Counseling University Instructors Based on Student Evaluations of Their Teaching Effectiveness: A Multilevel Test of its Effectiveness Under Consideration of Bias and Unfairness Variables

Markus Dresel · Heiner Rindermann

Abstract Counseling instructors using evaluations made by their students has shown to be a fruitful approach to enhancing teaching quality. However, prior experimental studies are questionable in terms of external validity. Therefore, we conducted a non-experimental intervention study in which all of the courses offered by a specific department at a German university were evaluated twice with a standardized student evaluation questionnaire (HILVE-II; overall 44 instructors, 140 courses, and 2,546 student evaluations). Additionally, twelve full time instructors received counseling after the first measurement point. Long-term effects over a period of 2 years and transfer effects to other courses were analyzed using multi-level analyses with three levels. Possible influences by bias and unfairness variables were controlled for. Our results indicate a moderate to large effect of counseling on teaching quality. In conclusion, if students' evaluations are accompanied by counseling based on the evaluation results, they present a useful method to assure and increase teaching quality in higher education.

Keywords Counseling of instructors · Students' evaluations of teaching quality · Bias and unfairness variables · Multi-level analysis

Introduction

Students' evaluations of courses which employ questionnaires covering different dimensions of teaching are a widely established procedure for measuring teaching quality and teaching effectiveness in higher education. The main objective, in terms of the main communicated aim in many cases, is the improvement of instructional quality. In reasonable evaluation approaches quality measurement should not be an end in itself, but rather a step along the path to high quality in teaching and higher education (for an

M. Dresel
Department of Psychology, University of Augsburg, Universitätsstr. 10, 86135 Augsburg, Germany
e-mail: markus.dresel@phil.uni-augsburg.de

H. Rindermann
Department of Psychology, Chemnitz University of Technology, Chemnitz, Germany

overview see Perry and Smart 2007). The majority of realized evaluation approaches are subject to the—at least implicit—assumption that improved teaching quality will result from either assessments of teaching which mediate an increased quality awareness among instructors ("sensitization or awareness model"; e.g. Will and Blickhan 1987), or from additional feedback of student evaluation results to instructors ("feedback model").

However, empirical research shows that a lack of follow-up activities will impair the development of substantial improvements to teaching quality (for overviews see Marsh 2007a; Rindermann 2009). One measure which has proven to be effective in a range of experiments is faculty counseling based on student course evaluations (Penny and Coe 2004). The majority of these studies used experimental designs with random assignments to experimental (counseling) or control (no counseling) groups. Despite a high internal validity among these studies, the external validity of their findings is generally problematic, moreover, the generalizability of results based on North-American and Anglo-Saxon institutional contexts to other tertiary education systems is questionable.

In order to overcome these shortcomings, the present study aims to demonstrate the long-term, cross-course, cross-course-format and "pure" (control of bias effects at different levels) effects of student evaluations of teaching quality in regular course instruction when accompanied by counseling, under the institutional conditions prevailing at German universities.

Findings for the Sensitization Model and the Feedback Model

Empirical studies which have tested the "sensitization model" (assumption: the mere use of teaching quality assessments increases the awareness of the relevance for teaching quality in instructors and thus leads to improvements in teaching quality) have consistently shown negligible or no effects (e.g. Erickson and Erickson 1979; Marsh and Hocevar 1991; McKeachie et al. 1980). The effect sizes in students' evaluations of teaching quality are around zero for this model (Rindermann 2009). Studies which examined the effects of the "feedback model" (assumption: feedback of students' evaluations lead to improvements in teaching quality) resulted in either no improvement in teaching quality or modest ones at best (see overviews in Marsh 2007a, b; Menges and Brinko 1986; l'Hommedieu et al. 1990; Rindermann 2009). Menges and Brinko (1986) conducted a meta-analysis of 23 studies which assessed the effects of evaluation feedback determined an increased teaching quality of only $d = 0.22$. These findings lead one to conclude that the goal of improving teaching quality with institutionalized evaluations is being missed at many universities: Nearly all of the evaluation approaches realized, at least in German-speaking universities in Germany, Austria and Switzerland, simply use student evaluations without the benefit of further measures such as counseling or training.

But why does feedback show rather small or zero effects? Three possible reasons are discussed in the literature: (1) Instructors ignore discrepancies between actual and desirable teaching behavior because feedback is registered only superficially, consequently the results are interpreted incorrectly or rejected entirely as instructors challenge the validity of course assessments made by students (e.g. Rotem 1978). (2) Instructors have a low motivation to improve their didactic skills, possibly due to a low perceived value to improve teaching or poor self-efficacy expectations (e.g. Marsh and Roche 1993). (3) Instructors have insufficient knowledge on how to improve their teaching e.g. by the use of specific didactic strategies (e.g. Marsh and Roche 1993; Murray and Lawrence 1980).

To rectify the first problem stated above, a "discourse model" was developed (Webler 1996): According to this model, instructors and students utilize a course-internal discussion

to assess and consider the results of the evaluation, specific problems and possibilities to improve the instructional process within the course. However, it has been shown by Webler (1996) and Rindermann (1996) that this approach is also less than sufficient in attaining the desired improvements, particularly since poorly evaluated teachers tend to avoid presentations and discussions of their evaluation results.

Counseling Instructors

Approaches in which feedback is combined with instructor counseling (for implementation see Brinko and Menges 1997; Knapper and Piccinin 1999; Rindermann and Kohler 2003; Rindermann et al. 2007) simultaneously address all three aspects—the perception of discrepancies, the motivation to improve teaching quality and the development of methods of enhancement: In a session with individual personal contact, counseling with a person who is familiar with both student evaluations of teaching effectiveness and university teaching practices can inform instructors of their results, help them overcome difficulties in the interpretation of these results, point out discrepancies between actual and aimed teaching behaviors, initiate an analysis of the possible causes for problematic findings, and replace patterns of explanations which are threatening to both motivation and self-esteem with functional causal explanations and thereby improve self-efficacy (attributional retraining; for a summary see Foersterling 1985). These strategies should emphasize the positive consequences of improvements in teaching behavior. Consultants should select aspects of teaching behavior which can be modified. Concrete suggestions for improvement should be developed in close cooperation with the counseled instructor, e.g. by the application of teaching strategies observed among excellent teachers (e.g. Feldman 2007; McKeachie 2007). Overall, the functions of the counseling sessions can be seen as providing feedback and a simplification of the results of the evaluation, confronting problem areas, focusing on dimensions of teaching quality which are behaviorally tangible and controllable, motivating and activating instructors as well as expanding their didactic behavioral repertoire, and offering suggests for further training to improve teaching quality.

Experimental and Quasi-Experimental Results for the Counseling Approach

A number of experimental and quasi-experimental studies conducted in North America, Australia and the UK have shown counseling with university teachers on the basis of student evaluations to be effective (e.g. Brinko 1990; Bray and Howard 1980; Cohen 1991; Hampton and Reiser 2004; Marsh and Roche 1993; McKeachie et al. 1980; Piccinin et al. 1999; Wilson 1986; overview in Penny and Coe 2004). For example, in a quasi-experiment conducted by Wilson (1986), student assessments were conducted in the middle of a semester. The instructors in the experimental group were counseled with regard to the teaching techniques of colleagues who had received particularly good evaluations. Instructors in a control group were only provided with written feedback on the results of the evaluation. At the end of the semester only the members of the first group were found to have significantly improved teaching quality in terms of students' evaluations. Marsh and Roche (1993) were able to replicate these findings experimentally: Their counseling sessions utilized "optimal teaching behaviors", which were drawn on the didactic activities addressed in their questionnaire. Using a comparable experimental design, Hampton and Reiser (2004) were also able to confirm similar results with student tutors who received counseling during the middle of a semester. Noteworthy here is that, following the counseling procedure, improved learning results were even recorded among the students

participating in the classes led by these tutors. In their meta-analysis, Penny and Coe (2004) examined the results of 12 experimental studies and came to the conclusion that using student evaluations as a basis for advising university teachers is an effective strategy for improving teaching quality, and quantified the effect size of counseling at $d = 0.69$. A previous meta-analysis conducted by Menges and Brinko (1986) actually yielded an effect size of $d = 1.10$.

To date, only one study has examined the effectiveness of the counseling approach in a German language university environment (Rindermann et al. 2007; see also Rindermann and Kohler 2003). Providing teachers with counseling on the basis of the results of student evaluations was able to generate substantial improvements on several dimensions of instructors' teaching behavior ($d = 0.65$) as well as overall teaching quality ($d = 0.64$). However, the study was performed at a private school for speech therapists where, in many respects, the learning conditions were more similar to the institutes examined in North American studies (small numbers of classroom participants, teaching quality is an important attribute of institutional quality, students are seen as customers) than those generally found at public German universities (e.g. large courses with up to 800 participants, almost always publicly financed). Therefore, a simple transfer of the evaluation context to that prevailing at German state universities can not be warranted, and the generalizability of the results presented above to universities outside of North America, Australia and the UK is still an open question.

In addition to the major difficulties inherent in generalizing the referenced experimental findings, three other problems complicate the interpretation of the experimental results presented above, namely the questionable external validity, the disregard of bias and unfairness variables and the inadequate representation created by using a single-level design.

Questionable External Validity

In most studies which examined the effectiveness of counseling university instructors, a randomly selected training group was counseled in the middle of a semester on the basis of course assessments completed by their students. The counseling effects were evaluated after another round of student assessments which was conducted for the same courses at the end of the semester and comparisons were made to similar assessments of instructors in a control group. This design is fraught with a number of problems in terms of external validity (see l'Hommedieu et al. 1990; Marsh 2007a). Probably the largest problem with this type of design is that it neglects the long-term effects of counseling. One of the few exceptions is the study done by Piccinin et al. (1999): They examined the effects of teacher counseling over a period of 3 years among university course instructors co-operating with a didactic counseling center on a voluntarily basis. The results of evaluation and counseling obtained in a natural context indicate a persistent efficiency of teacher counseling, even under circumstances which produced divergent intensities of counseling (see also Murray 1997). The work conducted by Piccinin et al. (1999) emphasizes the dilemma involved in securing both internal and external validity: To produce a high generalizability of results in natural contexts, the counseling sessions need to be implemented in the regular teaching environment and the long-term effects of counseling must be examined. One consequence of choosing this research context would be that ensuring internal validity with a control group (composed of randomly assigned university instructors who receive neither feedback on course evaluation nor counseling) would not only be difficult to justify, but also impossible to realize in practice.

Directly connected to the problem of ignoring the long-term effects of counseling are the difficulties arising from the failure to investigate transfer effects to courses other than those for which the teachers have been counseled. A cross-course transfer would primarily be hindered by differences in the characteristics and conditions pertinent to the courses (e.g. course topic, course format, position in curriculum, class size, prior knowledge of students and prior interest of students). So, for example, the teaching strategies recommended during counseling sessions which target evaluations from an introductory lecture (participants with e.g. low prior knowledge) may not be useful, or even effective, in advanced seminars with a smaller number of participants (e.g. extensive prior knowledge). The transfer of effects to other situations and demands is not guaranteed.

Methodological Prerequisites for Externally Valid Investigations

Securing proof of long term counseling effects across several semesters, and of transfer effects over different courses and course formats, represents a difficult methodological task since student assessments of teacher quality can be distorted by an array of bias and/or unfairness variables (e.g. Greenwald 1997; for an overview see Marsh 2007a). These variables include characteristics which are impossible (or at least difficult) to influence by the teacher, but do influence students' perceptions of teaching quality (bias; e.g. topic of the course) or teaching quality itself (unfairness; e.g. prior knowledge or interest; Centra 2003; Rindermann 2009). Discussions of bias and unfairness variables have been wide reaching in the literature and include, individual prior interest, prior knowledge, perceptions of course relevance, effort and semester of the students, age and gender of students and instructors, course size and course format, compulsory versus voluntary attendance, as well as the diversity of prior knowledge and interest in the course (see Marsh 2007a; Marsh and Roche 1997; Rindermann 1996). Research on bias in student course evaluations has shown that only few bias variables (such as the individual prior interest in the course topic) have a general influence on students' evaluations independent of specific features of courses; in contrast it appears that a vast number of (potential) bias variables cause students to make different judgments in different courses, i.e. the distorting influence of individual characteristics is apparently moderated by course characteristics (e.g. Spiel and Gössler 2000). Considering the large number of possible sources of bias and unfairness, the examination of the long-term and transfer effects of instructor counseling for which dependent samples of instructors and independent samples of courses and course participants are required, must ensure that evaluations of teaching quality can be compared across different courses (or the same courses led by the same instructors, but in different semesters, with different course participants). They should also encompass almost all (potential) bias and unfairness variables and control their influences.

Representations of Evaluation Data Nested in Multiple Levels

A general methodological problem facing the analysis of students' evaluations of teaching effectiveness in university courses is the determination of the appropriate level of analysis ("unit of analysis problem"; see Abrami et al. 2007; Cranton and Smith 1990; Marsh 2007a). The data collected in course evaluations represent a hierarchical structure which comprises at least two levels: Student assessments (Level 1) are nested within courses (Level 2), which are conducted by course instructors. Should these instructors be responsible for several courses in the same semester—as often is the case—course data is then also nested within the instructors (Level 3). In the vast majority of previous studies on

university course evaluations, in general and with specific regard to the effectiveness of university teacher counseling, the hierarchical data structure was not systematically taken into account. Instead, the analysis was typically conducted on an aggregated data level. When investigating the validity of student assessments the choice is usually the course level; in contrast, studies on the effectiveness of feedback or counseling normally use the instructor level (Marsh 2007a; Marsh and Roche 1997). Examples of the few exceptions on this point include studies conducted by Cranton and Smith (1990), Ting (2000) as well as Wendorf and Alexander (2005). Literature on the analysis of multi-level data has documented a series of problems and potential sources of errors which are associated with the neglect of the multilevel structure of data (see Snijders and Bosker 2002). Specifically with respect to the present context of analyzing counseling effects the aggregation of data on the instructor level results in a loss of information on differences in teaching quality in different courses led by the same instructors (i.e. the within-instructor variance). Moreover, working only at the course level cannot adequately control student bias variables (including possible interactions across levels, e.g. between prior interest and course format) and thus diminishes test power (see Ting 2000). In order to avoid these problems, the effects of counseling university instructors should be analyzed by means of multi-level analysis.

Aims of the Present Study

The central aim of the present study is to examine the effectiveness of instructor counseling based on student evaluations of their courses. The focus is on the long-term effects of counseling, including transfer to courses other than those targeted by the counseling sessions. From a methodological perspective, a further concern of the present work is to assess the usefulness of the multi-level approach in the analysis of student assessments, especially with regard to the control of bias variables.

Method

Procedure

At a public university of applied sciences in Southern Germany, all of the courses comprising the curriculum of the Architecture program were evaluated twice within a period of 2 years. Both evaluations took place in the middle of the winter semester. At both evaluations, the majority of the courses were being conducted by professional instructors (professors of Architecture with life-long and full-time appointments), with (compared to universities in North America, UK and Australia) a relatively high course load of 18 classroom hours per week, which is the norm at German universities of applied sciences (teachers at other forms of German public universities: average between 8 and 13, and up to 18 h per week). A number of part-time instructors were also leading courses in the department. Each student cohort comprised around 75 students. Course formats included lectures, laboratory courses, seminars, and internships. For several courses attendance was mandatory. Course evaluations were conducted in accordance with a decision by the departmental council.

For both evaluation dates, all faculty members were supplied with written feedback of their results. Professional full-time instructors additionally received counseling based on these results.

To analyze the effects of counseling, a multi-level analysis strategy was applied with three levels (students, courses, instructors) which included potential bias and unfairness variables at each level.

Sample

Across both evaluation points, students assessed a total of 140 courses held by 44 instructors. After exclusion of student data sets with missing entries (this occurred in 9.5% of all cases) our sample encompassed a total of 2,546 student assessments. In Table 1 the sample is broken down by measuring (evaluation) point, level of analysis, and whether or not the instructor was assessed at both of the measurement points in the evaluation. The data show that the 12 instructors who were evaluated at both measuring points had conducted the majority of the courses. This group consisted entirely of professional full-time instructors (eleven male and one female) and is identical to the group of teachers who received individualized counseling. The 32 course instructors who only participated at just one of the measuring points were primarily part-time teachers, and each of them had only conducted one course with a small number of participants (frequently elective courses). Also belonging to this second group were two professors who retired following the first evaluation period, as well as three professors (two male and one female) who were giving their first classes when the second wave of evaluations were being conducted.

With regard to both the comparability of instructor-related results at the two measuring points as well as the transfer of counseling effects, it is remarkable that a majority of the instructors held different courses at the second measuring point than at the first measuring point: On average, only one quarter of the courses offered by the instructors were repeated at the second measuring point (28%).

Instrument and Measures Included in the Present Analyses

The "Heidelberger Inventar zur Lehrveranstaltungsevaluation II" was used at both measuring points as the evaluation instrument (*HILVE-II*, Heidelberg Inventory for Evaluation of Teaching II; Rindermann and Schofield 2001). The HILVE is a multidimensional instrument to assess students' evaluations of teaching effectiveness and is frequently used

Table 1 Sample description with respect to instructor group, level of analysis and measuring point (MP)

Instructor group	Level of observation	MP 1	MP 2	Total
Instructors who were evaluated at both MPs	Instructors	12[a]	12[a]	12[a]
	Courses	44	54	98
	Students	941	1,093	2,034
Instructors who were evaluated at one MP[b]	Instructors	17	15	32
	Courses	23	19	42
	Students	234	279	513
Total	Instructors	29	27	44
	Course	61	73	140
	Students	1,176	1,372	2,546

[a] Same instructors, professional full-time instructors with high course load

[b] Instructors participated only at one MP due to appointment or retirement (full-time instructors) and due to beginning or terminating of an teaching assignment (instructors with temporary part-time agreements)

in various German-speaking university contexts (for an overview see Rindermann 2009). The inventory is similar to the instrument that has probably been used most widely in English-speaking post-secondary contexts, the *SEEQ* (Students' Evaluations of Educational Quality; Marsh 1982; see also Richardson 2005). It consists of a total of 48 items. Measured dimensions of teaching effectiveness concern the overall teaching quality (learning and general course evaluation, similar to the dimension "learning/value" in the SEEQ), the teacher and his or her behavior (e.g. instructor enthusiasm, structure and teaching competence, the latter two can be seen as represented in the SEEQ-dimension "organization/clarity"), the interaction between teacher and students (e.g. the encouragement and participation of students in discussions; similar to SEEQ's "group interaction"), conditions of the course topic (e.g. course demand similar to "workload/difficulty" in the SEEQ), as well as potential bias/unfairness variables (e.g. students' prior interest in the course topic, students' individual effort). Moreover, the HILVE contains three open questions that permit course participants to acknowledge positive and negative aspects of the course and to offer possible suggestions for improvement.

In a series of studies of the psychometric quality of the HILVE it was demonstrated that it is a reliable, valid and useful instrument to assess teaching effectiveness (e.g. Rindermann and Schofield 2001). Rindermann (2009) reported mean interrater-reliabilities of $r = 0.81$ for all subscales (calculated for 10 students), mean internal consistencies of Cronbachs $\alpha = 0.81$, mean within term retest-reliabilities of $r = 0.67$ and mean generalizability scores for teacher subscales of $r = 0.46$ (correlations between evaluations of different courses taught by the same teacher). Moreover, they provided evidence which supported the hypothesized factor structure of the inventory.

Subject of the written feedback and the counseling were all dimensions of the inventory. However, the statistical analyses on the improvements of teaching effectiveness through counseling instructors focused on the overall teaching quality (as general indicator of teaching effectiveness) and additionally incorporated only potential bias and fairness variables.

Teaching Quality

As a dependent variable, a scale consisting of five HILVE-II items measuring perceived overall teaching quality were used. Included are learning growth, which is considered the central goal of all teaching ("I learned a lot in this course" and "I have a more fundamental understanding of this topic than I had before the course"), the value of the learned content ("I am learning something useful and important"), the encouragement of interest ("This course stimulated my interest in pursuing further studies in this area") as well as the general overall evaluation of the course ("Taking this course is worthwhile"). The scale ranges from the poles 1 (*absolutely false*) to 7 (*absolutely true*). The internal consistency of the compounded scale reaches Cronbachs $\alpha = 0.91$.

Potential Bias and Unfairness Variables on Student Level

Basic student variables considered to have a potentially bias influence were student age (in years) and student gender (1 = male, 0 = female). Further variables were: The individually perceived relevance of the course topic ("The topic of this course is itself relevant (occupationally/in practice/for my exams)"), individual prior interest in the course topic ("I was very interested in this topic before I enrolled in this course"), individual effort ("I prepare material for this course (e.g. readings) intensively, either before class meetings

or as a follow-up", "The amount of effort I apply to this course is relatively high in comparison to the effort I give to my other courses", $\alpha = 0.83$). All of the items were presented with a scale ranging from 1 (*absolutely false*) to 7 (*absolutely true*). Additionally, individual prior topic knowledge was assessed with the item "My prior knowledge level:" and a response scale ranging from 1 (*was too low to be able to follow the course*) over 4 (*was just right*) to 7 (*I already knew everything, it was a waste of my time*).

Potential Bias and Unfairness Variables on Course Level

Structural, course-related variables were assessed with an instructor questionnaire. Among these were the number of course participants (relative to the number of students who were qualified to attend the course), the course format (1 = lecture, 0 = other format), the predetermined program semester, whether attendance was compulsory, and whether a final examination was required (1 = yes, 0 = no). Furthermore, all potential student bias and unfairness variables, with the exception of student age and gender, were averaged on the course level in order to establish variables that reflect the level of the characteristic under consideration. E.g. it was assumed that mean course prior knowledge influences teaching quality (Rindermann and Heller 2005). In contrast to the variables considered on the student level, these aggregates are to be interpreted as estimates of course characteristics (see Snijders and Bosker 2002; Ting 2000). Independent of the course means of the student characteristics, it is also presumed that the diversity and/or uniformity of specific within-course student characteristics can exert an influence on assessments of teaching quality. In particular, it is expected that a large degree of diversity in prior knowledge and prior interest among course participants will have a negative influence on teaching efficiency (e.g. Ting 2000). In generating indicators for the diversity of prior knowledge and prior interest, the standard deviations for these two variables were calculated for each class in the study.

Potential Bias and Unfairness Variables on Instructor Level

On the instructor level potential bias variables were instructor status (1 = professor, 0 = other kind of instructor position), age (in years) and gender (1 = male, 0 = female).

Feedback and counseling approach

Written reports on the evaluation results were prepared for all instructors. These contained both unnormed as well as normed quantitative results. The normed values were also illustrated by means of profile diagrams. Furthermore, transcribed and grouped (by question) answers to the open questions were included.

The counseling sessions between the two evaluations were realized in a day-long workshop (8 h) which occurred in the closing days of the semester in which the first evaluation took place.[1] All of the full professors participated in the counseling workshop. The workshop and the counseling of instructors were designed, conducted and moderated by the two authors using their expertise as researchers and instructors in the field of higher education. To start off the session, the teachers conducted self-assessments of their own teaching quality by completing an instructor's version of the HILVE-II. Subsequently, in a 2 h session in plenum, the average results of the student assessments for the entire faculty

[1] An analogous workshop including counseling instructors was also conducted following the second evaluation.

were presented, whereby general strengths and weaknesses were pointed out and discussed, and strategies aimed to improve overall teaching quality were developed. This was followed by individual sessions, which lasted between 30 and 45 min, in which each instructor was counseled by one of the two authors privately. In this context, the professors received their personal written results along with explanations (part-time instructors received their feedback from the dean of the department). Using the profile diagrams and self-assessments as a basis, individual strengths and weaknesses were identified and discussed with each instructor. The emphasis was on behavior relevant process characteristics of teaching (e.g. how the material covered is structured, clarity in presenting the material, incorporation of communicative and cooperative forms of teaching, social classroom atmospheres). The instructors were encouraged to develop a causal analysis of their relatively weakest points, whereby such explanatory patterns considered to be detrimental to motivation and self-esteem were replaced with functional causal explanations (see Foersterling 1985). With regard to these relative weaknesses, strategies to improve teaching quality were worked out jointly. At the conclusion of the individual sessions, the focus was placed on individual valuing of improvements in teaching quality. Following the individual counseling sessions, a 60 min all-inclusive discussion was conducted, which gave the instructors opportunities to discuss the individual strategies for improvement they had developed amongst themselves as a group, and examine the evaluation and counseling process. The vast majority of the teachers' assessments of the evaluation procedure were positive.

Results

Several hierarchical linear regression models with three levels (student, course, instructor) were specified and tested with the *HLM 5.04* (Raudenbush et al. 2001). We present the results of our investigation in three sections. The first section contains the descriptive statistics and results obtained with the null model. Both are of great value in putting the later results into context. In the second section we present the bias model which was identified with the help of an analytical strategy, comprised of several sub-steps on the basis of the entire sample and takes into consideration potential bias influences on all three analytic levels. Finally, and most important in the present context, the third section depicts the estimation of the effect of counseling on university instructors made with this bias model.

Descriptive Statistics and the Three-Level Null Model

Descriptive statistics pertaining to the variables considered at each of the three investigative levels are displayed in Table 2. Included here are bivariate correlations with the dependent variable, teaching quality. Temporal stability (not corrected for bias) of teaching quality over the period of two years was $r = 0.37$.

In order to estimate the proportions of criterion variance to be specifically accorded the student, course and instructor levels, we first analyzed the three-level null model for the dependent variable, which is defined through the three equations $Y_{ijk} = \pi_{0jk} + e_{ijk}$ (student level), $\pi_{0jk} = \beta_{00k} + r_{0jk}$ (course level) and $\beta_{00k} = \gamma_{000} + u_{00k}$ (instructor level). Using this model, it is possible to decompose the entire variance of the dependent variable into three levels, represented by the following equation $Var(Y_{ijk}) = Var(e_{ijk}) + Var(r_{0jk}) + Var(u_{00k})$. $E = Var(e_{ijk})$ corresponds to the variance within courses, $R_0 = Var(r_{0jk})$ depicts

Table 2 Descriptive statistics and correlations with teaching quality on the three levels of analysis

Characteristic	Min	Max	*M*	*s*	r^a
Assessment of teaching quality on student level ($N_{Students} = 2{,}546$)					
Teaching quality	1.00	7.00	4.87	1.32	
Potential bias variables on student level ($N_{Students} = 2{,}546$)					
Student age (years)	18	60	23.04	3.43	0.02
Student gender[b]	0	1	0.47		0.07*
Perceived topic relevance	1.00	7.00	5.72	1.45	0.44*
Interest in topic	1.00	7.00	4.19	1.71	0.25*
Prior topic knowledge	1.00	7.00	3.98	0.91	0.03
Effort	1.00	7.00	3.38	1.73	0.30*
Potential bias variables on course level ($N_{Courses} = 140$)					
Number of course participants[c]	0.04	1.00	0.46	0.40	−0.39*
Course format[d]	0	1	0.39		−0.24*
Semester	1	7	3.11	2.25	0.05
Mandatory attendance[e]	0	1	0.07		−0.08
Final exam[e]	0	1	0.49		−0.17*
CM Perceived topic relevance	3.13	7.00	5.81	0.81	0.56*
CM Interest in topic	2.86	6.10	4.38	0.73	0.43*
CM Prior topic knowledge	2.83	5.33	3.99	0.32	0.03
CM Effort	1.27	6.70	4.07	1.53	0.43*
CV Interest in topic	0.00	2.65	1.56	0.37	−0.10
CV Prior knowledge on the topic	0.00	1.89	0.78	0.35	−0.35*
Potential bias variables on instructor level ($N_{Instructors} = 44$)					
Instructor status[f]	0	1	0.61		−0.13
Instructor age (years)	23	72	46.74	12.87	0.29
Instructor gender[b]	0	1	0.80		0.08

CM course mean (arithmetic mean of students' ratings for the individual course), *CV* course variance (standard deviation of students' ratings for the individual course)

*$p < 0.05$

[a] Correlation with teaching quality. In order to calculate correlations on the course and instructor levels, teaching quality was aggregated on these two higher levels

[b] 1 = male, 0 = female

[c] Percentage of students in the appropriate semester of their studies

[d] 1 = lecture, 0 = other

[e] 1 = Yes, 0 = No

[f] 1 = professor, 0 = other

the variance between courses and $U_{00} = Var(u_{00k})$ corresponds to the variance between instructors. The results obtained with the three-level null model are illustrated in Table 3. Variance components which are significantly different from null could be confirmed at both the course and instructor levels and indicated that courses and instructors were being assessed differently. Thus 32.9% of the total variance in student course assessments can be attributed to differences between courses, and 32.6% of this between-course variance can be traced to differences between instructors. Even if these variance proportions are rather high in comparison to those usually observed with multi-level investigations (Snijders and

Table 3 Variance Components and intra-class correlations (ICC) for the three-level null model with students' perceptions of teaching quality as a dependent variable, estimated using the entire sample

Variance component	Variance	df	χ^2	ICC
Variance within courses	1.16			0.671
Variance between courses	0.39	96	331.98***	0.222
Variance between instructors	0.19	43	97.77***	0.107

Notes $N_{\text{Students}} = 2{,}546$. $N_{\text{Courses}} = 140$. $N_{\text{Instructors}} = 44$. For course and instructor level, results of the χ^2 tests indicate whether or not the accordant variance component is greater than zero. The intra-class correlations ICC quantify the proportion of each variance component relating to the entire variance

***$p < 0.001$

Bosker 2002, find 5 to 20% to be typical variance proportions on the higher levels), they imply that 67.1% of the entire variance is to be attributed to different perceptions of teaching quality within the courses. An exclusive consideration at only the course level, or just the instructor level, would ignore this large proportion of the total variance.

In addition, it is revealing how well the three-level null model fits to the data in comparison with other models using deviance tests: The three-level null model demonstrates a significantly better fit than either of the two-level null models, which either only consider the course level ($\chi^2(df = 2) = 18.8$; $p < 0.001$) or the instructor level ($\chi^2(df = 2) = 282.9$; $p < 0.001$).

Bias Model

In order to receive valid estimations of the effects of counseling instructors on overall teaching quality controlling for distorting influences of potential bias and unfairness, the null model was expanded over the course of four steps to a bias model. These steps follow the recommendations made by Snijders and Bosker (2002, p. 86) and are described in the Appendix. The final bias model is presented in Table 4 and is defined through the equations $Y_{ijk} = \pi_{0jk} + \Sigma_p \pi_{pjk} \cdot X_{pijk} + e_{ijk,}$ on student level (X_p represents the pth predictor), $\pi_{0jk} = \beta_{00k} + \Sigma_q \beta_{0qk} \cdot X_{qjk} + r_{0jk}$, $\pi_{pjk} = \beta_{p0k} + r_{pjk}$ or $\pi_{pjk} = \beta_{p0k} + \Sigma_l \beta_{plk} \cdot X_{ljk} + r_{pjk}$ on course level (with X_q as the qth predictor) as well as $\beta_{00k} = \gamma_{000} + u_{00k}$ and $\beta_{0qk} = \gamma_{0q0}$ on instructor level (no significant predictors). This model demonstrates a significantly better fit to the data than the three-level null model ($\chi^2(df = 18) = 699.9$; $p < 0.001$). A comparison of the residual variances of the bias model with the variance proportions of the null model (see Table 3) illustrates that the distortions which are attributed to the bias variables represent a substantial source of variance among the student course assessments. This is particularly true for two bias variables on the course level (mean interest in the course topic and diversity of prior knowledge), which explained more than half of the criterion variance accumulated on this level of analysis.

Remarkably, effects of several bias and unfairness variables on the student level varied between courses, e.g. the average positive effect of perceived topic relevance (see Table 4). Moreover, some of these effects varying between courses were dependant on course characteristics, indicated by significant cross-level interactions. For example, the positive effect of perceived topic relevance was reduced in larger classes. With respect to bias and unfairness variables on the course level, it could be demonstrated that evaluations of teaching quality not only depended on course level of certain student characteristics (here: course mean of interest), but also on the heterogeneity found in the course (here:

Table 4 Multi-level regression of teaching quality on potential bias variables on the student, course and instructor levels using the entire sample

Fixed effect	Level	Coefficient	SE	df	*t*
Intercept	S				
Intercept	C				
Intercept	I	4.96	0.080	43	62.052***
CM interest	C	0.21	0.061	137	3.346***
CV Prior topic knowledge	C	−0.61	0.154	137	−3.946***
Student gender[a]	S				
Intercept	C	0.20	0.045	139	4.448***
Perceived topic relevance	S				
Intercept	C	0.25	0.024	138	10.335***
CM effort[b]	C	−0.05	0.024	138	−2.136*
Prior topic knowledge	S				
Intercept	C	−0.02	0.039	138	−0.514
Number of course participants[b]	C	−0.14	0.067	138	−2.092*
Effort	S	0.13	0.020	2536	6.553***
Interest in topic	S	0.10	0.011	2536	8.470***

Variance component[c]	Level	Variance	df	χ^2
Residual variance within courses	S	0.85		
Residual variance between courses	C	0.18	75	234.8***
Variance of student gender effect	C	0.09	112	139.0*
Variance of perceived topic relevance effect	C	0.03	111	209.5***
Variance of prior topic knowledge effect	C	0.05	111	175.0***
Residual variance between instructors	I	0.11	43	105.8***

S student level ($N_{\text{Students}} = 2.546$), *C* course level ($N_{\text{Courses}} = 140$), *I* instructor level ($N_{\text{Instructors}} = 44$), *CM* course mean (arithmetic mean of students' ratings for the individual course), *CV* course variance (standard deviation of students' ratings for the individual course)

***$p < 0.001$. *$p < 0.05$

[a] 1 = male, 0 = female

[b] Course characteristic predicted the regression coefficient of the corresponding student variable (cross-level interaction)

[c] For course and instructor level, results of the χ^2 tests indicate whether or not the accordant variance component is greater than zero

heterogeneity with respect to prior topic knowledge). Finally, it was found that bias/unfairness variables can affect evaluations simultaneously on the student and the course level. Specifically, it was found that student course evaluations are simultaneously dependent on individual prior interest in the course topic and the course mean of individual interest, which can be interpreted as an indicator of the interestingness of the topic covered by the course.

Effects of Counseling

Using the three-level bias model with the partial sample of instructors for which data at both measuring points were available (see Table 1), in the final (and substantially

important) step, the effect of counseling on university instructors was estimated. To this end, the course level regression equation in the final bias model was expanded to incorporate measurement point (MP; 0 = first measuring point; 1 = second measuring point; no centering): $\pi_{0jk} = \beta_{00k} + \Sigma_q \beta_{0qk} \cdot X_{qjk} + \beta_{03k} \cdot MP_{jk} + r_{0jk}$. Variations between instructors were permitted with $\beta_{03k} = \gamma_{030} + u_{03k}$ and $U_{03} = Var(u_{03k})$. With this model specification, the regression coefficient γ_{000} estimates the mean teaching quality found at the first measurement point and the regression coefficient γ_{030} estimates the mean change in teaching quality between the first and second measurement points, whereby both parameters control for the influences of bias and unfairness variables on all three levels. Taking into account the related instructor-specific residuals u_{00k} and u_{03k}, for the kth instructor teaching quality at the first measurement point and its change to the second measurement point (both adjusted for distortions) are estimated by $\beta_{00k} = \gamma_{000} + u_{00k}$ and $\beta_{03k} = \gamma_{030} + u_{03k}$, respectively.

The resulting model proves to have a better fit to the observed data than the bias model estimated with the partial sample ($\chi^2(df = 3) > 13.8$; $p < 0.01$). Means and standard deviations for teaching quality estimated at both measurement points as well as individual improvements for the instructors are contained in Table 5. A statistically significant effect of measurement point could be proven, which indicates that at the second measurement point the instructors received, on average, significantly better student evaluations than at the first measurement point ($\gamma_{030} = 0.30$; $SE = 0.136$; $t(df = 11) = 2.228$; $p < 0.05$). This effect reflects the "true improvement" which is visible when distortions are held in check. These distortions result from the simple fact that the instructors conducted different courses with different students at the two measurement points. In addition to the mean improvement in teaching quality attributed to the counseling approach, significant variations between instructors could be isolated ($U_{03} = Var(u_{03k}) = 0.11$; $\chi^2(df = 11) = 20.3$; $p < 0.05$), this suggests that the individual faculty members experienced differential rates of improvement.

In using the final model, the effect size of counseling on teaching quality was estimated. Since it has not yet been clarified how effect sizes are to be determined in multi-level models with a dependent sample on the top level and independent samples on subsequent levels, three alternatives for calculating effect sizes were used (see Table 5): (1) In the first variant, the traditional standardized mean differences for two independent samples were

Table 5 Improvements in teaching quality among the instructors receiving counseling, estimated with the three-level model under consideration of bias and unfairness variables

Variables estimated with the three-level model	M	s
Teaching success at MP 1	4.81	0.43
Teaching success at MP 2	5.11	0.17
Individual improvement in teaching	0.30	0.33
Magnitude of effect changes	**Formula for calculation**	**d or d'**
Variant 1: d for independent samples	$d = \frac{\bar{x}_2 - \bar{x}_1}{s_1}$	0.68
Variant 2: d for dependent samples	$d = \frac{\bar{x}_2 - \bar{x}_1}{s_D}$	0.91
Variant 3: d for dependent samples	$d = \frac{\bar{x}_2 - \bar{x}_1}{s_1 \cdot \sqrt{1-r}}$	0.85

$N_{\text{Instructors}} = 12$. s_D standard deviation of estimated individual improvement, r correlation of unadjusted teaching quality at the two measurement points on instructor level ($r = 0.37$), *MP* measurement point

used. The application of this variant on repeated measurement data probably results in an underestimation of the counseling effect, because the positive correlation of teaching quality between the two measurement points will not be taken into consideration (see Rindermann and Kohler 2003). (2) In the second variant, effect sizes for dependent samples were calculated on the basis of the variance of the individual differences between the two measurement points. (3) Finally, in the third variant the standardized mean differences for the dependent samples were calculated, whereby the correlation between the two measurement points is directly incorporated into the calculations. The results of all three calculation variants indicate, with $d \geq 0.68$ a moderate to large effect of counseling on teaching quality (see Table 5). The two variants for dependent samples delivered, in accordance with expectations, larger values for the counseling effect than the formula for independent samples.

By using the residuals of the final three-level model it is possible to estimate the error which is associated with the application of aggregated values on the instructor level and filter out the distorting influences of bias and unfairness variables. Therefore, the instructor-specific residuals u_{03k} were correlated with the differences in aggregated teaching quality on the instructor level (measurement point 2 minus measurement point 1). The result of $r = 0.79$ (attenuation-corrected) indicates a substantial error rate for disregarding the three-level structure and bias/unfairness variables when analyzing students' evaluations of teaching quality (38% error variance).

Quite revealing is a comparison of the two measurements of instructor-specific change in teaching quality, namely changes estimated with the final three-level model and changes in terms of simple differences of aggregated values (Fig. 1). On the one hand, it is obvious that for some of the instructors the utilization of simple difference values resulted in a sizable overestimation of their improvement in assessed teaching quality. This is illustrated by the statistics calculated for instructor 19 who, at the second measurement point, was conducting a number of courses attended by students demonstrating high prior interest in the topics covered in his/her courses, these students could also be characterized by a low degree of diversity regarding their prior knowledge. On the other hand, the sample also

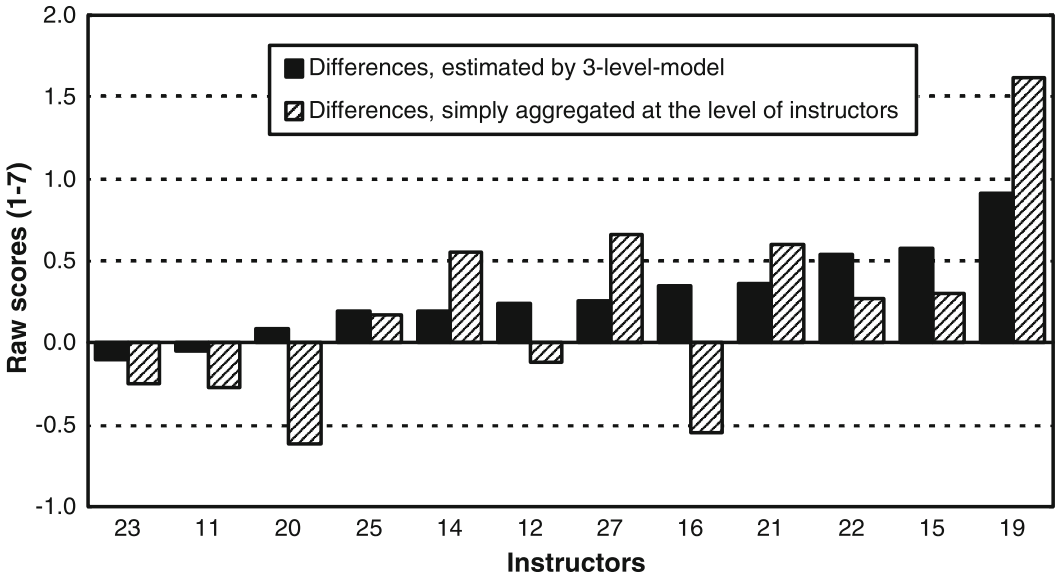

Fig. 1 Changes in assessed teaching quality between measurement points 1 and 2: comparison of values estimated through the three-level model (mean change γ_{030} plus instructor-specific residual u_{03k}), with differences in aggregated teaching quality values on the instructor level

contained instructors whose improvements in teaching quality were widely underestimated due to the application of simple difference values. For example, instructor 16 was giving courses at the second measuring point on topics considered to be of lower relevance than those he/she was giving during the evaluations at the first measurement point. Although the consideration of the three-level structure and the bias/unfairness variables made visible a significant improvement of his/her teacher quality, the application of simple difference values would lead to the conclusion that his/her teacher quality had deteriorated.

Discussion

The present study builds on prior experimental findings concerning the effectiveness of counseling instructors based on students' evaluations of teaching quality. Although these experimental findings demonstrated a high internal validity, their external validity raised concern, especially with respect to the long-term effects of counseling, to the generalizability to systems of higher education other than those found in North America, Australia and the UK, and to nested data evaluation. The present study examines how the implementation of a counseling approach can improve teacher quality for full-time instructors at a German university. The focus was on the long-term effects of counseling over the course of 2 years and transfer effects to courses other than those on which the counseling was directed. A multi-level approach was applied with three levels (student, course and instructor) and possible bias and unfairness effects were controlled for on all three levels.

The results show a significant and substantial improvement in student evaluations of teaching quality at the second measurement point of the evaluation for those teachers who received counseling on the basis of student course evaluations at the first measurement point. Applying a conservative effect size statistic (assuming independent instructor samples), a moderate counseling effect could be found; applying a more adequate effect size formula for dependent instructor samples resulted in a large counseling effect. This moderate to large effect indicates that the counseling of university instructors on the basis of student evaluations is an effective approach, one which is capable of generating long-term improvements in teaching quality. This is also true for courses which were not directly addressed during the counseling sessions. The effect sizes are comparable to those calculated in meta-analyses of experimental studies in the area (Penny and Coe 2004; Menges and Brinko 1986).

The findings here do form ecologically valid indications that counseling instructors is efficient in the long-term, and that their effects can be effectively transferred to different instructional contexts. Therefore, it can be described as an effective intervention to enhance teaching quality in higher education. Furthermore, empirical evidence was previously lacking to confirm that the approach of counseling university instructors, which had predominantly been investigated in universities in North America, is also effective in German institutes of higher learning, which are characterized by substantially different environmental conditions. Here, the present study complements the literature by pointing out that professors who work at public universities with lifetime appointments, maintain a heavy course load, conduct courses with a large number of participants, and who can not expect to be accorded any type of gratuity for high quality teaching, can also profit from counseling on the basis of students' evaluations of teaching quality.

In the present study, due to the typical limitations inherent to regular instruction in tertiary education, a control group could not be realized: Randomly selecting university instructors to be denied feedback and counseling on course evaluations for a period of

2 years was neither ethical, nor was it approved by the faculty advisory committee at the university hosting the investigation. Therefore, our design is characterized by a lower degree of internal validity than the experimental studies which defined the starting point of our research. At the same time, however, the external validity realized here, through consideration of long-term effects and transfer effects, clearly surpasses that of these studies. In this sense the present study meaningfully complements previous experimental studies. Indeed, given the practical limitations already described, it is highly improbable in practice that one study would simultaneously realize a high level of internal as well as external validity. In this respect, the separate implementation of experimental studies on the one hand and externally valid studies on the other, may provide the only feasible opportunity to meet both requirements. Moreover, extensive findings confirm that an exclusive course evaluation does not lead to improvements in teaching quality and that advances associated with the provision of feedback are small at best (e.g. Erickson and Erickson 1979; Marsh 2007b; Marsh and Hocevar 1991; Menges and Brinko 1986; l'Hommedieu et al. 1990; Rindermann 2009). Therefore it seems reasonable to assume that little to no improvement worth mentioning would surface in a control group. Consequently, one can conclude on a relatively safe basis, that the effects of counseling identified with the present study largely represent the net effects of this intervention.

One could criticize that the present results may lack of generalizability due to the small sample size of instructors and evaluation at only one university. However, another study conducted at a private German institute of speech therapy education in which teachers have a less formal status (which is more comparable to the status of instructors in English-speaking higher education) than senior university professors at German universities has come to similar results (Rindermann et al. 2007). This study was also able to demonstrate that teaching quality may be improved substantially by the use of counseling and the sizes of the counseling effects were also moderate to high ($d > 0.60$, although not corrected for instructuors rated as very good at the first time for which no substantial improvements could be expected). Taken together, the converging results of the present study and the study of Rindermann et al. (2007) can be interpreted as sound evidence that the (up to now: merely experimentally proven) counseling approach to enhance teaching effectiveness has long-term and transfer effects in culturally different contexts to those of North American, Australian and British higher education institutions. Nevertheless, additional studies are desirable to prove the effectiveness for Anglo-American contexts as well. Although we would expect even stronger effects of the counseling approach in those contexts since we see the institutional conditions for improving teaching quality in those contexts as substantially better than in German-speaking contexts.

Another aspect of the generalizability of the present results one may raise, concerns potential dependencies on the evaluation instrument. However, the instrument used is a multidimensional inventory which is similar to the SEEQ (Marsh 1982), has good psychometric properties with respect to reliability and validity and is frequently used in German-speaking institutions of higher education. Moreover, research done with the inventory revealed similar results as research with the SEEQ in English-speaking countries (e.g. Rindermann and Schofield 2001). The use of a well-established, reliable and valid instrument additionally substantiates the generalizability of the present results.

From a methodological perspective, one aim of the present paper was to examine the applicability of a three-level model to data collected with students' evaluations of university courses. This proved to be particularly useful in the representation of the bias and unfairness variables situated on the different levels of analysis. Four findings contributed to this conclusion: First, and consistent with other literature (Spiel and Gössler 2000), it was

shown that the strength and direction of bias and unfairness effects can vary between courses. Building on the variance found concerning the course-specific significance of bias and unfairness variables were, second, observances of two cross-level interactions, which indicated that the distorting influence of bias/unfairness variables on the student level could be moderated by bias/unfairness variables on the course level. Third, it was demonstrated that bias/unfairness variables can have effects on several analytical levels simultaneously. Fourth, evidence was provided that not only course averages of student characteristics, but also the diversity of relevant characteristics among course participants can influence evaluations of teaching effectiveness—such effects have been rarely investigated (see Ting 2000). All four findings suggest that data from student course evaluations should be represented on multiple levels when being analyzed.

The three-level model also proved its utility in the verification of counseling effects. It was not only useful in ensuring the comparability of students' course evaluations despite differential course attendances, the model could also accommodate the complex data structure comprised of independent samples of courses and students, as well as dependent samples of instructors who were leading more than one course. One task facing future methodological work is to determine the appropriate method to calculate effect sizes for a design of this type. In conclusion, it can be maintained that the use of a multi-level approach in the analysis of data from student course evaluations is advisable in constellations when statements concerning aggregate levels of various units are to be made simultaneously (courses, lecturers, departments, institutes, schools), or if the results from a variety of different courses and instructors are to be compared, indicating a need to control for the influence of (potential) bias and unfairness variables on the different levels.

Acknowledgment We would like to thank all of the students and faculty at the university involved for their willingness to participate in the course evaluations.

Appendix

Estimation of the Bias Model in Four Steps

In the multi-level estimation of the distorting influences of potential bias and unfairness variables, the null model was expanded to a bias model in four steps described below. This takes into consideration bias and unfairness variables on all three levels (see Table 4) in accordance with recommendations made by Snijders and Bosker (2002, p. 86).

Step 1 (Bias Variables on the Student Level)

In the first step all six potential bias variables associated with the student level were added to the model (centered on the grand mean) and variations of the six regression coefficients between courses were allowed for (random slopes). The regression equation on the student level of the null model was therefore expanded to $Y_{ijk} = \pi_{0jk} + \Sigma_p \pi_{pjk} \cdot X_{pijk} + e_{ijk}$, whereby X_p represents the pth of the six predictors. For each of the p regression weights π_{pjk}, the equation $\pi_{pjk} = \beta_{p0k} + r_{pjk}$ with $R_p = Var(r_{pjk})$ was additionally formulated on the course level in order to permit predictors to vary between courses. The estimation revealed that the effects of some variables vary between courses. The corresponding effects were retained as random effects, all remaining effects were fixed (i.e. setting $r_{pjk} = 0$). Variables which revealed no significant relationship to the criterion and which effect did not vary between courses were removed from the model.

Step 2 (Bias Variables on the Course Level)

The second step comprised the block-wise addition of the five structural course characteristics, the four course mean scores and the two course distributions into the model as predictors of the intercept (centered on the grand mean; variations between instructors were permitted). Formally speaking, the parameters were estimated using the equation $\pi_{0jk} = \beta_{00k} + \Sigma_q \beta_{0qk} \cdot X_{qjk} + r_{0jk}$ on the course level (with X_q as the qth term of the 11 predictors) as well as the eleven equations $\beta_{0qk} = \gamma_{0q0} + u_{0qk}$ on the instructor level with $U_{0q} = \mathrm{Var}(u_{0qk})$. The influences of all course characteristics were constant across instructors; consequently, all course level effects were fixed (i.e. setting $u_{0qk} = 0$). Course-level variables which did not predict the criterion variable significantly were therefore removed from the model.

Step 3 (Cross-Level Interactions Between Bias Variables on Course and Student Level)

In the third step, a test was made to determine whether the between course variations in the regression coefficients π_{pjk} for three student variables (gender, perceived topic relevance and prior knowledge) could be explained by characteristics of the courses themselves. To this end, the regression coefficients for the three bias variables named were treated as dependent variables (slopes as outcomes). In this case, a significant effect indicated a cross-level interaction in which the effect of a variable on the student level is moderated by a variable on the course level. Formally, these effects are represented by three regression equations $\pi_{pjk} = \beta_{p0k} + \Sigma_l \beta_{plk} \cdot X_{ljk} + r_{pjk}$ for the pth of the three student bias variables, whereby X_l depicts the lth of the course characteristics included (coefficients β_{plk} fixed on instructor level). Results confirmed two cross-level interactions.

Step 4 (Bias Variables on the Instructor Level)

In the fourth step, a final expansion incorporated the three potential bias variables at the instructor level. They were introduced in the intercept regression equation (grand mean centered): $\beta_{00k} = \gamma_{000} + \Sigma_m \gamma_{00m} \cdot X_{mk} + u_{00k}$ with X_m as the mth of the three predictors. None of them significantly biased student evaluations of teaching quality. They were therefore removed from the model.

References

Abrami, P. C., d'Apollonia, S., & Rosenfield, S. (2007). The dimensionality of student ratings of instruction: What we know and what we do not. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 385–456). Dordrecht: Springer.

Bray, J. H., & Howard, G. S. (1980). Methodological considerations in the evaluation of a teacher-training program. *Journal of Educational Psychology, 72*, 62–70.

Brinko, K. T. (1990). Instructional consultation with feedback in higher education. *Journal of Higher Education, 67*, 65–83.

Brinko, K. T., & Menges, R. J. (Eds.). (1997). *Practically speaking: A sourcebook for instructional consultants in higher education*. Stillwater, OK: New Forum.

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.

Cohen, P. A. (1991). Effectiveness of student ratings feedback and consultation for improving instruction in dental school. *Journal of Dental Education, 55*, 145–150.

Cranton, P., & Smith, R. A. (1990). Reconsidering the unit of analysis: A model of student ratings of instruction. *Journal of Educational Psychology, 82*, 207–212.

Erickson, G. R., & Erickson, B. L. (1979). Improving college teaching. An evaluation of a teaching consultation procedure. *Journal of Higher Education, 50*, 670–683.

Feldman, K. A. (2007). Identifying exemplary teachers and teaching: Evidence from student ratings. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 93–143). Dordrecht: Springer.

Foersterling, F. (1985). Attributional retraining: A review. *Psychological Bulletin, 98*, 495–512.

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

Hampton, S. E., & Reiser, R. A. (2004). Effects of a theory-based feedback and consultation process on instruction and learning in college classrooms. *Research in Higher Education, 45*, 497–527.

Knapper, C., & Piccinin, S. (Eds.). (1999). *Using consultants to improve teaching*. San Francisco: Jossey-Bass.

l'Hommedieu, R. L., Menges, R. J., & Brinko, K. T. (1990). Methodological explanations for the modest effects of feedback from student ratings. *Journal of Educational Psychology, 82*, 232–241.

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

Marsh, H. W. (2007a). 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.

Marsh, H. W. (2007b). Do university teachers become more effective with experience? A multilevel growth model of students' evaluations of teaching over 13 years. *Journal of Educational Psychology, 99*, 775–790.

Marsh, H. W., & Hocevar, D. (1991). Students' evaluations of teaching effectiveness: The stability of mean ratings of the same teacher over a 13-year period. *Teaching and Teacher Education, 7*, 303–341.

Marsh, H. W., & Roche, L. A. (1993). Effects of grading leniency and low workload on students' evaluations of teaching. *Journal of Educational Psychology, 92*, 202–228.

Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective. *American Psychologist, 52*, 1187–1197.

McKeachie, W. J. (2007). Good teaching makes a difference—and we know what it is. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 457–474). Dordrecht: Springer.

McKeachie, W. J., Lin, Y., Daugherty, M., Moffett, M., Neigler, C., Nork, J., et al. (1980). Using student ratings and consultation to improve instruction. *British Journal of Educational Psychology, 50*, 168–174.

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

Murray, H. G. (1997). Effective teaching behaviors in the college classroom. In R. P. Perry & J. C. Smart (Eds.), *Effective teaching in higher education: Research and practice* (pp. 171–204). New York: Agathon Press.

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

Penny, A. R., & Coe, R. (2004). Effectiveness of consultation on student ratings feedback: A meta-analysis. *Review of Educational Research, 74*, 215–253.

Perry, R. P., & Smart, J. C. (Eds.). (2007). *The scholarship of teaching and learning in higher education: An evidence-based perspective*. Dordrecht: Springer.

Piccinin, S., Cristi, C., & McCoy, M. (1999). The impact of individual consultation on student ratings of teaching. *International Journal for Academic Development, 4*, 75–88.

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., & Congdon, R. T. (2001). *HLM 5 (Version 5.04)*. Chicago: Scientific Software International.

Richardson, J. T. E. (2005). Instruments for obtaining student feedback: A review of literature. *Assessment & Evaluation in Higher Education, 30*, 387–415.

Rindermann, H. (1996). Zur Qualität studentischer Lehrveranstaltungsevaluationen [On the quality of student evaluations of college courses]. *German Journal of Educational Psychology, 10*, 129–145.

Rindermann, H. (2009). *Lehrevaluation – Einführung und Überblick zur Forschung und Praxis der Lehrveranstaltungsevaluation an Hochschulen* [Evaluation of teaching—an introduction and overview of research and practice] (2nd ed.). Landau, Germany: VEP.

Rindermann, H., & Heller, K. A. (2005). The benefit of gifted classes and talent schools for developing students' competences and enhancing academic self-concept. *German Journal of Educational Psychology, 19*, 133–136.

Rindermann, H., & Kohler, J. (2003). Lässt sich die Lehrqualität durch Evaluation und Beratung verbessern? Überprüfung eines Evaluations-Beratungs-Modells [Does evaluation and consulting improve quality of instruction? Test of an evaluation-consulting-model]. *Psychologie in Erziehung und Unterricht, 50*, 71–85.

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.

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

Rotem, A. (1978). The effects of feedback from students to university instructors: An experimental study. *Research in Higher Education, 9*, 303–318.

Snijders, T. A. B., & Bosker, R. J. (2002). *Multilevel analysis. An introduction to basic and advanced multilevel modeling*. London: Sage.

Spiel, C., & Gössler, P. M. (2000). Zum Einfluss von Biasvariablen auf die Bewertung universitärer Lehre durch Studierende [On the influence of bias variables on assessments of university teaching by students]. *German Journal of Educational Psychology, 14*, 38–47.

Ting, K.-F. (2000). Cross-level effects of class characteristics on students' perceptions of teaching quality. *Journal of Educational Psychology, 92*, 818–825.

Webler, W.-D. (1996). Qualitätssicherung in Lehre und Studium an deutschen Hochschulen [Quality Assurance in Teaching and Academic Studies at German Universities]. *Zeitschrift für Sozialisationsforschung und Erziehungssoziologie, 16*, 119–148.

Wendorf, C. A., & Alexander, S. (2005). The influence of individual- and class-level fairness-related perceptions on student satisfaction. *Contemporary Educational Psychology, 30*, 190–206.

Will, H., & Blickhan, C. (1987). Evaluation als Intervention [Evaluation as an intervention]. In H. Will, A. Winteler, & A. Krapp (Eds.), *Evaluation in der beruflichen Aus- und Weiterbildung* [Evaluation in Vocational Training] (pp. 43–59). Heidelberg: Sauer.

Wilson, R. C. (1986). Improving faculty teaching. Effective use of student evaluations and consultants. *Journal of Higher Education, 57*, 196–211.