselling	35	1.9%	0	0.0%
Biological psychology	0	0.0%	169	10.9%
Diagnostics	90	5.0%	146	9.4%
Personality psychology	126	7.0%	108	7.0%
Developmental psychology	122	6.8%	181	11.7%
Clinical psychology	114	6.4%	125	8.0%
Research methods	196	10.9%	0	0.0%
Educational psychology	83	4.6%	64	4.1%
Social psychology	157	8.7%	200	12.9%
Statistics	418	23.3%	38	2.4%
Environmental psychology	89	5.0%	93	6.0%
Industrial psychology	48	2.7%	92	5.9%
Total	1795	100%	1553	100%

Note. N = 3348.

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* = 1.81, *SD* = 0.77; seminars: *M* = 1.83, *SD* = 0.80).

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" or "The lecture is clearly structured." The scale consists of five items in lectures (*M* = 4.11, *SD* = 0.67, Cronbach's α = .88) and eight items in seminars (*M* = 4.13, *SD* = 0.64, Cronbach's α = .78).

Predictor Variables

General preconception of psychology. This scale assesses the clarity of a first-semester student's concept of psychology as a study subject and profession. It contains items such as "I have a clear concept of the contents of the psychology program." Students rated responses on a Likert scale ranging from 1 (*strongly disagree*) to 5 (*strongly agree*) during the first day of orientation, one week before the courses begin. The scale consists of five items that are provided in Appendix A (*M* = 3.63, *SD* = 0.62, Cronbach's α = .59).

Specific preconceptions of psychology. This scale measures students' preconceptions of the subfields in psychology. The scale consisted of 10 items such as "I have a precise idea about the contents of social psychology." Likert scale responses range from 1 (*strongly disagree*) to 5 (*strongly agree*). All items are provided in Appendix B. Responses to this scale were also collected on the first day of orientation. The scale had an internal consistency of .77 (Cronbach's α, *M* = 3.20, *SD* = 0.53).

Prior subject interest. Prior subject interest was measured with one item ("What was your level of interest in course subject before the course started?") in the evaluation questionnaire for each course. Response options were from 1 (*very low*) to 5 (*very high*; *M* = 3.43, *SD* = 0.96).

Results

Analyses were performed with cross-classified multilevel models (Baayen et al., 2008), which allowed the separation of the variance components teacher, course, and student. These three sources of variance and the interaction between teacher and student were included as random effects in the analyses. Separate models were estimated for teacher performance and planning and presentation as criterion variables. The models were estimated with the statistical software R version 3.3.2 (R Core Team, 2016) and the full Maximum Likelihood estimation procedure included in the lmer function of the R-package lme4 (Bates et al., 2015). Models were compared with the anova function of the R-package stats (R Core Team, 2016), which compares the fit of nested models that differ in the structure of fixed or random effects. Data were analysed separately for lectures and seminars.

Estimated Models

We estimated a sequence of five nested models for both criterion variables. In the first step, we estimated a null model with no fixed effects but the variance components student, teacher, course and the interactions of teacher and student (Feistauer & Richter, 2017a). Figures 3.1 and 3.2 display the variances and the proportions of variance of both criterion variables.

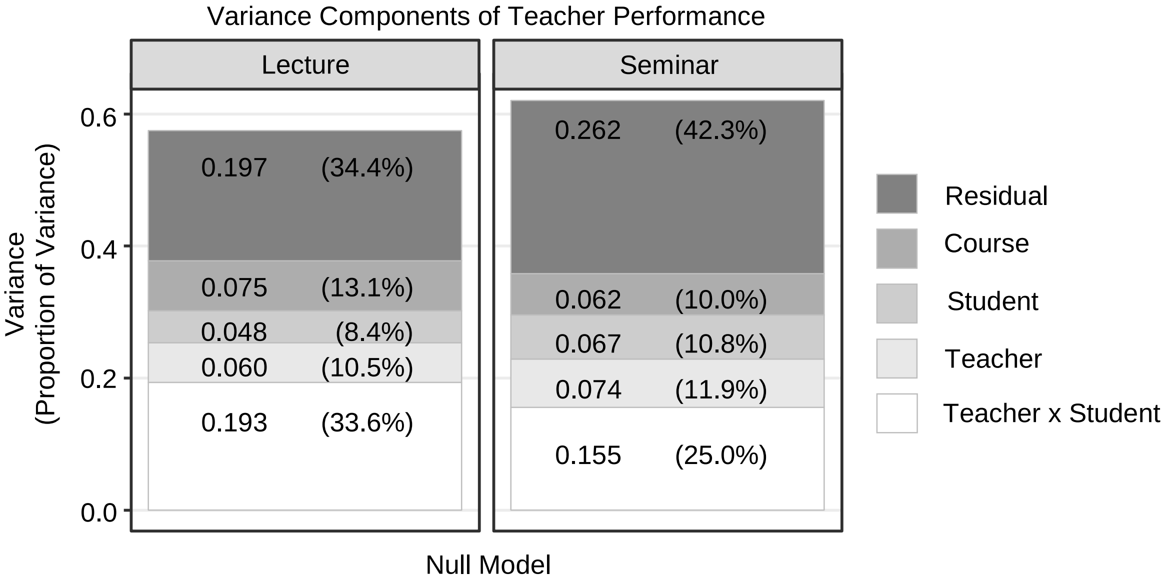

Figure 3.1. Variance and proportions of variance reflecting the different variance components for the criterion variable teacher performance (null model).

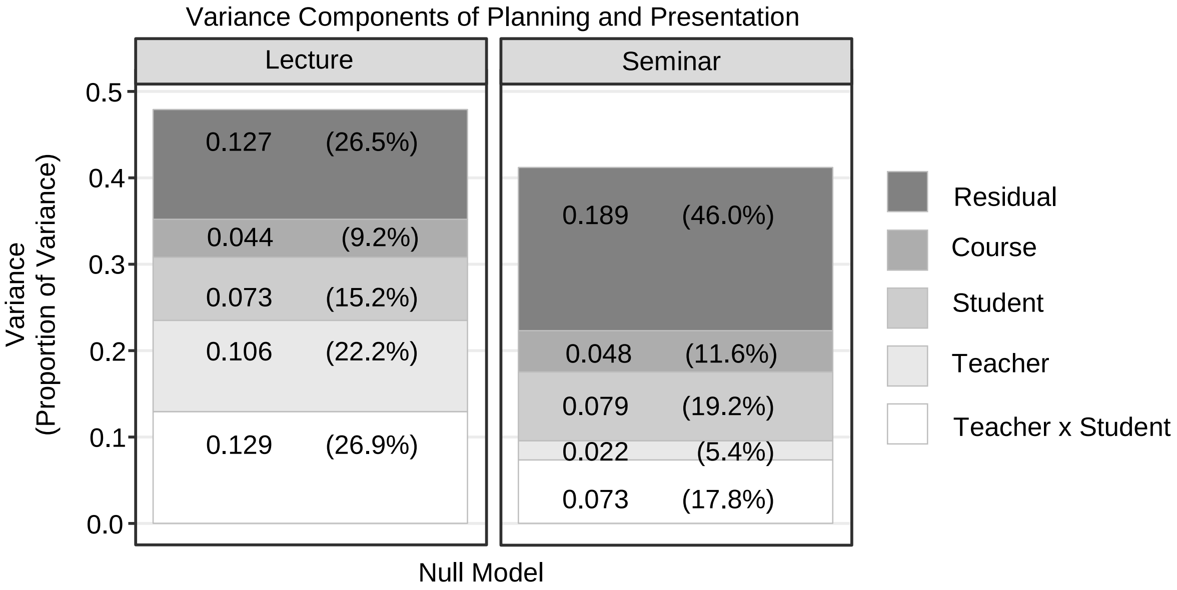

Figure 3.2. Variance and proportions of variance reflecting the different variance components for the criterion variable planning and presentation (null model).

We used the null model as a background for testing the effects of student background characteristics, which were entered as fixed effects and centered at the grand mean. We added students' general preconception of psychology in Model 1, specific preconceptions of psychology in Model 2, and prior subject interest in Model 3 as fixed effects. In the Models 1-3, the effects of the student-level predictors were assumed to be fixed effects that are constant across all students. However, it may be that the effects of student characteristics on SETs vary randomly across students. To address this possibility, we ran preliminary models with random coefficients for the student characteristics. Only the random coefficient of prior subject interest was significantly different from zero in these analyses, whereas general and specific preconceptions of psychology did not vary across students. Therefore, we estimated Model 4.

Teacher Performance

Results for the models with teacher performance in lectures as criterion variable are shown in Table 3.2. The overall mean of 1.805 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").

Table 3.2

Estimates for the Cross-Classified Multilevel Models for Teacher Performance in Lectures

	Model 0 (Null Model)		Model 1		Model 2		Model 3		Model 4	
Fixed effects	*Estimate* *(SE)*	*t*	*Estimate* *(SE)*	*t*	*Estimate* *(SE)*	*t*	*Estimate* *(SE)*	*t*	*Estimate* *(SE)*	*t*
(Intercept)	1.805 (0.078)	23.05	1.805 (0.078)	23.07	1.805 (0.078)	23.16	1.807 (0.079)	22.80	1.811 (0.079)	22.94
General preconception			-0.076* (0.035)	-2.14	-0.024 (0.039)	-0.62	-0.023 (0.039)	-0.60	-0.017 (0.039)	-0.43
Specific preconceptions					-0.131** (0.045)	-2.94	-0.126** (0.044)	-2.85	-0.130** (0.044)	-2.95
Prior subject interest							-0.040* (0.018)	-2.21	-0.040* (0.021)	-1.93
Random effects	*Variance*		*Variance*		*Variance*		*Variance*		*Variance*	
Residual	0.197		0.197		0.196		0.195		0.182	
Course (Intercept)	0.075		0.075		0.076		0.076		0.075	
Student (Intercept)	0.048		0.047		0.043		0.041		0.041	
Teacher (Intercept)	0.060		0.060		0.059		0.063		0.062	
Teacher x Student (Intercept)	0.193		0.194		0.195		0.196		0.188	
Student (Slope Interest)									0.024	
Covariance Student (Intercept) x Student (Slope Interest)									0.314	
Fit statistics										
-2LL	3553.0		3548.5+		3540.0++		3535.1+		3514.6+++	
AIC	3565.0		3562.5		3556.0		3553.1		3536.6	
BIC	3597.9		3600.9		3599.9		3602.6		3597.0	

Note. General preconception, specific preconceptions, and prior subject interest were all grand-mean centered before entering the predictors into the model. The number of observations that the variance components are based on are: Residual: $N = 1793$, Course: $n = 60$, Student: $n = 285$, Teacher: $n = 17$, Teacher x Student: $n = 1415$.

-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 nested models (χ^2-difference tests with 1 *df* based on the deviances): + $p < .05$, ++ $p < .01$, +++ $p < .001$ (two-tailed).

Including students’ general preconception of psychology as predictor in Model 1 led to a significantly improved model fit. The more precise that students were about general preconceptions, the higher they evaluated teacher performance (β = -0.076; *t*(289.8) = -2.14; *p* < .05). Adding this predictor led to a decrease of 2.1% in the variance component of students.

In Model 2, we entered specific preconceptions of psychology as an additional predictor. The correlation between general and specific preconceptions was .44. Inclusion of specific preconceptions led to a considerable increase in model fit. The more precise students were about specific preconceptions, the higher they evaluated teacher performance (β = -0.131; *t*(275.8) = -2.94; *p* < .01). Inclusion of this predictor led to a decrease of 8.5% in the student variance component. However, the effect of general preconception was no longer significant after including specific preconceptions as an additional predictor.

In Model 3, we added students’ prior subject interest as a predictor. This third predictor was basically uncorrelated with general preconception (*r* = .03) and specific preconceptions (*r* = .09). Adding prior subject interest also led to a significantly improved model fit. Greater student interest indicated more favourable evaluations of teacher performance in lectures (β = -0.04; *t*(1755.4) = -2.21; *p* < .05). Inclusion of this predictor led to an additional decrease of 4.7% in the student variance component but an increase of 6.8% in the teacher variance component.

In Model 4, we included prior subject interest as a random coefficient. A significant random coefficient would mean that students were influenced to different degrees (and possibly in different directions) by their prior subject interest while evaluating teacher performance. Including the random coefficient of prior subject interest significantly improved the model fit. The slope variance was 0.024, indicating that the magnitude effect of this variable on teacher performance varied considerably between students, with some students even showing negative effects (Leckie, 2013).

In seminars, none of the effects specified in Model 1 to 4 for the teacher performance criterion variable was significant.

Planning and Presentation

Results for the models with the criterion variable planning and presentation are shown in Table 3.3 for lectures and in Table 3.4 for seminars. The overall mean (lectures: 4.083;

seminars: 4.147; maximum: 5) in Model 0 suggests that students perceived their courses on average as well prepared, structured, and presented in a meaningful way.

Including students' general preconception as a predictor in Model 1 led to a significantly improved model fit. The more precise that students were about general preconceptions, the higher they evaluated planning and presentation (lectures: β = 0.116; *t*(266.8) = 3.31; *p* < .001, seminars: β = 0.085; *t*(246.4) = 2.35; *p* < .01). Adding this predictor led to a decrease in the student variance component (lectures: 6.8%; seminars: 3.8%).

In Model 2, we added specific preconceptions as a further predictor. Including it led to an improved model fit in lectures but not in seminars. The more precise specific preconceptions students held, the higher they evaluated planning and presentation in lectures (β = 0.137; *t*(260.3) = 3.12; *p* < .001). Including this predictor in lectures led to a decrease of 4.4% in the student variance component. The effect of general preconception was also no longer significant after including specific preconceptions as an additional predictor in Model 2.

Entering prior subject interest as an additional predictor in Model 3 significantly improved the model fit. Higher students' prior subject interest indicated more favourable evaluations of planning and presentation (lectures: β = 0.076; *t*(1728.4) = 5.13; *p* < .001; seminars: β = 0.056; *t*(1514.6) = 3.35; *p* < .001). Adding this predictor to the model led to a decrease in the student variance component (lectures: 4.6%, seminars: 5.3%), to a decrease of 4.1% in the seminar's course variance component, to an increase of 10.5% in the lecture's teacher variance component but to a decrease of 4.5% in seminar's teacher variance component, and to an increase in the Teacher x Student variance component (lectures: 4.7%, seminars: 2.8%).

Including the prior subject interest as a random coefficient in Model 4 led to a significantly improved model fit in lectures but not in seminars. The slope variance in lectures was 0.013, implying that the individual effects of prior subject interest on planning and presentation varied and were even negative for some students.

Table 3.3

Estimates for the Cross-Classified Multilevel Models for Planning and Presentation in Lectures

	Model 0		Model 1		Model 2		Model 3		Model 4	
	(Null Model)									
Fixed effects	*Estimate*	*t*	*Estimate*	*t*	*Estimate*	*t*	*Estimate*	*t*	*Estimate*	*t*
	(SE)		*(SE)*		*(SE)*		*(SE)*		*(SE)*	
(Intercept)	4.083	45.04	4.083	45.03	4.082	45.16	4.080	43.42	4.079	43.50
	(0.091)		(0.091)		(0.090)		(0.094)		(0.094)	
General preconception			0.116***	3.31	0.061	1.57	0.059	1.56	0.057	1.51
			(0.035)		(0.039)		(0.038)		(0.038)	
Specific preconceptions					0.137***	3.12	0.128**	2.94	0.128**	2.98
					(0.044)		(0.044)		(0.043)	
Prior subject interest							0.076***	5.13	0.077***	4.63
							(0.015)		(0.017)	
Random effects	*Variance*		*Variance*		*Variance*		*Variance*		*Variance*	
Residual	0.127		0.127		0.127		0.120		0.115	
Course (Intercept)	0.044		0.043		0.044		0.045		0.044	
Student (Intercept)	0.073		0.068		0.065		0.062		0.060	
Teacher (Intercept)	0.106		0.106		0.105		0.116		0.116	
Teacher x Student (Intercept)	0.129		0.129		0.129		0.135		0.129	
Student (Slope Interest)									0.013	
Covariance Student (Intercept) x Student (Slope Interest)									0.216	
Fit statistics										
-2LL	2909.7		2898.9++		2889.3++		2863.7+++		2851.2++	
AIC	2921.7		2912.9		2905.3		2881.7		2873.2	
BIC	2954.7		2951.4		2949.3		2931.1		2933.6	

Note. General preconception, specific preconceptions, and prior subject interest were all grand-mean centered before entering the predictors into the model. The number of observations that the variance components are based on are: Residual: N = 1795, Course: n = 60, Student: n = 285, Teacher: n = 17, Teacher x Student: n = 1416.

-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 nested models (χ^2-difference tests with 1 *df* based on the deviances): + p < .05, ++ p < .01, +++ p < .001 (two-tailed).

Table 3.4

Estimates for the Cross-Classified Multilevel Models for Planning and Presentation in Seminars

	Model 0 (Null Model)		Model 1		Model 2		Model 3		Model 4	
Fixed effects	*Estimate* *(SE)*	*t*	*Estimate* *(SE)*	*t*	*Estimate* *(SE)*	*t*	*Estimate* *(SE)*	*t*	*Estimate* *(SE)*	*t*
(Intercept)	4.147 (0.043)	96.72	4.146 (0.043)	97.02	4.146 (0.043)	97.01	4.141 (0.043)	97.30	4.143 (0.042)	97.54
General preconception			0.085** (0.036)	2.35	0.070* (0.041)	1.72	0.072* (0.040)	1.77	0.069* (0.040)	1.73
Specific preconceptions					0.036 (0.046)	0.78	0.028 (0.046)	0.60	0.026 (0.046)	0.57
Prior subject interest							0.056*** (0.017)	3.35	0.058*** (0.018)	3.29
Random effects	*Variance*		*Variance*		*Variance*		*Variance*		*Variance*	
Residual	0.189		0.191		0.190		0.188		0.182	
Course (Intercept)	0.048		0.049		0.049		0.047		0.047	
Student (Intercept)	0.079		0.076		0.076		0.072		0.072	
Teacher (Intercept)	0.022		0.022		0.022		0.023		0.022	
Teacher x Student (Intercept)	0.073		0.072		0.072		0.074		0.076	
Student (Slope Interest)									0.005	
Covariance Student (Intercept) x Student (Slope Interest)									-0.349	
Fit statistics										
-2LL	2711.4		2706.0+		2705.4		2694.3+++		2691.9	
AIC	2723.4		2720.0		2721.4		2712.3		2713.9	
BIC	2755.5		2757.5		2764.2		2760.4		2772.8	

Note. General preconception, specific preconceptions, and prior subject interest were all grand-mean centered before entering the predictors into the model. The number of observations that the variance components are based on are: Residual: N = 1553, Course: n = 102, Student: n = 284, Teacher: n = 40, Teacher x Student: n = 1464.

-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 nested models (χ^2-difference tests with 1 *df* based on the deviances): + p < .05, ++ p < .01, +++ p < .001 (two-tailed).

Discussion

Our study examined the validity of SETs by analysing effects of students' expectations of course content, operationalized via self-reported general and specific preconceptions of psychology one week before studying psychology, and their prior subject interest. The results revealed that content expectations (general and specific preconceptions) and prior subject interest positively affected the evaluations of teacher performance in lectures and the evaluations of planning and presentation in lectures and seminars. In more detail, the two moderately correlated components of content expectations explained overlapping proportions of variance in SETs. After including specific preconceptions in the model, general preconception was no longer significant. In contrast, prior subject interest, basically uncorrelated with both components of content expectations, provided a unique contribution to the explained variance in SETs.

These results support our hypothesis that the potential biasing effects of content expectations and prior subject interest are more likely to occur in lectures than in seminars. In seminars, neither content expectations nor prior subject interest were significantly related to teacher performance. This finding might be interpreted as support for a bias-free measurement. However, biasing effects were found for the scale planning and presentation in seminars. A possible explanation for the lack of effects might also be the low reliability of teacher performance compared to the planning and presentation scale in seminars (Feistauer & Richter, 2017a).

Our hypothesis that evaluations of teacher performance would be more prone to biasing effects of content expectations and prior subject interest than evaluations of planning and presentation was not supported. Apparently, the detailed items used for the assessment of planning and presentation failed to foster more reflective judgments (Merrit, 2008) that would prevent biasing effects. Our results are more in line with research by Pinto and Mansfield (2010) who found in group discussions that students seem to rely mostly on emotionally-charged ("gut reactions") and not so much on reflective judgments when evaluating teaching quality. Even when students were prompted with the question, "Rate the logical arrangement of the course material" (that corresponds to the planning and presentation scale used in our study), 33% of the responses were of the emotionally-charged type.

Our results show that planning and presentation, as a central dimension of SETs for measuring teaching quality, which should be completely dependent on teacher performance, is biased to some extent by content expectations and student interest. Although this conclusion casts some doubt on the validity of these evaluations, note that the models in which content expectations and prior subject interest exerted significant effects on SETs, the three predictors explained small proportions of the total variance of the criterion variables.

In this study, we did not investigate the psychological mechanisms that underlie the observed biasing effects of content expectations and prior subject interest. However, the pattern of variance components that were reduced or not reduced after including the predictors suggest specific hypotheses concerning mechanisms that could be tested in further research. First, students with more expectations concerning the course content might also have more knowledge of the subject, follow learning goals, and therefore experience faster progress in learning. Such learning goals possibly lead to a higher course commitment (Mikkonen, Ruohoniemi, & Lindblom-Ylänne, 2013; Seijts & Latham, 2011). The considerable decrease in the student variance component (9%) when content expectations were included and also the absence of a decrease in the variance component that reflected the interaction of teacher and student is consistent with this explanation. Apparently, content expectations affect SETs regardless of the teacher. In contrast, including prior subject interest explained variance on all variance components. The effects of this variable cannot be explained by the commitment to a course and a teacher but instead indicates that other factors must be present. Further research is necessary to identify these factors.

In the random coefficient models estimated in the final step of analyses, only the random coefficient of prior subject interest in lectures was significantly different from zero. In other words, the degree to which prior subject interest was related to SETs varied between students, from a mostly strong positive association to a slightly negative association in some students. The positive associations might have occurred because of higher student interest and the resulting higher positive evaluations, whereas the negative associations could be due to highly interested students who were disappointed with the way the subject matter was presented in the lecture. Future research is necessary to evaluate these assumptions. In contrast to the random coefficient of prior subject interest, the random coefficients of the two components of content expectations were not significant in any of the models. Although null results should be interpreted with caution, this finding provides some indication that the positive associations of content expectations with SETs, albeit small, vary little among students.

Limitations of the Present Study

We believe that our results are informative, but the conclusions that may be drawn from them are limited in several ways. One limitation is that prior subject interest was measured retrospectively at the time of the SETs. Thus, the measurement of prior subject interest could have been affected by the teacher performance to an unknown extent. This makes a causal interpretation of the relationship between prior subject interest and SETs difficult. In a future study, we plan to assess subject interest at the beginning of a course, which then cannot be influenced by teacher performance.

A second limitation concerns the relatively low effect sizes found for content expectations and prior subject interest. One possible interpretation of this finding is that the validity of SETs is only slightly compromised by these bias variables. However, another factor that might have artificially lowered the effect sizes was that SETs commonly suffer from ceiling effects (Menges & Brinko, 1986; Zhao & Gallant, 2012). Our study is no exception. The grand means of both criterion variables were relatively close to the maximal obtainable values, which limited the variability and, hence, possibly the effect sizes.

A limitation shared with most other studies in the field is that the standards students apply in their evaluation are not clear and likely vary among students (maybe even between different evaluations given by the same student). Previous research suggests that students compare the evaluated course with other courses (Darby, 2008), with an ideal course (Dunegan & Hrivnak, 2003; Goldstein & Benassi, 2006), or with an average course (Grimes, Millea, & Woodruff, 2004). In many contexts, it is common to instruct or even train raters in using the same standards (Lehmann, Ban, & Donald, 1965), but this procedure is normally not practiced in educational institutions that use SETs. Researchers therefore need to develop evaluation instruments that include standards in the items to avoid ceiling effects in SETs.

Finally, our study was based on data from psychology courses at one university in a few student cohorts. We included students, teachers, and courses in our models as random effects, because they were drawn from larger populations. However, whether the results generalize to other study programs remains an open question. This problem is complicated, because our study was based on a convenience sample of only the students who were present when the evaluation was conducted. Nevertheless, the effects of this study are comparable to other studies (e.g., Spooren, 2010; Rantanen, 2013; Staufenbiel et al., 2016), which makes us optimistic that the results can be generalized.

Conclusion

Our study has shown that content expectations and prior subject interest can affect SETs to some extent. These variables can be regarded as bias variables, because teaching quality should be influenced only by the performance of the teacher, not by students' individual characteristics. However, the biasing effects seem to be rather small, implying that the threat they pose to the validity of SETs is not very large.

Previous research on effects of expectations on SETs has focused on effects of expectations concerning grades (e.g., Centra, 2003; Marsh, 2007), course difficulty (Addison, Best, Warrington, 2006), and workload (e. g. Gursoy & Umbreit, 2005; Kreber, 2003). To our knowledge, this study is the first one examining content expectations as a potential bias of SETs. Expectations concerning the subject matter of psychology (and other study subjects) can be influenced by giving prospective students a realistic preview, for example, by allowing them to attend a trial study program or informative meetings before they sign up for the study program. Our results suggest that such measures might increase the fit of students' expectations and course content, thereby improving the validity of SETs.

CHAPTER 4 Study 3 -

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

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

Daniela Feistauer & Tobias Richter

Submitted to: Research in Higher Education

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

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). These doubts are as old as SETs, and there is extensive research on this topic (Barr, 1943; Marsh & Roche, 1997; Olivares, 2003; Sears, 1921; Spooren et al., 2013; Stalnaker & Remmers, 1928; Staufenbiel et al., 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. In the following literature review, we discuss earlier research on likability and prior subject interest in the context of SETs to derive the research questions examined in our study.

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 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, & 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, 1982b) 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, 1982b, 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 (2017b) reported a weak relationship with two similar dimensions, teacher performance and planning and presentation.

Conceptually, prior subject interest is a personal disposition of individual students and not an aspect of teaching quality. Therefore, it should not influence the ratings of SETs and may 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, 1982b; 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; 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 = .61) between pre-course motivation and post-course ratings of pre-course motivation of single courses. In addition, Clayson and Sheffet (2006) 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. However, the relationship between likability and SETs could had already been influenced by, for example, teaching quality in an early session of the course. Unfortunately, the likability scores at Week 0 were not reported. Likability of teachers measured at the beginning of the course could be influenced by students' familiarity with the teacher, especially when they have attended courses taught by the same teacher. Thus, familiarity is an important covariate that needs to be considered.

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. We investigated 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 et al., 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, 2017a). 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. 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 related 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. 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. 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.

Method

This study analysed a dataset of 517 student evaluations (questionnaire data) of courses 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. 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). They were given 5-10 minutes of the course time to complete the questionnaires. In addition to providing evaluations, students rated at the beginning of the course 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 α = .85) and eight items in seminars (*M* = 4.11, *SD* = 0.82, Cronbach’s α = .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 4.1 for lectures and in Table 4.2 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 α) of .92 in lectures (Table 4.1) and .89 in seminars (Table 4.2) and is provided in Appendix C. The single item and the scale correlated to .98 in lectures (Table 4.1) and 89 in seminars (Table 4.2). 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.

Table 4.1

Intercorrelations Between Predictor Variables for Lectures

	M	*SD*	Likability	Likability	Scale	Prior subject
			T1	T2	Likability	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* < .001 (two-tailed).

Table 4.2

Intercorrelations Between Predictor Variables for Seminars

	M	*SD*	Likability	Likability	Scale	Prior subject
			T1	T2	Likability	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* < .05, *** *p* < .001 (two-tailed).

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 4.1 for lectures and in Table 4.2 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 et al., 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:

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

In Equation 0, Y_{sct} represents the evaluation score provided by student *s* for courses *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 $h_{00\mathrm{s}}$ 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 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 $LT1_s$ 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.

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

Ratings of Teacher Performance in Lectures

Results for the six models with teacher performance in lectures are shown in Table 4.3. 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 4.1 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.

Table 4.3

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

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

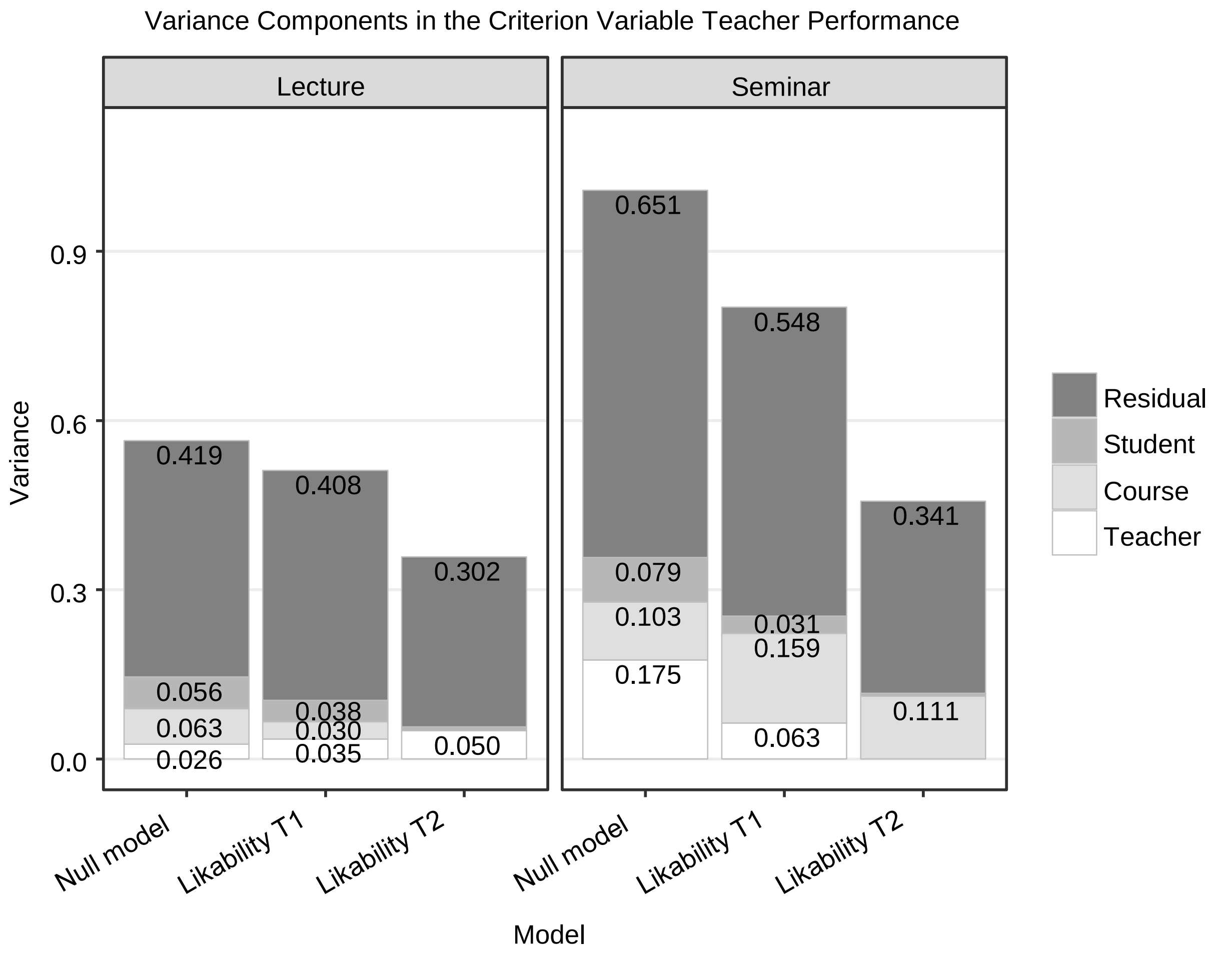

Figure 4.1. Comparison of the different likability model variance components in stapled bar plots for the teacher performance criterion variable. Null model: Model without predictors, Likability T1: Model with the likability predictor at the beginning of the course, Likability T2: Model with the likability predictor at the time of evaluation.

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 (β_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 (β_1 = -0.23, *t*(238.7) = -4.03, *p* < .001) but not prior interest (β_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.

Planning and Presentation in Lectures

Results for the six models with planning and presentation in lectures are shown in Table 4.4. 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 4.2 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.

Table 4.4

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

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

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.

Ratings of Teacher Performance in Seminars

Results for the six models with teacher performance in seminars are shown in Table 4.5. 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 ($\beta_1 = -0.49$, $t(241.4) = -8.05$, $p < .001$). 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 4.1 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

Table 4.5

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

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

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 (β_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 (β_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 (β_1 = -0.48, $t(242.2)$ = -7.89, $p < .001$) and more interesting (β_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.

Planning and Presentation in Seminars

Results for the six models with planning and presentation in seminars are shown in Table 4.6. 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 (β_1 = 0.29, $t(234.2)$ = 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 4.2 shows the total variance and the differences in the variance components of the null model compared to Model 1a and 1b. Inclusion of familiarity

Table 4.6

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

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

had no influence on planning and presentation in seminars (β_5 = -0.06, *t*(204.4) = -0.39, *p* > .05) but the interaction between likability and familiarity was significant (β_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 (β_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 (β_3 = 0.16, *t*(247) = 3.54, *p* < .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 (β_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 (β_1 = 0.28, *t*(232.7) = 5.82, *p* < .001) and more interesting (β_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.

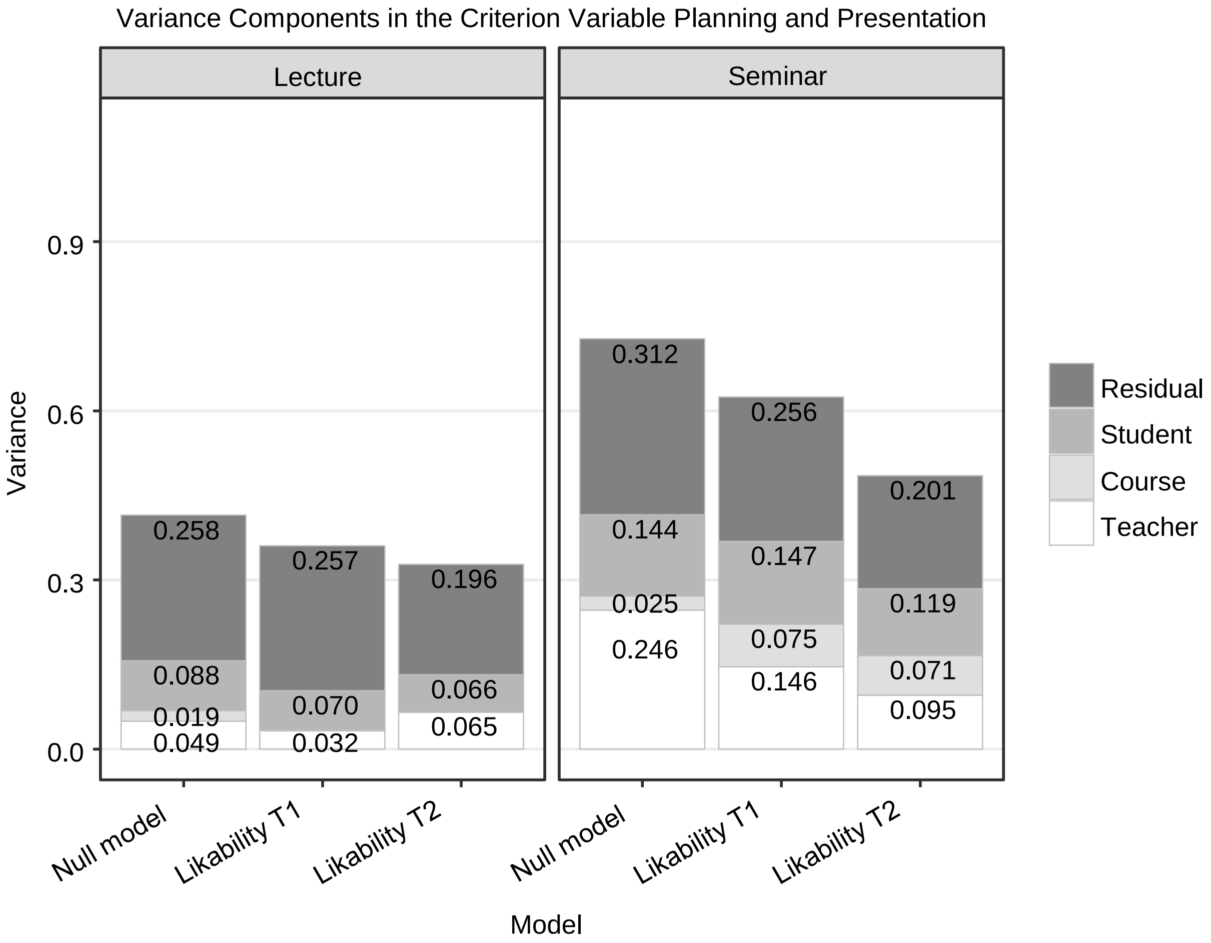

Figure 4.2. Comparison of the different likability model variance components in stapled bar plots for the planning and presentation criterion variable. Null model: Model without predictors, Likability T1: Model with the likability predictor at the beginning of the course, Likability T2: Model with the likability predictor at the time of evaluation.

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.

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, because of the implications that familiarity is not an important factor when analysing the various effects on SETs. This result is 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 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 (1982b) 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 is a consistent but relatively harmless bias that only slightly compromises the validity of SETs.

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 students voluntarily took part in the SETs, yielding a convenience sample (a shortcoming shared by most other studies in this area).

Another possible limitation is that we based our analyses on online SETs. Results from SETs provided online could be different from results based on paper-pencil SETs. Dommeyer, Baum, Hanna, and Chapman, (2004), among others, directly examined this question and reported no significant differences between the two types of SET scores. To overcome the typically low response rates (e.g., Dommeyer et al., 2004) of online SETs, we arranged 10 minutes in each course that was reserved for filling 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.

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).

CHAPTER 5

General Discussion

General Discussion

The aim of this dissertation was to test whether the application of SETs is appropriate to measure teaching quality. Based on the German SET questionnaire FEVOR (Staufenbiel, 2000), the reliable and valid assessment of teaching quality through SETs was tested in three empirical studies. In the following chapter, the findings are shortly summarized and discussed under the three initial research questions. This chapter closes with consequences, recommendations for future research and an overall conclusion.

Research Question 1: Are Student Evaluations of Teaching a Reliable Instrument for Measuring Teaching Quality?

The first research question of this dissertation emphasised the reliability of SETs. A reliable SET questionnaire is one requirement for a valid application of SETs for measuring teaching quality. To be able to conclude about the reliability of the FEVOR, this dissertation was based on multi-level analyses that allowed to partition the total variance into the three variance components, teacher, student and course. Under the assumption that the intra-class correlations of the teacher and course variance components exceed the cut-off score of .20 (Marsh & Roche, 1997) an instrument can be considered reliable. The findings of Study 1 showed that teachers and courses were substantial sources of variance (*Range* = .16-.35) on all SET dimensions. The investigation of the multidimensionality of SETs also allowed a detailed insight into each dimension’s reliability (see Figure 2.3) in comparison to other studies (e.g., Spooren, 2010; Rantanen, 2013). The mixed result pattern shown in Figure 2.3 provided evidence that some dimensions (e.g., global rating of teacher performance) had a higher reliability than other dimensions (e.g., planning and presentation in seminars). The data also provided evidence that suggests separate analyses for different course formats (lectures/ seminars), because the single-rater reliability was higher, on average, in lectures than in seminars. This difference was likely due to the different teaching styles. The less reliable dimensions in seminars might be affected by the different instruction format that varied between teacher instruction and student presentations.

To compare the results of the first study with other research, a median was generated for all dimensions and course formats. The resulting mean intra-class correlation of .27 is in line with findings from other studies (Rantanen, 2013; Spooren, 2010; Staufenbiel et al., 2016) and exceeded the suggested cut-off score of .20 (Marsh & Roche, 1997). In sum, these findings suggest that teaching quality can be reliably assessed by SETs, if the SET results are based on at least 24 students. Study 1 results can be supported by findings of a similarly designed study from another time frame (Feistauer & Richter, 2016) and calculations of intra-class correlations from the null models of Study 2. However, the null models of the Study 3 showed a reverse result pattern. In contrast to seminars (.20) that were equal to the cut-off score, lectures showed a smaller amount of variance (.16) explained by teachers and courses. Given the large number of students who attend lectures, obtaining evaluations from more than 24 students will likely get a reliable SET measurement of the teaching quality. Based on an intra-class correlation of .16, the number of students should be at least 46 for a reliable SET result. This number is also acceptable, because typically 60 or more students attend a lecture. However, logic implies that such a calculation could even be calculated with, for example, an unacceptable low intra-class correlation of .05. Such a correlation score would suggest that teaching quality is not reliably measured under such conditions, and any observable effects are more the result of chance (Clayson, 2017) than systematic differences in teaching.

In addition to these findings that can be interpreted in reference to the reliability of SETs, a considerable amount of variance was explained by students and by the interaction of students and teachers. The variance explained by students ranged between 7-21% and was comparable to results of other SET researchers (16%, Rantanen, 2013; 28%, Spooren, 2010). The variances were only slightly lower than the variance explained by teachers and courses, which has implications for the validity of SETs. This amount of variance indicates that the FEVOR instrument also assesses a considerable amount of individual differences in students, the result of which can be a sign that SETs are merely a measurement of student satisfaction (Pötschke, 2010; Prosser, 2011; Spooren et al., 2013; Uttl et al., 2017) or even chance (Clayson, 2017). However, the differences could also be explained by diverse response patterns of students (Ory & Ryan, 2001) or individual characteristics that influence how students experience their surroundings. Possible response patterns might be caused by leniency or centrality effects (Bernardin, 1978; Wolfe, 2004), a different base for comparisons (Darby, 2008; Dunegan & Hrivnak, 2003; Goldstein & Benassi, 2006; Grimes, Millea, & Woodruff, 2004; Metje, 2009), or the current mood (Schwarz, 2011) of students before evaluating the course. Several attempts to avoid such response patterns were employed in previous studies. Bernardin (1978) attempted to prevent leniency and centrality effects of students while they rated their teachers. However, he found that an hour-long and a 5-min long training only helped in a first rating period. In later rating periods, students in both training groups rated no differently than students in the control group (without training). An application of this result would mean that students should be trained before each evaluation session. The training might decrease the differences between students and their SETs, because presumably all (or at least most) students would apply the same standards on SETs. However, continuous training would probably lose its effectiveness if applied more than once. A cost-effective approach would be to remodel the SET questionnaire with more detailed and clearer instructions. Another more pragmatic approach would be to include items that assess leniency or centrality effects into the SET instrument. These items can then be interpreted as covariates in SET reports.

The second possible reason for response patterns in SETs could be a different base of comparisons in evaluations (Darby, 2008; Dunegan & Hrivnak, 2003; Goldstein & Benassi, 2006; Grimes, Millea, & Woodruff, 2004; Metje, 2009). Such different response patterns also could be avoided by rephrasing SET items by including comparison standards into the instruction of how to evaluate. Comparable to Bernardin’s (1978) attempt, these standards can be seen as a brief training session every time a SET is administered. However, before remodelling SET instruments with new instructions or rephrased items, further research is needed to clarify what students’ evaluations are exactly based on. For example, including items that assess the time frame which students use to evaluate teachers and courses could be fruitful. Do students focus on the most recent sessions or the current session when evaluating, or do they focus on all previous sessions, which is the intention of SETs? An additional item could address the base of comparison (e.g., ideal, other, or average course) used by most students. Analysing the responses to these items could provide suggestions for a prospective redraft of SET items and instructions, resulting in a more target-orientated redraft of SETs. In consideration of Tagomori and Bishop’s findings (cited in Bonitz, 2011), scrutinizing SETs for additional flaws would be prudent, given that they reported 79% of over 4,000 checked items were flawed in their wording or response options.

The last reason for different response patterns are individual characteristics of students. Students’ characteristics increase the variability in SET data (Gillmore et al., 1978), which affects the interrater reliability index among students. Some student characteristics such as content expectations lead to a differentiated perception of the teaching. This different perception leads then to more variability in SETs. Hence, reliability is higher among students who evaluate teaching with similar expectations than among students who have different expectations. A further possibility is that the students in some courses know each other better than in other courses. Familiarity could breed more conversations about the teacher, the course, or the content, which might affect the interrater agreement. Under such conditions, SET scores are less independent from each other than in courses in which students do not know each other well. Hence, interrater reliability covaries negatively with independence among the students (Amelang & Schmidt-Atzert, 2006). Such a hypothesis could be tested with separate analyses of monovalent and polyvalent courses. This analysis was not possible in the current dissertation, because the data contained too few polyvalent courses.

Apart from the variance explained by students, a remarkable result was the considerable decrease in the dimension's unexplained variance in SETs (Study 1: 8-32%, Study 2: 18-37%) by including the interaction of students and teachers as a further variance component. This finding is in line with results found by Leamon and Field (2005). The inclusion of this interaction did not reduce the variance of the other components but was solely supplied by the unexplained variance. This finding suggests that in addition to the main effects of students and teachers that might be caused by special behaviour or personality characteristics, different students evaluated the same teacher in a systematically different way. Therefore, SETs seem to also assess, next to the actual aim of teaching quality, the fit between students and teachers. This fit might be interpreted as support for a two-way-process of teaching (Fenstermacher & Richardson, 2005; Helmke, 2007; Rantanen, 2013). If students and teachers fit well together, it could indicate that the teachers heed, for example, prior knowledge and expectations of the students (Fenstermacher & Richardson, 2005; Helmke, 2007; McKeachie, 1997, Rantanen, 2013) and integrate them into their teaching. This integration could make it easier for students to accept and use the teaching offer (Helmke, 2007). An integration of, for example, prior knowledge can occur by using special teaching methods or providing varying amounts of detailed explanations. Nevertheless, the process of creating a better fit can be interpreted as a sign of teaching quality.

Other factors are unrelated to teaching quality such as attractiveness, likability, prior subject interest, motivation, and gender or age that might affect the fit between students and teachers. A short look in the validity analysis of Study 2 provides evidence that the effect of content expectations and prior subject interest varied only slightly (0.5-4.5%) in this variance component. The sample in Study 3 was not large enough for the interaction variance component of students and teachers. Therefore, the effect of teacher's likability could not be estimated. However, considering its effect on the other variance components and the effect of likability in interpersonal relationships (Lease, Musgrove, & Axelrod, 2002), the effect of likability on this interaction variance component might also be high.

Reliability analysis statistically assesses the consistency of SETs, but even a reliable SET instrument could consistently measure constructs other than teaching quality. For this reason, the other two research questions concentrate on the validity of SETs.

Research Question 2: To What Extent Does Student Characteristics, Such as Content Expectations, Prior Subject Interest, and the Perceived Likability of a Teacher, Influence Student Evaluation of Teaching Scores?

The analysis of some student characteristics provided evidence to the debate over the validity of SETs. In Studies 2 and 3, the student characteristics, content expectations, prior subject interest, and likability of the teacher perceived by students, were introduced as predictors into the multi-level analyses. Then, the effect of each predictor was interpreted based on the theoretical relationship of the predictor to teaching quality and the effect size as a measure of validity threat strength. Validity is a theoretical construct (Olivares, 2003). Therefore, the interpretation of a student characteristic as having had a biasing effect was made when theoretically no effect of the predictor should have been found on the SET dimension, because the predictor is not a conceptional part of the construct of teaching quality. Additionally, the teacher variance component should decrease after including this biasing predictor even though it is outside of the teacher's sphere of influence. The variance attributed to the teacher should only be influenced by teachers' characteristics that are a conceptional part of the construct of teaching quality.

Overall, results revealed effects of all three student characteristics on SETs. Content expectations and prior subject interest had small effects on SETs, but likability had a strong effect in comparison. The effect of content expectations was found in the two examined SET dimensions (global rating of teacher performance and planning and presentation) in lectures and in planning and presentation in seminars. Considering that content expectations was measured on the first day of the students' first semester, and therefore were not in the teacher's sphere of influence, they should not have had any effect on SETs. Yet, Study 2 revealed a small effect (< 2% total variance explained) of these expectations on SETs and a small decrease (< 2%) in the teacher variance component and a small decrease in the student variance component (< 8%) compared to the null models. Given that the effect of content expectations on SETs is theoretically unwarranted, the predictor should be classified as a biasing effect, albeit a nonthreatening effect because of its small effect size.

The second predictor, prior subject interest, was considered a possible biasing effect in Studies 2 and 3. It can be theoretically seen as a personal disposition that may colour students' evaluation of teachers and courses. Prior subject interest is also theoretically outside of the teacher's sphere of influence. Thus, its effect on SETs should also be interpreted as a biasing effect. Results of the second study showed a small effect of the additional predictor prior subject interest that led to a decrease in the teacher variance component (4-11%) and the student variance component (4-5%). However, the effect of prior subject interest, assessed retrospectively, could not be found in Study 3. This finding and the inconsistent results of former studies might indicate that the effect of prior subject interest is so small that only a large sample of SETs (as in Study 2) can provide sufficient power to detect it. A study conducted by Olivares (2001), with a similar sized sample as in Study 3, supports this argument. He also found no significant effect of prior subject interest. Nevertheless, the small effect size found in Study 2 is in line with the effect sizes found by other researchers (e.g., Marsh, 1982b; Marsh & Cooper, 1981; Wolbring & Treischl, 2016). The small effect size of prior subject interest should be interpreted as a nonthreatening biasing effect on SETs.

Likability of a teacher perceived by students was the third predictor that was the focus of the analysis in Study 3. It is theoretically unrelated to teaching quality, because teachers can only marginally influence how likable students find them. Therefore, likability was classified as a potential biasing effect. The results showed a strong effect of likability on SETs, explaining between 22-51% of the total variance, and it decreased the teacher variance component by 92% or more in the teacher performance dimension. The planning and presentation dimension also changed substantially through the inclusion of likability. The student variance component decreased also by 20-61%, but only in the dimension of planning and presentation of seminars. The size of the effect is sufficient reason to interpret likability as a very threatening biasing effect on the validity of SETs. This finding is in line with former research (Clayson & Haley, 1990; Clayson & Sheffet, 2006; Delucchi, 2000; Wolbring & Riordan, 2016) and contrary to Murray (1983). He reported that the prediction of SETs by external observers (trained student assistants) was based on teacher behaviour instead of popularity and personality.

The results discussed above showed effects not only on the teacher performance dimension of SETs as hypothesized but also on the planning and presentation dimension that theoretically should be less prone to biasing effects. Considering the high face validity of the planning and presentation dimension as a measure of teaching quality compared to the more vague operationalization of the teacher performance global rating, the current results do not favour the validity of SETs, especially given that all three student characteristics have no conceptional relationships with teaching quality. This finding also highlights the need to further explore the distinction between ratings based on intuition, which were expected when evaluating the global rating of teacher performance, and more reflective, rational-based ratings, which were expected for the planning and presentation dimension (Merrit, 2008). The found rating distinction is more consistent with Pinto and Mansfield’s (2010) emotionally-charged judgements. To test this hypothesis, an assessment of the student’s current mood (Schwarz, 2011) while evaluating courses is planned in a further study.

A noteworthy aspect of Studies 2 and 3 was the separate analyses of lectures and seminars. All biasing effect analyses were significant for lectures but only in some conditions for seminars. This finding is in line with Spiel and Gössler (2000) who found that lectures were more affected by biasing effects than other course formats. Possible reasons for this finding are the low reliability of seminars (see Study 1) or the more teacher-centered format of lectures that is theoretically more prone to biasing effects. Based on the didactical differences in the course format, lectures arguably show a clearer picture of teaching quality in the meaning of good teaching (Fenstermacher & Richardson, 2005) than seminars. Lectures allow for mostly only teacher’s behaviours (i.e., in the teacher’s sphere of influence) because of its teacher-centered format, whereas seminars are a mixture of student and teacher behaviours. Interestingly, a difference between lectures and seminars was not found by Staufenbiel et al. (2016) who also used the FEVOR questionnaire for their analyses. In their sample of over 60,000 SETs from different departments of one German university, a nearly identical fixed effect of prior subject interest in lectures and seminars was found. This difference in results might be due to the weaker reliability of seminars or to the wider variety of departments in their data. Some departments outside of psychology might offer seminars that are more teacher-centered and therefore more similar to lectures than the seminars in the current studies.

In sum, all student characteristics can be interpreted as having biasing effects on SETs, but only likability was found to be a strong threat to its validity. However, prior subject interest and likeability discussed in this section were assessed retrospectively at the end of the course. Therefore, the time of measurement itself could be a reason for faulty estimated effects. Whether prior subject interest affected SETs (and therefore the measurement of teaching quality) or another unknown factor affected the retrospective measurement of prior subject interest remains unclear. Maybe, prior subject interest had already had an effect on SETs. Hence, to come to a clearer understanding, the measurement time of both predictors, likability and prior subject interest, was the focus of the next research question.

Research Question 3: Does the Time of Measurement of Prior Subject Interest and Likability Influence Student Evaluation of Teaching Scores?

Although the inclusion of student characteristics is a widely used approach when assessing the validity of SETs, measuring the characteristics at the correct time is essential to obtain evidence that can be interpreted causally and that is not already affected by other factors. In most SET research, possible biasing effects were assessed at the typical time of evaluation at the end of the course (e.g., Barth, 2008; Dresel & Rindermann, 2011; Marsh, 1981, 1982a, 2007; Staufenbiel et al., 2016). Given that opinions about the teacher's likability or prior subject interest tend to be formed before or at the beginning of a course, such an approach might lead to faulty estimations of these effects. Those estimations can be affected by other factors such as recall biases (Ross, 1989) or interestingly even by teaching quality itself. Therefore, a measurement of retrospective variables is most likely already biased, and a theoretical assumption about the real effect might be false, because it suggests a wrong direction of the effect. For example, Staufenbiel et al. (2016) discussed a probable overestimation of the effect of prior subject interest measured at the time of evaluation. In contrast, Study 3 of this dissertation showed that the effect was underestimated. More precisely, effects of prior subject interest were weak but significant when measured at the beginning of the course, but they were not significant when measured at the time of evaluation. An opposite effect of measurement time was found for the predictor likability. The effect of likability measured at the beginning of course was strong (9-20%), but the effect of likability measured at the time of evaluation was vastly overestimated (21-55%). These findings show evidence that prior subject interest and likability are clearly affected over time by other undefined factors. The study provided no evidence for other factors that could have influenced the change over time. However, the results provided clear evidence that the time of measurement is an important factor to consider when administering SETs.

Consequences and Recommendations for the Future

The three empirical studies reported in this dissertation revealed new findings on the reliability and validity of the SET instrument FEVOR (Staufenbiel, 2000) as a measurement of teaching quality. The similarity of SET instruments (Schmidt & Loßnitzer, 2010) and comparable findings of other studies concerning the reliability and validity of SETs suggest that the results reported here are generalizable to other SET instruments. However, the findings of the three studies need to be interpreted in terms of their relevance for future research and the application of SET questionnaires in general.

Adequacy of students’ evaluations of teaching. Conclusions about the FEVOR questionnaire can be drawn based on the results of this dissertation. The instrument has acceptable reliability when at least 24 students evaluate a teacher and a course, and likability of the teacher has a strong effect on SETs even when it is measured at the beginning of the course. The effect of likability is unrelated to the theoretical concept of teaching quality. Therefore, a consequence of these results is that SETs results must be interpreted with caution. SETs should not be used to make human resource decisions, and if so, they should be used only to underpin other more valid sources of information (Penny, 2003). However, the problem is that even if “student ratings are not perfect measures of effective teaching […] they […] are more reliable and valid than most other indicators of teaching quality” (Penny, 2003, p. 400). To overcome this problem, Ory (2001, p. 8) suggested a “multiple-source and multiple-method approach” which combines data from peer reviews, teaching portfolios, course observations, and self-evaluations, although Marsh and Roche (1997) argued that evaluations of colleagues and administrators who visited the course once were not reliable because of a small rater agreement. Even trained observers could only reliably predict SET results when provided by 15 or more observers (Murray, 1983). Evaluations completed by teachers and students were moderately correlated (.45-.49; Marsh & Roche, 1997). Teachers are perhaps too close to their own teaching to impartially evaluate it (Metz-Göckel, Kamphans, & Scholkmann, 2012). Pham et al. (2012) found that teachers often underestimate their own teaching quality. The most cost-efficient option of measuring teaching quality is to ask teachers and students for completing evaluation questionnaires. A comparison of such self and other evaluations might even provide insight into dimensions and items that are evaluated differently in SETs.

Another reason for caution is that the administration of SETs has not been investigated as often as the instrument. For example, the appropriate time for measuring potential moderators and the way of generating SET reports for teachers are two areas worth exploring. Administrators should also be instructed on how to interpret SET reports correctly (e.g., Boysen, Kelly, Raesly, & Casner, 2014; Theall & Franklin, 2001), because even a reliable and valid instrument can be handled wrongly. The effect of one type of wrong application was examined in Study 3. Measuring all variables (e.g., prior subject interest and likability) at the appropriate time is essential, because otherwise the results can be flawed. A further type of wrong application is the recommendation that SET reports should be generated based on at least six students (Servicestelle Lehrevaluation, 2007), because otherwise the validity of the instrument and the anonymity of students would be in danger. Application of the Spearman-Brown-Formula (Brown, 1910; Spearman, 1910) on a sample of six students predicts an interrater reliability of only .60. With such a small reliability, the SET cannot be rated as a reliable instrument. Given the difficulty in getting enough students to participate in the evaluation, the recommendation of at least 24 student evaluations makes practical sense. To understand this suggestion from actual data, the following data were obtained from the Department of Psychology at the University of Kassel. From a total of 432 seminars (between summer 2011 until winter 2017), 409 had at least six students participating in the evaluation of the course, 244 seminars had at least 16 students (interrater reliability of .80), and only 85 seminars had at least 24 students (interrater reliability of .90). In comparison, 119 of 149 lectures fulfilled the criterion of at least 24 students. Considering reliability as an argument for providing quality feedback to teachers, only 19.6% of seminars (79.8% of lectures) had a high enough response rate to generate a SET report that fulfils the reliability criterion.

Finally, a main disadvantage that many SETs share should be considered. SETs suffer from ceiling effects (e.g., Keeley, English, Irons, & Henslee, 2013; Renaud & Murray, 1996). Keeley et al. (2013) attempted to increase the variance of SET data by changing 5-point response scale to seven or nine response possibilities, but the variance did not change much. They therefore recommended to keep the 5-point scale until a solution is found to prevent ceiling effects. Considering ceiling effects of SETs in a practical application would mean that SETs fail to distinguish between average and excellent teaching. They can only identify poor from average teaching, assuming that the instrument is valid. Even scores that indicate poor quality teaching should be interpreted with caution, because a poor SET result of the teaching could have been caused by an unfortunate group dynamic among the students and the teacher. Therefore, Gillmore et al. (1978) already suggested that at least five SET reports should be considered before taking any actions, for example, suggesting that the teacher teach another topic (Fenstermacher & Richardson, 2005; Gillmore et al., 1978; Rantanen, 2013) or receive additional training (Dresel & Rindermann, 2011).

In sum, the incorrect handling of SETs could lead to similar false decisions as an invalid instrument (Boysen et al., 2014). To my current knowledge, no special training exists (apart from some books) that provides administrators and superiors with information on how to handle SETs at the University of Kassel and other German universities. The questionable validity of the instrument combined with its incorrect handling might lead to decisions (personal decisions, change of curriculum) that should not be made by an untrained administrator.

Value of students’ evaluations of teaching. Another discussion thread that needs to be addressed is the value of SETs if they were judged to not have sufficient validity. Even if clear evidence warrants the conclusion that SETs are not a valid measurement of teaching quality, the data still provide information that can be used for improving teaching quality. This improvement can be viewed from the perspective of the teacher, student, or external stakeholder.

1) The teacher perspective: Teachers receive detailed (mostly quantitative) and structured feedback about how students like their course. A typical SET questionnaire provides students with a structure of different aspects of teaching, including an open space for comments if students want to provide information outside of the given structure. The

advantage of this design is that students can provide feedback anonymously without fearing reprisal for negative or unpopular feedback from the teacher or the other students. Despite the advantage of verbal feedback, students seem to perceive it as the least effective feedback (Ferguson, 2011). The argument that SETs provide more information about students than about teachers (Clayson, 2017; Kromrey, 1994b) can also be interpreted in the way that only teachers who know their students can provide better teaching than teachers who have little knowledge about the general needs of their students. Teachers can use all this information (about students) to improve their teaching (Kromrey, 1994b).

However, the application of SETs alone does not lead to an improvement of teaching (Marsh & Hocevar, 1991). The combination of SET results and follow-up discussions and additional training is found to be most effective (Overall & Marsh, 1979; Schmidt & Loßnitzer, 2010). Further development can be achieved through supervision, peer-vision, off-the-job- and on-the-job training, and discussions of the results with peers and students (Schmidt & Loßnitzer, 2010). Dresel and Rindermann (2011) found in their study, which included a two-year follow-up assessment, that counselling teachers based on SETs led to a moderate to large increase in subsequent SET results. Murray and Lawrence (1980) also found improved SETs for teachers who participated in a training designed to improve speech and presentation methods. Even if some trainings fail to directly improve teaching quality, they can raise the teacher's confidence, which in turn can lead to improved teaching performance (Ertmer & Ottenbreit-Leftwich, 2010; Stanton, 1995). One necessary point of training for inexperienced teachers might also be how to react to contradictory feedback among students, because the students pursue different goals (McKeachie, 1997) or they have different prior knowledge (Thompson & Zamboanga, 2003).

Although some teachers are proactive about monitoring their teaching through SETs (Staufenbiel, 2000), the results should not be used to rank teachers as some researchers do (e.g., Wolbring & Treischl, 2016). The individual course conditions vary greatly (Kromrey, 1994b). Some courses are mandatory and deliver unpopular content (e.g., statistics), and other courses are voluntary and can be attended because of interest in the topic (Feldman, 1978). A further difference is the type of course, for example, if it is delivered weekly or in a block course. Each course has unique conditions (e.g., Rindermann, 2003, 2009) that could affect SET results, which cannot be entirely attributed to the teacher.

2) The student perspective: Most educators believe that students should have the option to state their opinion on how successful they think a course was taught. Students generally have the ability to distinguish between teaching quality and entertainment (e.g., Bargel & El Hage, 2000). Some researchers, however, argue that students should not evaluate teaching (e.g., Kromrey, 1994b; Metje, 2009; Metje & Kelle, 2010; Moosbrugger & Schweizer, 2002). Possible reasons in support of this argument is that students focus more on the delivery than the content aspect of teaching (Murray, 1983), and they have insufficient knowledge about the content taught (Moosbrugger & Schweizer, 2002). Understandably, students lack content knowledge at the beginning of the course, but they should have a rough idea about the taught content at the end of the course to be able to provide feedback as to whether or not the content and the delivery was appropriate.

Another frequently used argument against SETs is that they only assess student satisfaction (e.g., Uttl et al., 2017). If teaching only serves a purpose when students use the offered teaching (Helmke, 2007), their satisfaction might indicate acceptance (Spiel, Wolf, & Popper, 2002). Satisfied students are presumed to be willing and have the right mind set for learning, whereas unsatisfied students might spend cognitive resources partly on reasons why they are unsatisfied and therefore have less resources for learning (Kirschner, 2002; van Merrienboer & Sweller, 2005). Since learning is the main goal of teaching, it should be in the interest of teachers that students use all possible cognitive resources for learning.

Most students evaluate their courses willingly, but some students decline from participating in SETs. Possible reasons might be that even when they evaluate a course, they believe that their ratings or feedback will have no impact. For this reason, Ory (2001) suggested that teachers explain to students at the beginning of a course the changes that were made as a result of the evaluations from the previous semester. Until better instruments for measuring teaching quality are developed, I think it is necessary to give all student the opportunity to give feedback to the teacher.

3) The external stakeholder perspective: A teaching environment is mostly a closed system with a teacher and students as the only actors. Thus, gaining information about the teaching conditions as an outsider is difficult. An external stakeholder could be a supervisor, an administrator, a parent, or a state agency that oversees higher education. SETs provide quantitative and qualitative information for all of these stakeholders as an indicator of quality

monitoring (Penny, 2003), although the information is normally only directly seen by supervisors and administrators. SETs can provide important information, for example, information in the comments section about an accidental duplication of content in two courses designed to teach different content. In such situations, the teachers of both courses (mostly newly employed teachers) can be asked to discuss and adjust their content so that the two courses have less content overlap. Another information is provided by one dimension in the FEVOR that typically assesses the difficulty of a course. If feedback indicates that a course has been perceived as very difficult, the teaching could be assisted through additional sources of information such as additional tutorials.

SETs also provide general information about the number of students who are still attending the course. This information can be used to check for room size, for possible problems with overlapping courses, for the interest in an elective course (does it justify its costs?) or for absenteeism rates. The response rate of SETs varies strongly (34-86%; Dommeyer et al., 2004). One possible way to interpret absenteeism is that it might be classified as an indicator of teaching quality (Wolbring, 2012; Wolbring & Treischl, 2016). The theoretical relationship between teaching quality and absenteeism can be that students have to be willing to learn (Fenstermacher & Richardson, 2005) and want to use the teaching offer (Helmke, 2007). However, some of the permanently absent students probably left the course because they did not like the teaching offer. Unfortunately, they often do not give the teacher feedback on what they do not like. Some teachers are quite flexible and willing to change topics or methods in their courses if they know it is desired. However, administrators should consider absenteeism rates while interpreting SET results because it can be assumed that some absent students evaluate a course and teacher in a different way than present students (Adams & Umbach, 2012; Kherfi, 2011). Consequently, teaching quality would be measured incorrectly. Another interpretation might be that absenteeism rates are unrelated to teaching quality. For example, students could be absent because of scheduling conflicts with work or familial responsibilities. Wolbring and Treischl (2016) found that at least one third of over 1,300 students were absent at the time of evaluation. Reported reasons for absenteeism were “course quality, students’ motivation, course topic, climate among course participants, course- and workload, and timing of the course” (Wolbring & Treischl, 2016, p. 66) or course size because large courses were attended less regularly (Wolbring, 2012). The effect of the different reasons for absenteeism on SETs should be clarified in future research.

Recommendations for the future. Future application of SETs could follow different directions. One direction is to improve the validity of SETs as a measure of teaching quality. The most expedient way to accomplish this goal is to provide administrators and policy makers with up-to-date research results and information on how to use and interpret SET results. A more difficult approach to improving SETs is to increase its reliability to the extent that it can be used with only six evaluations as discussed in a previous section, to minimize the strong impact of perceived likability of teachers, eliminate strong ceiling effects, and address absenteeism.

Administrators should be knowledgeable of the extensive literature of SET research. In view of the findings previously discussed that SET instruments consist of flawed items and response scales, a remodelling of the SET questionnaire is warranted, particularly the instructions. Students have no additional training to complete a SET questionnaire. Although survey duration is an important criterion, SETs should be developed with detailed instructions that allow no tolerance to individual interpretation of students. For example, it should be clear on what base of comparison (average vs. ideal course) an item needs to be interpreted (Darby, 2008; Dunegan & Hrivnak, 2003; Goldstein & Benassi, 2006; Grimes, Millea, & Woodruff, 2004; Metje, 2009). A clear base of comparison in response scales and the items could also prevent ceiling effects in SETs. In this context, one possibility is to take advantage of new technological developments like online-evaluations. Dommeyer et al. (2004) found no significant differences between online and traditional paper-pencil SETs. A descriptive comparison of the null models in Study 2 (paper-pencil SETs) and Study 3 (online SETs) support their finding. The application of online-SETs allows then adaptive testing (Amelang & Schmidt-Atzert, 2006) without lengthening the questionnaire over an acceptable limit. Adaptive testing can also be used for a better differentiation between average and excellent teaching. However, items for such a differentiation must be developed in the future. One disadvantage of online SETs is low response rates (Dommeyer et al., 2004; Dresel & Tinsner, 2008). Though, this shortcoming can be countered by in-class online SETs in which the teacher provides the students with 5-10 minutes of the course time for the evaluation via smart phones, tablets or laptops, which is the same time frame that is usually used for paper-pencil SETs. Such in-class online SETs were successfully administered in Study 3. The response rate was not much lower than with paper-pencil SETs.

Another possible addition to SET results often requested by teachers is data on active participation. At the moment, the FEVOR has one item that assesses the individual weekly workload in hours of the evaluated course and one item that asks about the participation of peers in the exercise class version (this version of the FEVOR was not analysed in this dissertation). The requesting teachers think they can only deliver high quality teaching when students are active participants in the course. This belief is consistent with the theoretical background of teaching quality as a two-way-process affected by teachers and students (Fenstermacher & Richardson, 2005; Helmke, 2007; Rantanen, 2013). Therefore, such items can reflect students' use of the teaching offer (Helmke, 2007). Possible items could be "How much did you invest in the success of the course?" or "How did you contribute to the success of the course?"

As Olivares (2003) noted, the application of SETs for decades has not directly affected teaching effectiveness or the quality of education. Thus, another possible option is to avoid the application of SETs. This approach would require the application of other methods for measuring teaching quality, but few alternative approaches are advisable. The following four alternatives are discussed: (1) Evaluations completed by other people than students attending the course. This method was already discussed in research and was rejected as mostly unreliable (Marsh & Roche, 1997). (2) The individual investigation of exam questions and final grades would also be an option, but this method requires expert knowledge and time and might therefore be financially unaffordable. (3) A standardized exam, such as the Programme for International Student Assessment (PISA), a college aptitude test, or a central high school diploma allows to test all students with the same exam and therefore would indicate whether students attending one course are less prepared than others. Nevertheless, this method could foster the ranking of teachers when necessary, based on the average results of their students. Teacher rankings, however, could lead to undesirable consequences like competition between teachers or departments. Moreover, standardized exams are very difficult or even impossible to implement on a single course level. This approach also disagrees with the general philosophy in higher education that universities or individual study programs should develop their own areas of expertise. (4) A smaller variant of the standardized exam is that one teacher teaches the students and another teacher tests them. This option is sometimes necessary when a teacher has unexpectedly become ill for an extended period of time. Yet, students dislike this option, because they fear being tested on content that was not taught by the original teacher.

Each of these methods has at least three disadvantages: They cost more money and would require more administration than SETs, and it is unclear whether they would provide a more valid approach to assessing teaching quality than SETs. The other disadvantage is that assessing teaching quality is only possible in courses that finish with a measurable outcome such as an exam. However, the German system in higher education has also courses that finish without graded exams. A further difference of these alternative approaches to SETs is that the focus is on learning success as the outcome variable instead of teaching itself.

Other methods that could be used in courses without exams are qualitative methods, such as observations of teaching situations, or even interviews with course participants (teachers and students). These methods are interesting approaches that could be systematically applied, but they cannot replace the comprehensive use of SETs, mainly because they would need too much time to execute and could possibly incur more costs. However, maybe assessing teaching quality in every course of every semester is not essential. Evaluating only a random selection of courses each semester is plausible. Students would likely be satisfied if they were not asked to complete SETs in numerous courses each semester. It would particularly reduce the effect of fatigue on SETs (Adams & Umbach, 2012; Peiffer, Rach, Rosanowitsch, Wörl, & Schneider, 2015). However, maybe a combination of several methods, suggested by Ory (2001), is the only option to come to more informed and valid decisions concerning, for example, the future of teachers.

Conclusion

The aim of this dissertation was to provide further evidence to the reliability and validity of SETs. The three studies of this dissertation provided evidence that SETs can be seen as reliable instruments for measuring teaching quality when a sufficient number of students (at least 24) provide an evaluation. In spite of this promising result, the reader should note that the current studies revealed that seminars were less reliable than lectures and that SETs were as informative about students as they were about teachers and courses. This latter finding was reinforced by the fact that in the first and second study a large interaction effect of teachers and students could reduce the normally high amount of unexplained variance. This interaction is a sign for the important role of the individual fit between student and teacher.

Starting with the given reliability of the FEVOR, Studies 2 and 3 revealed significant biasing effects on SETs. The student characteristics, content expectations, prior subject interest, and the likability of teachers perceived by students, were classified as biasing effects, because they are not theoretical tied to teaching quality and are outside of the teacher’s sphere of influence. Content expectations and prior subject interest, however, had no large impact on SETs, which suggests that they pose no threat to the validity of SETs. In contrast, likability showed a strong effect on SETs that could not be solely attributed to the retrospective measurement. Therefore, its effect indicates a strong threat to the validity of the FEVOR as a measure of teaching quality. Although the time of measurement can easily (but with administrative effort) be corrected to the appropriate time at the beginning of the course, the effect of the teacher’s likability cannot easily be modified. This one strong biasing effect is sufficient to question the validity of SETs, and therefore administrators should not use SET results to make decisions of high or far-reaching importance.

However, if SETs are applied, they should be only used for individual improvement of teachers. For example, they can provide a basis for students and teachers to discuss about teaching, and they can be a useful indicator for further teacher development (Dresel & Rindermann, 2011).

REFERENCES

References

Adams, M. J. D., & Umbach, P. D. (2012). Nonresponse and online student evaluations of teaching: Understanding the influence of salience, fatigue, and academic environments. *Research in Higher Education, 53*, 576-591.

Addison, W. E., Best, J., & Warrington, J. D. (2006). Students' perceptions of course difficulty and their ratings of the instructor. *College Student Journal, 40*, 409-417.

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.

Amelang, M., & Schmidt-Atzert, L. (2006). *Psychologische Diagnostik und Intervention [Psychological diagnostics and intervention]*. Springer: Heidelberg.

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.

Bargel, T., & El Hage, N. (2000). Evaluation der Hochschullehre [Evaluation in higher education]. In A. Helmke, W. Hornstein, & E. Terhart (Eds.): *Zeitschrift für Pädagogik, Qualität und Qualitätssicherung im Bildungsbereich. Schule, Sozialpädagogik, Hochschule* (pp. 207-224). Weinheim: Beltz.

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.

Basow, S. A., Codos, S., & Martin, J. L. (2013). The effects of professors' race and gender on student evaluations and performance. *College Student, 47,* 352-363.

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

Bejar, I. I., & Doyle, K. O. (1976). The effect of prior expectations on the structure of student ratings of instruction. *Journal of Educational Measurement, 13,* 151-154.

Bernardin, H. J. (1978). Effects of rater training on leniency and halo errors in student ratings of instructors. *Journal of Applied Psychology, 63*, 301-308.

Bligh, D. (1993). Learning to teach in higher education. *Studies in Higher Education, 18,* 105-111.

Bonitz, V. S. (2011). *Student evaluation of teaching: Individual differences and bias effects*. (Doctoral dissertation). Iowa State University.

Boysen, G. A., Kelly, T. J., Raesly, H. N., & Casner, R. W. (2014). The (mis) interpretation of teaching evaluations by college faculty and administrators. *Assessment & Evaluation in Higher Education*, *39*, 641-656.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 2,* 296-322.

Bühner, M. (2004). *Einführung in die Test- und Fragebogenkonstruktion [Introduction to the design of tests and questionnaires]*. München: Pearson Studium.

Campbell, D. T. (1960). Recommendations for APA test standards regarding construct, trait, or discriminant validity. *American Psychologist, 15*, 546-553.

Campbell, H., Gerdes, K., & Steiner, S. (2005). What's looks got to do with it? Instructor appearance and student evaluations of teaching. *Journal of Policy Analysis and Management, 24*, 611-620.

Carbone, A., Evans, J., & Ye, J. (2016). Beyond teaching quality: Towards a framework for course unit quality. *HERDSA Review of Higher Education, 3,* 57-72.

Centra, J. A. (1993). *Reflective faculty evaluation: Enhancing teaching and determining faculty effectiveness* (1st ed.). San Francisco: Jossey-Bass.

Centra, J. A. (2003). Will teachers receive higher student evaluations by giving higher grades and less course work? *Research in Higher Education*, *44*, 495-518.

Centra, J. A., & Gaubatz, N. B. (2000). Is there gender bias in student evaluations of teaching? *The Journal of Higher Education, 71*, 17-33.

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

Clayson, D. E. (2017). Student evaluation of teaching and matters of reliability. *Assessment & Evaluation in Higher Education*. Advance online publication. doi: 10.1080/02602938.2017.1393495

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.

Cohen, P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. *Review of Educational Research, 51*, 281-309.

Colby, J., & Witt, M. (2000, June). *Defining quality in education*. Paper presented at the meeting of The International Working Group on Education, Florence, Italy. Retrieved from: https://www.unicef.org/education/files/QualityEducation.PDF

Crosby, P. (1979). *Quality is Free.* New York: McGraw-Hill.

Darby, J. A. (2008). Course evaluations: A tendency to respond "favourably" on scales? *Quality Assurance in Education*, *16*, 7-18.

de Weert, E. (1990). A macro-analysis of quality assessment in higher education. *Higher Education*, *19*, 57-72.

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. (2002). College students' attitudes toward methods of collecting teaching evaluations: In-class versus on-line. *Journal of Education for Business*, *78,* 11-15.

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.

Döring, N., & Bortz, J. (2016). *Forschungsmethoden und Evaluation für Human-und Sozialwissenschaftler [Research methods and evaluation for human and social scientists]* (5th ed.). Berlin Heidelberg: Springer-Verlag.

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.

Dresel, M., & Tinsner, K. (2008). Onlineevaluation von Lehrveranstaltungen: Methodeneffekte bei der Onlineevaluation von Lehrveranstaltungen [Online evaluation of courses: Methodical effects]. *Zeitschrift für Evaluation*, *2*, 183-212.

Dunegan, K. J., & Hrivnak, M. W. (2003). Characteristics of mindless teaching evaluations and the moderating effects of image compatibility. *Journal of Management Education, 27*, 280-303.

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.

Ellis, R. (1995). Quality assurance for university teaching: Issues and Approaches. In R. Ellis (Ed.), *Quality assurance for university teaching* (pp. 3-15). Buckingham: Open University Press.

Ertmer, P. A., & Ottenbreit-Leftwich, A. T. (2010). Teacher technology change: How knowledge, confidence, beliefs, and culture intersect. *Journal of Research on Technology in Education, 42,* 255-284.

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. (2016): Wie zuverlässig sind studentische Einschätzungen der Lehrqualität? Eine Analyse mit kreuzklassifizierten Mehrebenenmodellen [How Reliable are Student Assessments of the Quality of Teaching? An Analysis with Cross-Classified Multi-Level Models]. In: M. Krämer, S. Preiser, & K. Brusdeylins (Eds.): *Psychologiedidaktik und Evaluation XI* (pp. 299-306). Aachen: Shaker.

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

Feistauer, D., & Richter, T. (2017b). Content expectations and prior subject interest affect psychology students’ evaluations of teaching. Manuscript submitted for publication.

Feldman, K. A. (1976). Grades and college students’ evaluations of their courses and teachers. *Research in Higher Education, 4,* 69-111.

Feldman, K. A. (1978). Course characteristics and college students’ ratings of their teachers: What we know and what we don’t. *Research in Higher Education, 9,* 199-242.

Feldman, K. A. (1989). The association between student ratings of specific instructional dimensions and student achievement: Refining and extending the synthesis of data from multisection validity studies. *Research in Higher Education, 30,* 583-645.

Feldman, K. A. (1997). Identifying exemplary teachers and teaching: Evidence from student ratings. In R. P. Perry & J. C. Smart (Eds.), *Effective teaching in higher education: Research and practice* (pp. 368-395). New York: Agathon.

Feldt, L. S., & Brennan, R. L. (1989) Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105-146). New York: The Oryx Press.

Fenstermacher, G. D., & Richardson, V. (2005). On making determinations of quality in teaching. *The Teachers College Record, 107,* 186-213.

Ferguson, P. (2011). Student perceptions of quality feedback in teacher education. *Assessment & Evaluation in Higher Education*, *36*, 51-62.

Fisch, R., Orlik, P., & Saterdag, H. (1970). Warum studiert man Psychologie? [Why do you study psychology?]. *Psychologische Rundschau, 21*, 239-256.

Fletcher, G. J. O. (1984). Psychology and common sense. *American Psychologist, 39,* 203-213.

Fondel, E., Lischetzke, T., Weis, S., & Gollwitzer, M. (2015). Zur Validität von studentischen Lehrveranstaltungsevaluationen [Validity of student evaluations of teaching]. *Diagnostica, 61,* 124-135.

Frazer, M. (1995). Quality in higher education: An international perspective. In D. Green (Ed.), *What is quality in higher education?* (pp. 101-111). Buckingham, Bristol PA: Open University Press.

Freeman, H. R. (1994). Student evaluations of college instructors: Effects of type of course taught, instructor gender and gender role, and student gender. *Journal of Educational Psychology, 86,* 627-630.

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.

Gaither, G. A., & Butler, D. L. (2005). Skill development in the psychology major: What do undergraduate students expect? *College Student Journal, 39,* 540-553.

Gardner, R. M., & Dalsing, S. (1986). Misconceptions about psychology among college students. *Teaching of Psychology, 13,* 32-34.

Gigliotti, R. J. (1987). Expectations, observations, and violations: Comparing their effects on course ratings. *Research in Higher Education*, *26,* 401-415.

Gillmore, G. M., Kane, M. T., & Naccarato, R. W. (1978). The generalizability of student ratings of instruction: Estimation of the teacher and course components. *Journal of Educational Measurement*, *15,* 1-13.

Ginns, P., Prosser, M., & Barrie, S. (2007). Students' perceptions of teaching quality in higher education: The perspective of currently enrolled students. *Studies in Higher Education, 32*, 603-615.

Goedeke, S., & Gibson, K. (2011). What do new psychology students know about psychology? *Australian Psychologist, 46*, 133-139.

Goldstein, G. S., & Benassi, V. A. (2006). Students' and instructors' beliefs about excellent lecturers and discussion leaders. *Research in Higher Education*, *47,* 685-707.

Goodyear, P. (2015). Teaching as design. *HERDSA Review of Higher Education*, *2,* 27-50.

Green, D. (1995). What is quality in higher education? Concepts, policy and practice. In D. Green (Ed.), *What is quality in higher education?* (pp. 3-20). Buckingham, Bristol PA: Open University Press.

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

Greenwald, A. G., & Gillmore, G. M. (1997). Grading leniency is a removable contaminant of student ratings. *American Psychologist*, *52*, 1209-1217.

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

Grimes, P. W., Millea, M. J., & Woodruff, T. W. (2004). Grades - who's to blame? Student evaluation of teaching and locus of control. *The Journal of Economic Education*, *35*, 129-147.

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.

Gursoy, D., & Umbreit, W. T. (2005). Exploring students’ evaluations of teaching effectiveness: What factors are important? *Journal of Hospitality & Tourism Research, 29*, 91-109.

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

Handerer, J. (2014). Zwischen Natur-und Geisteswissenschaft [Between natural science and humanities]. In M. Krämer, U. Weger, & M. Zupanic (Eds.): *Psychologiedidaktik und Evaluation X* (pp. 3-9). Aachen, Germany: Shaker.

Harvey, L., & Green, D. (1993). Defining quality. *Assessment & Evaluation in Higher Education*, *18,* 9-34.

Hattie, J., & Marsh, H. W. (1996). The relationship between research and teaching: A meta-analysis. *Review of Educational Research, 66*, 507-542.

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

Helmke, A. (2007). *Unterrichtsqualität erfassen, bewerten, verbessern [Assess, rate and improve the quality of teaching]* (6th ed.). Seelze: Klett Kallmeyer.

Hertwig, R., & Stoltze, A. (2001). *Beweggründe, Psychologie zu studieren: unterliegen sie einem zeitlichen Wandel und sind sie fachspezifisch?* [*Reasons for studying psychology: Are they subject to a temporal change and are they subject specific?*]. Retrieved from https://www.mpib-berlin.mpg.de/volltexte/institut/dok/full/hertwig/hrbew__01/hrbew __01.html

Hessisches Hochschulgesetz [Hessian Universities Act] (2004).

Hidi, S. (2001). Interest, reading, and learning: Theoretical and practical considerations. *Educational Psychology Review*, *13*, 191-209.

Hofmann, H., & Stiksrud, A. (1993). Wege und Umwege zum Studium der Psychologie III [Paths and detours to the study of psychology III]. *Psychologische Rundschau*, *44*, 250-256.

Holmes, D. S. (1972). Effects of grades and disconfirmed grade expectancies on students’ evaluations of their instructor. *Journal of Educational Psychology, 63*, 130-133.

Holmes, J. D., & Beins, B. C. (2009). Psychology is a science: At least some students think so. *Teaching of Psychology, 36,* 5-11.

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.

Keeley, J. W., English, T., Irons, J., & Henslee, A. M. (2013). Investigating halo and ceiling effects in student evaluations of instruction. *Educational and Psychological Measurement*, *73*, 440-457.

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

Kenny, D. A. (1994). *Interpersonal perception: A social relations analysis.* New York: Guilford Press.

Kherfi, S. (2011). Whose opinion is it anyway? Determinants of participation in student evaluation of teaching. *Journal of Economic Education*, *42*, 19-30.

Kirschner, P. A. (2002). Cognitive load theory: Implications of cognitive load theory on the design of learning. *Learning and Instruction, 12*, 1-10.

Kreber, C. (2003). The relationship between students' course perception and their approaches to studying in undergraduate science courses: A Canadian experience. *Higher Education Research & Development*, *22*, 57-75.

Kromrey, H. (1994a). Wie erkennt man „gute Lehre“? Was studentische Vorlesungsbefragungen (nicht) aussagen [How do you recognize “good teaching”? What student evaluations of teaching in lectures do (not) reveal]. *Empirische Pädagogik*, 8, 153-168.

Kromrey, H. (1994b). Evaluation der Lehre durch Umfrageforschung? Methodische Fallstricke bei der Messung von Lehrqualität durch Befragung von Vorlesungsteilnehmern. [Evaluation of the teaching through surveys research? Methodical pitfalls in the measurement of teaching quality by interviewing participants of lectures.] In P. Mohler (Ed.), *Universität und Lehre. Ihre Evaluation als Herausforderung an die empirische Sozialforschung* (pp. 105-128). Münster: Waxmann-Verlag.

Kulik, J. A. (2001). Student ratings: Validity, utility and controversy. *New Directions for Institutional Research, 27*, 9-25.

Leamon, M. H., & Fields, L. (2005). Measuring teaching effectiveness in a pre-clinical multi-instructor course: A case study in the development and application of a brief instructor rating scale. *Teaching and Learning in Medicine*, *17,* 119-129.

Lease, A. M., Musgrove, K. T., & Axelrod, J. L. (2002). Dimensions of social status in preadolescent peer groups: Likability, perceived popularity, and social dominance. *Social Development*, *11*, 508-533.

Leckie, G. (2013). Cross-classified multilevel models: Concepts. *LEMMA VLE Module 12*, 1-60.

Lehmann, H. E., Ban, T. A., & Donald, M. (1965). Rating the rater: An experimental approach to the methodological problem of interrater agreement. *Archives of General Psychiatry*, *13*, 67-75.

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. (1982a). The use of path analysis to estimate teacher and course effects in student ratings of instructional effectiveness. *Applied Psychological Measurement, 6,* 47-59.

Marsh, H. W. (1982b). 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. (1983). Multidimensional ratings of teaching effectiveness by students from different academic settings and their relation to student/course/instructor characteristics. *Journal of Educational Psychology, 75*, 150-166.

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

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

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., & Overall, J. U. (1980). Validity of students’ evaluations of teaching effectiveness: Cognitive and affective criteria. *Journal of Educational Psychology*, *72,* 468-475.

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.

Marsh, H. W., Fleiner, J., & Thomas, C. S. (1975). Validity and usefulness of student evaluations of instructional quality. *Journal of Educational Psychology, 67,* 833-839.

Marsh, H. W., Muthén, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, Alexandre J. S., & Trautwein, U. (2009). Exploratory structural equation modeling, integrating CFA and EFA: Application to students' evaluations of university teaching. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*, 439-476.

McKeachie, W. J. (1997). Student ratings: The validity of use. *American Psychologist, 52,* 1218-1225.

Menges, R. J., & Brinko, K. T. (1986, April). *Effects of student evaluation feedback: A meta-analysis of higher education research*. Paper presented at the meeting of the 70th American Educational Research Association, San Francisco, CA.

Merritt, D. J. (2008). Bias, the brain, and student evaluations of teaching. *St. John’s Law Review, 82,* 235-287.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: The Oryx Press.

Metje, B. (2009). *Validitätsprobleme von Lehrevaluationen. Eine Mixed Methods Studie [Validity problems of student evaluations of teaching. A mixed methods study]* (Doctoral dissertation). Philipps-Universität, Marburg.

Metje, B., & Kelle, U. (2010). Qualitätsentwicklung von Lehrveranstaltungsevaluationen durch Methodenkombinationen [Quality development of student evaluations of teaching through combination of methods]. In P. Pohlenz (Ed.), *Wie viel Wissenschaft braucht die Evaluation? Eine Einführung* (pp. 97-107). Bielefeld: UVW Universitätsverlag.

Metz-Göckel, S., Kamphans, M., & Scholkmann, A. (2012). Hochschuldidaktische Forschung zur Lehrqualität und Lernwirksamkeit [Didactical research on the quality of teaching and learning effectiveness]. *Zeitschrift für Erziehungswissenschaft, 15*, 213-232.

Mikkonen, J., Ruohoniemi, M., & Lindblom-Ylänne, S. (2013). The role of individual interest and future goals during the first years of university studies. *Studies in Higher Education*, *38*, 71-86.

Moosbrugger, H., & Kelava, A. (2008). Qualitätsanforderungen an einen psychologischen Test (Testgütekriterien) [Quality requirements for a psychological test (quality criteria of tests)]. In H. Moosbrugger & A. Kelava (Eds.), *Testtheorie und Fragebogenkonstruktion* (pp. 7-26). Heidelberg: Springer.

Moosbrugger, H., & Schweizer, K. (2002). Evaluationsforschung in der Psychologie [Evaluation research in psychology]. *Zeitschrift für Evaluation, 1*, 19-37.

Murray, H. G. (1983). Low-inference classroom teaching behaviors and student ratings of college teaching effectiveness. *Journal of Educational Psychology, 75*, 138-149.

Murray, H. G., & Lawrence, C. (1980). Speech and drama training for lectures as a means of improving university teaching. *Research in Higher Education*, *13*, 73-90.

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.

Onwuegbuzie, A. J., Daniel, L. G., & Collins, K. M. T. (2009). A meta-validation model for assessing the score-validity of student teaching evaluations. *Quality & Quantity, 43*, 197-209.

Orlik, P., Fisch, R., & Saterdag, H. (1971). Soziale Orientierung bei Studienanfängern der Psychologie [Questions of social orientation of freshman in psychology]. *Psychologische Rundschau*, *22*, 17-37.

Ory, J. C. (2001). Faculty thoughts and concerns about student ratings. *New Directions for Teaching and Learning*, 87*,* 3-15.

Ory, J. C., & Ryan, K. (2001). How do student ratings measure up to a new validity framework? *New Directions for Institutional Research, 109,* 27-44.

Overall, J. U., & Marsh, H. W. (1979). Midterm feedback from students: Its relationship to instructional improvement and students' cognitive and affective outcomes. *Journal of Educational Psychology*, *71*, 856-865.

Paget, N. (1984). An interesting bias on evaluation. In J. Lublin (Ed.), *Research and development in higher education* (Vol. 7, pp. 282-286). Sydney: HERDSA.

Patrick, C. L. (2011). Student evaluations of teaching: Effects of the Big Five personality traits, grades and the validity hypothesis. *Assessment & Evaluation in Higher Education, 36,* 239-249.

Peiffer, H., Rach, H., Rosanowitsch, S., Wörl, J., & Schneider, M. (2015). Lehrevaluation [Evaluation of courses]. In M. Schneider & M. Mustafić (Eds.), *Gute Hochschullehre: Eine evidenzbasierte Orientierungshilfe. Wie man Vorlesungen, Seminare und Projekte effektiv gestaltet* (pp. 153–184). Berlin, Heidelberg: Springer.

Penny, A. R. (2003). Changing the agenda for research into students' views about university teaching: Four shortcomings of SRT research. *Teaching in Higher Education*, *8*, 399-411.

Perry, P. B. (1995). Defining and measuring the quality of teaching. In D. Green (Ed.)*, What is quality in higher education?* (pp. 31-38). Buckingham, Bristol PA: Open University Press.

Pham, G., Koch, T., Helmke, A., Schrader, F. W., Helmke, T., & Eid, M. (2012). Do teachers know how their teaching is perceived by their pupils? *Procedia-Social and Behavioral Sciences*, *46*, 3368-3374.

Pinto, M. B., & Mansfield, P. M. (2010). Thought processes college students use when evaluating faculty: A qualitative study. *American Journal of Business Education, 3*(3), 55-62.

Pötschke, M. (2010). Mehrebenenanalyse: Angemessene Modellierung von Evaluationsdaten [Multi-Level Analysis: Adequate modelling of evaluation data]. In P. Pohlenz & A. Oppermann (Eds.), *Wie viel Wissenschaft braucht die Evaluation? Eine Einführung* (pp. 109-122). Bielefeld: UVW Universitätsverlag.

Pounder, J. S. (2007). Is student evaluation of teaching worthwhile? An analytical framework for answering the question. *Quality Assurance in Education, 15*, 178-191.

Power, M. (1997). *The Audit Society*. Oxford: Oxford University Press.

Prosser, M. (2011). Student "satisfaction" or student "experience": Interpreting and acting on evaluation results. In M. Saunders, P. Trowler, & V. Bamber (Eds.), *Reconceptualising evaluation in higher education: The practice turn* (pp. 46-50). Berkshire: Open University Press.

Pruitt, J. R., Dicks, M. R., & Tilley, D. S. (2010). Do students have fixed classroom perceptions? *NACTA Journal, 54,* 39-44.

R Core Team (2015). *R: A language and environment for statistical computing* [Computer program]. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

R Core Team (2016). *R: A language and environment for statistical computing* [Computer program]. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

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

Ramsden, P. (1996). *Learning to teach in higher education*. London: Routledge.

Rantanen, P. (2013). The number of feedbacks needed for reliable evaluation. A multilevel analysis of the reliability, stability and generalisability of students' evaluation of teaching. *Assessment & Evaluation in Higher Education, 38*, 224-239.

Rasbash, J., & Browne, W. J. (2008). Non-hierarchical multilevel models. In J. de Leeuw & E. Meijer (Eds.), *Handbook of multilevel analysis* (pp. 301-334). New York: Springer.

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.

Remedios, R., & Lieberman, D. A. (2008). I liked your course because you taught me well: The influence of grades, workload, expectations and goals on students' evaluations of teaching. *British Educational Research Journal, 34*, 91-115.

Renaud, R. D., & Murray, H. G. (1996). Aging, personality, and teaching effectiveness in academic psychologists. *Research in Higher Education*, *37*, 223-240.

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

Richter, T. (2006). What is wrong with ANOVA and multiple regression? Analyzing sentence reading times with hierarchical linear models. *Discourse Processes, 41*, 221-250.

Rindermann, H. (2003). Lehrevaluation an Hochschulen: Schlussfolgerungen aus Forschung und Anwendung für Hochschulunterricht und seine Evaluation [Teaching evaluation at universities: Conclusions from research and application for higher education and its evaluation]. *Zeitschrift für Evaluation, 2,* 233-256.

Rindermann, H. (2009). Qualitätsentwicklung in der Hochschullehre [Quality development in university teaching]. *Beiträge zur Lehrerbildung - Qualifizierung von Dozierenden der Hochschulstufe, 27,* 64-73.

Rindermann, H., & Schofield, N. (2001). Generalizability of multidimensional student ratings of university instruction across courses and teachers. *Research in Higher Education, 42,* 377-399.

Rindermann, H., Kohler, J., & Meisenberg, G. (2007). Quality of instruction improved by evaluation and consultation of instructors. *International Journal for Academic Development*, *12*, 73-85.

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

Rowley, M., Hartley, J., & Larkin, D. (2008). Learning from experience: The expectations and experiences of first-year undergraduate psychology students. *Journal of Further and Higher Education*, *32,* 399-413.

Sax, L. J., Gilmartin, S. K., & Bryant, A. N. (2003). Assessing response rates and non-response bias in web and paper surveys. *Research in Higher Education, 44,* 409-432.

Schacter, J., & Thum, Y. M. (2004). Paying for high-and low-quality teaching. *Economics of Education Review*, *23*, 411-430.

Schmidt, B., & Loßnitzer, T. (2010). Lehrveranstaltungsevaluation: State of the art, ein Definitionsvorschlag und Entwicklungslinien [Student Evaluation of Teaching: State of the Art, a Proposal for Definition and Guidelines for Development]. *Zeitschrift für Evaluation*, *9,* 49-72.

Schwarz, N. (2011). Feelings-as-information theory. In: P. A. M. Van Lange, A. W. Kruglanski, & E. T. Higgins (Eds.), *Handbook of theories of social psychology* (Vol. 1, pp. 289-308). London: Sage.

Sears, J. (1921). The Measurement of Teaching Efficiency. *The Journal of Educational Research, 4,* 81-94.

Seijts, G. H., & Latham, G. P. (2011). The effect of commitment to a learning goal, self-efficacy, and the interaction between learning goal difficulty and commitment on performance in a business simulation. *Human Performance*, *24,* 189-204.

Servicestelle Lehrevaluation (2007). FAQ - Warum kann ich meine Veranstaltung erst ab einer Mindestzahl von sechs Teilnehmern evaluieren lassen? [FAQ – Why can I only have my course evaluated from a minimum of six participants?]. Retrieved from https://www.psycho.uni-osnabrueck.de/service%20&%20beratung/servicestelle%20lehrevaluation/lehrevaluation/faq/Lehreval-FAQ.html#frage1

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420-427.

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.

Solomon, D. J., Speer, A. J., Rosebraugh, C. J., & DiPette, D. J. (1997). The reliability of medical student ratings of clinical teaching. *Evaluation & the Health Professions, 20,* 343-352.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3,* 271-295.

Spiel, C., & Gössler, P. M (2000). Zum Einfluß von Biasvariablen auf die Bewertung universitärer Lehre durch Studierende [On the influence of bias variables on student evaluation of teaching]. *Zeitschrift für Pädagogische Psychologie, 14,* 38-47.

Spiel, C., Wolf, P., & Popper, V. (2002). Lehre und Lehrevaluation – (un)geliebt? Die Perspektive der Universitätslehrenden [Teaching and teaching evaluation - (un)loved? The perspective of university teachers]. *Zeitschrift für Psychologie*, 210, 27–39.

Spooren, P. (2010). On the credibility of the judge: A cross-classified multilevel analysis on students' evaluation of teaching. *Studies in Educational Evaluation, 36,* 121-131.

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.

Stanton, H. E. (1995) Lecturer Confidence Building: The Use of Age Progression and Ego Enhancement. *Higher Education Research & Development, 14,* 121-127.

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.

Theall, M., & Franklin, J. (2001). Looking for bias in all the wrong places: A search for truth or a witch hunt in student ratings of instruction? *New directions for institutional research*, *109*, 45-56.

Thompson, R. A., & Zamboanga, B. L. (2003). Prior knowledge and its relevance to student achievement in introduction to psychology. *Teaching of Psychology, 30,* 96-101.

Tigelaar, D. E., Dolmans, D. H., Wolfhagen, I. H., & van der Vleuten, C. P. M. (2004). The development and validation of a framework for teaching competencies in higher education. *Higher Education, 48*, 253-268.

Toland, M. D., & De Ayala, R. J. (2005). A multilevel factor analysis of students' evaluations of teaching. *Educational and Psychological Measurement, 65*, 272-296.

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.

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

van Merrienboer, J. J., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational Psychology Review*, *17,* 147-177.

Vervecken, D., Ulrich, I., Braun, E., & Hannover, B. (2010). Lehre und Studium professionell evaluieren: Kompetenzorientierte Lehrevaluation mit BEvaKomp [Professionally evaluating teaching and academic studies: Competence-oriented teaching evaluation with BEvaKomp]. In P. Pohlenz (Ed.), *Qualität - Evaluation - Akkreditierung: Vol. 4. Wie viel Wissenschaft braucht die Evaluation? Eine Einführung* (pp. 153-165). Bielefeld: UVW Universitäts Verlag.

Volante, L. (2007). Educational quality and accountability in Ontario: Past, present, and future. *Canadian Journal of Educational Administration and Policy*, *58,* 1-21.

Weinberg, B. A., Hashimoto, M., & Fleisher, B. M. (2009). Evaluating teaching in higher education. *The Journal of Economic Education*, *40*, 227-261.

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. (2010). Attraktivität, Geschlecht und Lehrveranstaltungsevaluation [Attractiveness, gender and student evaluation of teaching]. *Zeitschrift für Evaluation*, *9*, 29-48.

Wolbring, T. (2012). Class attendance and students' evaluations of teaching: Do no-shows bias course ratings and rankings? *Evaluation Review*, *36*, 72-96.

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.

Wolfe, E. W. (2004). Identifying rater effects using latent trait models. *Psychology Science*, *46*, 35-51.

Wood, K., Linsky, A. S., & Straus, M. A. (1974). Class size and student evaluations of faculty. *The Journal of Higher Education*, *45,* 524-534.

Woodhall, M. (1987). Econimics of education: A review. In G. Psacharopoulos (Ed.). *Economics of education: Research and studies* (pp. 1-10). Oxford: Pergamon.

Yunker, P. J., & Yunker, J. A. (2003). Are student evaluations of teaching valid? Evidence from an analytical business core course. *Journal of Education for Business*, *78*, 313-317.

Zhao, J., & Gallant, D. J. (2012). Student evaluation of instruction in higher education: Exploring issues of validity and reliability. *Assessment & Evaluation in Higher Education*, *37*, 227-235.

APPENDIX

Appendix A: Items Measuring General Preconception of Psychology

(German/English Translation)

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

(1) Mit dem Begriff “Psychologie” verbinde ich klare Vorstellungen.
I have a clear concept of the term “psychology”.

(2) Ich habe eine klare Vorstellung von den Inhalten des Psychologiestudiums.
I have a clear concept of the contents of the psychology program

(3) Ich habe eine klare Vorstellung über mögliche Tätigkeitsbereiche von PsychologInnen.
I have a clear idea about where psychologists can work.

(4) Ich habe bereits klare Vorstellungen über meine eigene angestrebte Berufstätigkeit als Psychologe/in.
I have a clear idea about what I want to do later with my degree.

(5) Für mich kam nur ein Psychologiestudium in Frage.
I only wanted to study psychology.

Appendix B: Items Measuring Specific Preconceptions of Psychology

(German/English Translation)

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

(1) Ich habe eine klare Vorstellung von den Inhalten des Studienfaches Allgemeine Psychologie.

I have a precise idea about the contents of cognitive, motivational and emotional psychology.

(2) Ich habe eine klare Vorstellung von den Inhalten des Studienfaches Biologische Psychologie.

I have a precise idea about the contents of biological psychology.

(3) Ich habe eine klare Vorstellung von den Inhalten des Studienfaches Entwicklungspsychologie.

I have a precise idea about the contents of developmental psychology.

(4) Ich habe eine klare Vorstellung von den Inhalten des Studienfaches Sozialpsychologie.

I have a precise idea about the contents of social psychology.

(5) Ich habe eine klare Vorstellung von den Inhalten des Studienfaches Psychologische Methodenlehre.

I have a precise idea about the contents of psychological methods.

(6) Ich habe eine klare Vorstellung von den Inhalten des Studienfaches Pädagogische Psychologie.

I have a precise idea about the contents of pedagogical psychology.

(7) Ich habe eine klare Vorstellung von den Inhalten des Studienfaches Arbeits- und Organisationspsychologie.

I have a precise idea about the contents of Business psychology.

(8) Ich habe eine klare Vorstellung von den Inhalten des Studienfaches Klinische Psychologie.

I have a precise idea about the contents of clinical psychology.

(9) Ich habe eine klare Vorstellung von den Inhalten des Studienfaches Umweltpsychologie.

I have a precise idea about the contents of environmental psychology.

(10) Ich habe eine klare Vorstellung von den Inhalten des Studienfaches Psychologische Diagnostik.

I have a precise idea about the contents of psychological diagnostics.

Appendix C: 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.

Acknowledgements

At this point I would like to thank all those who supported and motivated me during this dissertation. At first and foremost, I thank Tobias Richter for his advice, support, feedback and his contribution to this dissertation.

I also thank several people for reading and commenting my drafts: Benjamin Franklyn Burton, Janett Zacher, Julia Schindler, Marion Stein, Martina Vogelsang, Michael Cofrin, Susanne Jurkowski, and my anonymous reviewers. In addition, I thank Bettina Müller, Christiane Braun, Elke Döring-Seipel, Janett Zacher, Johanna Maier, Johannes Zimmermann, Juliane Wilcke, Marion Stein, Martina Vogelsang, Silja Kotte, Sarah von der Mühlen, and Susanne Jurkowski for their input, emotional support, lots of patience and time. I will be always grateful.

This dissertation would have been impossible without the use of the FEVOR questionnaire and the many participants. Therefore I thank Thomas Staufenbiel for the permission to use it at the University of Kassel. I also thank Verena Lössl for the provided the scales general and specific preconceptions of psychology (Appendices A and B) and my student assistants for their support in scanning and controlling data. I am grateful for the patient participation of students and colleagues of the department of psychology. They helped immensely in the creation of this dissertation.

My final thanks go to Martin Hänze and Holger Horz for their interest in this topic and their willingness to examine it.