Instructional Science (2006)
DOI 10.1007/s11251-006-9006-1

Regular Feedback from Student Ratings of Instruction: Do College Teachers Improve their Ratings in the Long Run?

JONAS W. B. LANG* & MARTIN KERSTING
*Institute of Psychology, RWTH Aachen University, Jägerstr. 17-19, 52056 Aachen, Germany (*Author for correspondence: e-mail: Jonas.Lang@rwth-aachen.de)*

Received: 29 November 2005; accepted: 16 August 2006

Abstract. The authors examined whether feedback from student ratings of instruction not augmented with consultation helps college teachers to improve their student ratings on a long-term basis. The study reported was conducted in an institution where no previous teaching-effectiveness evaluations had taken place. At the end of each of four consecutive semesters, student ratings were assessed and teachers were provided with feedback. Data from 3122 questionnaires evaluating 12 teachers were analyzed using polynomial and piecewise random coefficient models. Results revealed that student ratings increased from the no-feedback baseline semester to the second semester and then gradually decreased from the second to the fourth semester, although feedback was provided after each semester. The findings suggest that student ratings not augmented with consultation are far less effective than typically assumed when considered from a long-term perspective.

Keywords: feedback, long-term effects, student ratings, teaching effectiveness

The assessment of student ratings of instruction is common practice in higher education around the globe. Usually, it is assumed that student ratings serve three main purposes (Cohen, 1980; Howell & Symbaluk, 2001). The first is to aid administrative evaluations by measuring teaching effectiveness, which is an important criterion for decisions on matters such as pay increases, promotion, and tenure of college faculty (Carter, 1989). The second purpose of student ratings is to help students select courses and instructors. Most students consider student ratings a valuable resource for such decisions, though faculty members often raise concerns about the publication of these evaluations (Coleman & McKeachie, 1981; Howell & Symbaluk, 2001; Wilhelm, 2004).

The focus of the present article lies on the third purpose of student ratings of instruction, which is to help teachers improve their teaching by providing them with feedback. Several authors have pointed out

that both short- and long-term effects should be taken into account when considering the impact of student ratings on the improvement of teaching (Armstrong, 1998; Greenwald & Gillmore, 1998; Stevens & Aleamoni, 1985). Long-term effects are particularly important because the majority of teachers remain in higher education institutions for several years and receive regular feedback in the form of student ratings. Nevertheless, previous research has focused almost exclusively on the short-term effects of feedback from student ratings; empirical research on the long-term effects is very limited. This prompted Greenwald and Gillmore (1998) to call for new research on the long-term effects of feedback from student ratings. In response, we conducted a study to systematically investigate the effects of regular student feedback on the effectiveness of college teaching over a period of several semesters.

Short-Term Effects of Feedback from Students' Ratings

Since the 1960s, there has been extensive research into the short-term effects of feedback from student ratings on teaching effectiveness. Two meta-analyses have reviewed the available literature (Cohen, 1980; L'Hommedieu et al., 1990). The authors of both meta-analyses found moderate effects of students' feedback on instructors' teaching effectiveness (Cohen, 1980: $d = 0.38$; L'Hommedieu et al., 1990: $d = 0.34$). These findings are in line with meta-analytic outcomes from other areas of psychology. For example, Kluger and DeNisi (1996) examined the effects of various feedback interventions from all areas of psychology on performance. Guzzo et al. (1985) analyzed the effects of various psychologically based intervention programs on worker productivity. Congruent with the findings of Cohen (1980) and L'Hommedieu et al. (1990), Kluger and DeNisi (1996) found an effect of $d = 0.41$, and Guzzo et al. (1985) an effect of $d = 0.35$, for feedback interventions.

In the studies covered in the meta-analyses by Cohen (1980) and L'Hommedieu et al. (1990), the typical research design used to determine the effectiveness of student ratings for instructional improvement was a two-group, pretest/posttest study. In most cases, a comparison was made between one group of teachers, who received feedback from the pretest, and another group of teachers, who did not receive any feedback. Typically, the studies covered a time frame of just one semester, with feedback from mid-term evaluations being provided to the experimental group only. None of the studies

analyzed ratings given later than the subsequent semester. Both meta-analyses found that the effects of feedback were stronger when feedback was supplemented by some kind of consultation intervention (Cohen, 1980: $d = 0.20$ for feedback only, $d = 0.64$ for feedback with consultation; L'Hommedieu et al., 1990: $d = 0.18$ for feedback only, $d = 0.55$ for feedback with consultation).

Long-term effects of feedback from students' ratings

The only available study to have examined the long-term effects of feedback from student ratings of instruction was conducted by Stevens and Aleamoni (1985), who compared a no-feedback baseline with longitudinal data obtained over several semesters. The study analyzed ratings at four points/intervals over the course of 10 years (no-feedback baseline rating, after which the first set of feedback was provided, and 0.5 years, 4–7.5 years and 7.5–10 years after the first feedback). Although Stevens and Aleamoni found some differences between teachers who received feedback and teachers who received feedback and additional interventions (interpretation of results, or interpretation of results plus consultation targeting improvement), there was no significant effect of time on student ratings. Considering the results from the two meta-analyses mentioned above, the finding of no significant improvement between the first and second sets of feedback is rather atypical.

In sum, the literature on the effects of feedback from student ratings suggests that there is an initial increase in student ratings after teachers have received this kind of feedback for the first time. However, there is little evidence on how student ratings develop after this initial increase. From a theoretical perspective, there are three feasible possibilities (other patterns of change are, of course, conceivable, but would be difficult to interpret). First, the overall level of student ratings stagnates after the second set of feedback and remains stable although feedback is provided on a regular basis. Second, ratings further improve after the second set of feedback and reach a stable level only after several additional sets of student feedback. Third, there is a decline in student ratings after the second set of feedback, suggesting that teaching does not improve in the long run in response to feedback from student ratings. The first two patterns would be consistent with classic learning curves of different magnitudes, indicating that feedback from student ratings primarily enables college teachers to abort ineffective teaching practices. The third pattern is in line with

theoretical proposals by Wood and Locke (1990) as well as Kluger and DeNisi (1996), who argued that the first response to feedback that is not entirely positive is to work harder. This means expending effort, persisting, focusing attention on the task, and activating task-specific programs or scripts for action that are available from past experience with the task or an analogous task. When working harder does not lead to satisfactory results, people may try to work smarter. One aspect of working smarter is to search for a new task-specific strategy. If this new strategy fails, people may try to develop yet another new task-specific strategy. Experimenting with the task in this way may lead to detrimental feedback effects, as the new task-specific plans may be less appropriate than the previous strategy (Wood & Locke, 1990; Kluger & DeNisi, 1996). Detrimental effects are especially likely when a task is well-learned, as individuals typically use well-evaluated and highly automated strategies to perform these kind of tasks (Kluger & DeNisi, 1996).

In the present study, we investigated the effects of student-rating feedback over several semesters to determine the pattern of change associated with this intervention. Although the literature suggests that feedback supplemented by consultation is more effective than feedback alone, interventions of this kind are rarely used in higher education institutions because they require trained consultation personnel. Therefore, we focused on a typical feedback-only intervention, where written feedback was provided to teachers and ratings were published. In line with all previous studies, we used student ratings of instruction not only as a source of feedback, but also as a measure of teaching effectiveness (e.g., Cohen, 1980; Stevens & Aleamoni, 1985; L'Hommedieu et al., 1990; Marsh & Roche, 1993).

Method

We took the opportunity to investigate the effects of student-rating feedback on instruction when the psychology department of a large German university started to collect student feedback for the first time. No previous student evaluations had taken place in the department and all faculty members either had been in the department for at least 10 years or had no previous college teaching experience. Thus, it was reasonable to assume that the ratings of the first semester captured a no-feedback baseline. These baseline ratings were then compared with students' evaluations in subsequent semesters.

Sample

Data were collected over four semesters (i.e., 2 years). Due to fluctuation in the department's teaching staff, complete sets of useable data are available for 12 faculty members. A total of 3122 student rating questionnaires were completed for these 12 faculty members over the 4-semester period. Four faculty members were full professors (German C3 and C4 positions), one was an associate professor (German C2 position), and six were assistant professors (German C1 and BAT II positions). Ten faculty members were male and 2 were female. Their mean age at the beginning of the study was 42.67 years (SD = 11.75) and they had an average of 13.83 years' teaching experience (SD = 10.11). Of the questionnaires, 1549 were completed by female students and 1543 by male students.

Questionnaire

A questionnaire developed by Diehl (VBVOR; Diehl, 2002, 2003; Diehl & Kohr, 1977) was used to tap student ratings of instruction. The measurement instrument is well evaluated by the teaching staff of German universities and contains 18 items.

The questionnaire consists of both global and specific ratings of teachers' effectiveness (items are listed in the Appendix). Two global ratings cover the instructor's performance and the quality of the class as a whole. Both items are rated on a 6-point scale (excellent, good, satisfactory, sufficient, unsatisfactory, poor). The advantage of this 6-point scale is that it is equivalent to the grading scale used in German schools and colleges, and is thus very familiar to students. In addition to the 2 global items, the questionnaire contains 16 specific items. These items are rated on 4-point Likert scales ranging from 3 (strongly agree) to 0 (strongly disagree). The 16 items comprise the subscales rapport, teaching skill, difficulty, and content (4 items each). The first three scales cover the instructor's performance. The content scale assesses whether students regard the class in question as an effective and well-integrated part of the overall curriculum.

To evaluate the general teaching effectiveness of the teachers in the data analyses, we formed an overall index of student ratings. Although there have been some controversies in the literature as to whether student ratings are multidimensional (Marsh, 1991, 1994; Marsh & Hocevar, 1991; Marsh & Roche, 1997, 2000) or dominated by one general factor (Abrami & d'Apollonia, 1991; Cashin &

Downey, 1992; d'Apollonia & Abrami, 1997; Greenwald & Gillmore, 1997a; McKeachie, 1997), even proponents of multidimensionality agree that it is appropriate to form an overall index of student ratings to judge teaching effectiveness (Marsh, 1991, 1994; Marsh & Hocevar, 1991; Marsh & Roche, 1997). The overall index in the present study consisted of the 2 global items and the 12 specific items covering the instructor's performance (Cronbach's $\alpha = 0.87$). As the curriculum of the institution was not directly drawn up by the instructors under investigation, we did not include the content items in our analysis. Because different scales were used for the global and specific items, we first used *z*-transformations to convert the global items to the same scale as the specific items. In the present study, the formation of an overall index was also statistically appropriate. An exploratory factor analysis using maximum likelihood estimation revealed one dominating factor with an eigenvalue of 5.29 and a steep decline between the first and the second factor (eigenvalue = 1.78), suggesting a one-factor solution according to the scree criterion. The dominant factor explained 33% of the overall variance, in line with typical results of exploratory factor analyses on student ratings of instruction (see d'Apollonia & Abrami, 1997, for a review). Furthermore, confirmatory factor analysis revealed that the items used to form the overall index provided a good fit to a hierarchical model with one higher-order and four lower-order factors ($\chi^2[73, N = 3122] = 1615.92$, $p < 0.001$, Comparative Fit Index = 0.91, Standardized Root-Mean-Square Residual = 0.07).

Procedure

Faculty members asked their students to fill out the evaluation questionnaires at the end of each of four consecutive semesters. The students completed the questionnaires before any exams were taken, and all efforts were made to ensure the anonymity of the students. Shortly after the questionnaires were administered, feedback was sent to instructors via email. The results of the evaluation were also made available in the university library, allowing students to gather information about their instructors' performance. Some authors have argued that specific feedback may be superior to global feedback in helping instructors to improve their teaching (Marsh & Roche, 1993; McKeachie, 1997). Therefore, the feedback provided to instructors and students contained descriptive statistics (means and standard deviations) for both specific and global scales and items.

Data aggregation and analysis

In previous research on student ratings, either aggregated ratings at the teacher level (e.g., Stevens & Aleamoni, 1985; Marsh & Roche, 1993) or ratings at the individual student level have been used to evaluate change in student ratings of instruction. In recent years, researchers have recognized the value of incorporating a multilevel perspective into educational and organizational research (e.g., Rousseau, 1985; Klein et al. 1994; Bliese, 2000; Bliese & Jex, 2002). Taking a levels perspective means considering and correctly specifying at which hierarchical level a hypothesized process occurs. It is particularly important to determine the appropriate level of analysis, as an aggregated variable at a higher-order level might well measure a different construct than does its namesake at the individual level (Firebaugh, 1978). Because teaching effectiveness is assumed to be a variable that is a function of the performance of teachers rather than students, we decided to conceptualize teaching effectiveness measured in terms of student ratings as a teacher-level construct by aggregating the student ratings at the teacher level. This decision is also in line with the recommendations of L'Hommedieu et al. (1990). To evaluate whether aggregation was statistically appropriate, we assessed within-group agreement (James et al., 1984; Bliese, 2000; Bliese & Jex, 2002) and intra-class correlations (ICCs; Bliese, 2000; Bliese & Jex, 2002). ICC values for the overall index were ICC1 = 0.16 and ICC2 = 0.93, suggesting that teachers' performances could be reliably differentiated using mean levels of student ratings. The average $r_{wg(j)}$ value for the overall index was $r_{wg(j)} = 0.95$ using a rectangular distribution. This value indicates good within-group agreement and further justifies aggregation at the teacher level.

Research in all areas of psychology has long been plagued by the lack of appropriate methods for analyzing data on change in variables across time (Cronbach & Furby, 1970; Hofmann et al., 1993). ANOVA techniques are not suitable for modeling individual change over time because they assume that there are no individual differences in systematic changes over time and that the correlations between all possible pairs of criterion values measured at different times are equal (Cohen et al., 2003). In recent years, the use of multilevel random coefficient modeling (RCM; Longford, 1993; Bliese & Ployhart, 2002; Cohen et al., 2003), also known as hierarchical linear modeling (HLM; Bryk & Raudenbush, 1987; Hofmann et al., 1993; Raudenbush & Bryk, 2002), has provided solutions to the problem of

analyzing change. In the present study, we tested RCM models to examine the nature of the relationship between feedback and mean student ratings of instruction in subsequent semesters. All models were two-level models, with measurement occasions (4 semesters $\times$ 12 teachers = 48) at Level 1 (L1) nested within teachers at Level 2 (L2). Analyses were conducted using the Linear and Nonlinear Mixed Effects package (Pinheiro & Bates, 2000) included in the program R (R Development Core Team, 2005) and restricted maximum likelihood estimation (REML).

Results

Descriptive data

Figure 1 presents means for the aggregated overall ratings of teaching effectiveness at each semester. There was a marked increase in students' ratings of instruction from the first to the second semester. This increase was followed by decreases from the second to the third semester and from the third to the fourth semester. As shown in Table 1, the effect size of the difference in the aggregated ratings from the first to the second semester (Cohen's $d = 0.65$) was stronger than the previously reported meta-analytic findings (Cohen, 1980; L'Hommedieu et al., 1990). However, the effect size of the difference

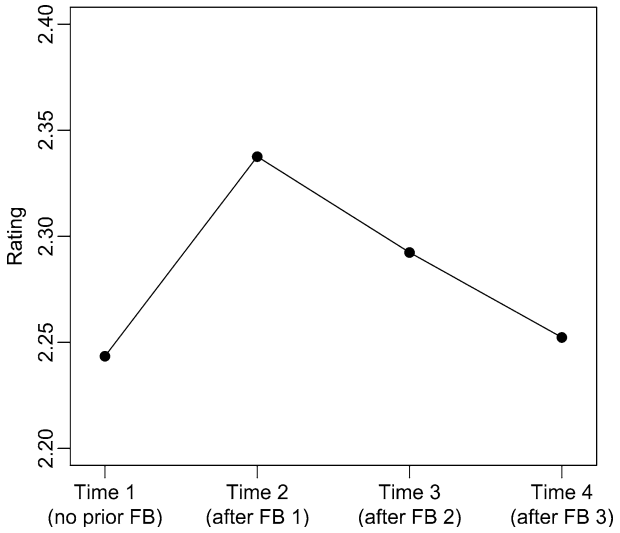

Figure 1. Means for aggregated ratings of teaching effectiveness across four semesters (2 years). FB = Feedback.

Table 1. Descriptive statistics for aggregated and non-aggregated ratings

Variable	Time 1 (no prior FB)	Time 2 (after FB 1)	Time 3 (after FB 2)	Time 4 (after FB 3)
Aggregated ratings				
M	2.24	2.34	2.29	2.25
SD	0.13	0.15	0.14	0.13
No. of teachers	12	12	12	12
d (vs. Time 1)		0.65	0.35	0.07
Non-aggregated ratings				
M	2.25	2.30	2.28	2.27
SD	0.29	0.28	0.27	0.29
No. of ratings	818	621	767	916
d (vs. Time 1)		0.17	0.11	0.06

FB = Feedback. *d* = Cohen's *d*.

in the non-aggregated ratings ($d = 0.17$) was in line with the effect sizes reported in the meta-analyses (see Table 1). These results are not surprising because most studies in the meta-analyses used non-aggregated ratings. The differences between the effect sizes for aggregated and non-aggregated ratings were largely the result of higher standard deviations at the student level (Time 1: $SD = 0.29$; Time 2: $SD = 0.28$) than at the teacher level (Time 1: $SD = 0.13$; Time 2: $SD = 0.15$).

RCM analyses

Before we modeled change in the aggregated ratings using RCM, we determined the amount of non-independence due to the nesting of measurement occasions in teachers by calculating the ICC1. We found an $ICC1 = 0.43$, indicating that teacher properties explained 43% of the variance in the variable across time.

To evaluate the pattern of change, we first tested a series of models with polynominal trends. Because students' ratings were assessed at four points in time (coded 0 = Time 1 to 3 = Time 4), polynominal models with linear, quadratic, and cubic trends were considered. We found that a linear L1 model revealed no significant linear change (coefficient = -0.002, $SE = 0.015$, $df = 33$, $t = -0.124$, $p = 0.902$), indicating that there was no significant improvement in student ratings across the full four-semester period. Therefore, we added a quadratic term to the model. The quadratic term was

significant. Adding a cubic trend to the model revealed no significant cubic change (coefficient = 0.024, SE = 0.025, df = 33, t = 0.963, p = 0.343). Consequently, we decided to use the quadratic model as the final polynominal change model.

In multilevel models with a logical ordering of the L1 variable, it is crucial to control for autocorrelation and heteroskedasticity if accurate estimations are to be made (DeShon et al., 1998; Bliese & Ployhart, 2002). Therefore, we assessed the error structure of the model by contrasting models with and without autocorrelation and heteroskedasticity using log-likelihood ratio tests. We found evidence for both significant autocorrelation (χ^2_{diff} [1] = 10.515, p = 0.001) and heteroskedasticity (χ^2_{diff} [1] = 4.982, p = 0.026). Thus, we accounted for both error structures in the final polynominal model. Results for the final polynominal RCM model, which are shown in Table 2, indicated that the quadratic trend was highly significant in the final model.[1]

Visual inspection of the pattern of change illustrated in Figure 1 suggests a strong increase in student ratings between the first and the second semester, and a steady decline in student ratings as of the second semester. In an attempt to model this pattern of change better than with the quadratic model, we also examined a piecewise RCM (Raudenbush & Bryk, 2002; Hernández-Lloreda et al., 2004). In this model, one linear component accounted for the strong increase in student ratings between the first and the second semester and a second linear term modeled the steady decline in student ratings after the second set of feedback (i.e., from the second semester to the fourth semester). The coding of the time variables that modeled the increase (period 1) and the decline (period 2) in students' ratings is shown in Table 3. Results for the piecewise random coefficient model are presented in Table 4. Again, we found both significant autocorrelation (χ^2_{diff} [1] = 9.829, p = 0.002) and significant heteroskedasticity (χ^2_{diff} [1] = 5.793, p = 0.016). Thus, these error structures were also included in the final model. As shown in Table 4, both linear components of the piecewise model were significant in the model.

Table 2. Quadratic random coefficient model predicting change in student ratings as a function of repeated feedback

Variable	Coefficient	*SE*	*df*	*t* test	*p*
(Intercept)	2.254	0.036	34	62.839	0.000
Linear Trend	0.107	0.025	34	4.234	0.000
Quadratic Trend	−0.039	0.010	34	−3.887	0.000

Table 3. Coding of linear terms in the piecewise random coefficient model

Variable	Time 1 (no prior FB)	Time 2 (after FB 1)	Time 3 (after FB 2)	Time 4 (after FB 3)
Period 1	0	1	1	1
Period 2	0	0	1	2

FB = Feedback.

Table 4. Piecewise random coefficient model predicting change in student ratings as a function of repeated feedback

Variable	Coefficient	*SE*	*df*	*t* test	*p*
(Intercept)	2.244	0.035	34	64.025	0.000
Period 1	0.093	0.022	34	4.268	0.000
Period 2	−0.043	0.018	34	−2.444	0.020

To compare the piecewise model with the quadratic model, we reestimated both models using full maximum likelihood estimation (FML) and then examined Akaike's information criterion (AIC; Akaike, 1973) and Schwartz's information criterion (BIC; Schwarz, 1978; Raftery, 1995).[2] Both information criteria indicated that the piecewise linear model (AIC = −64.31, BIC = −41.86, $\chi^2[12] = 44.16$) provided a slightly better fit to the data than the quadratic model (AIC = −63.86, BIC = −41.40, $\chi^2[12] = 43.93$).

Discussion

The purpose of the present study was to enhance the sparse literature on the long-term effects of feedback from student ratings of instruction on the teaching effectiveness of college instructors. We conducted a study lasting four semesters in an institution where no previous evaluations of teaching effectiveness had taken place. Feedback was provided regularly at the end of each of the four semesters. We found a strong initial increase in student ratings from the first semester (no-feedback baseline) to the second semester, followed by declines in student ratings from the second to the third and the third to the fourth semester. This pattern of change was adequately described by a polynominal random coefficient model incorporating a quadratic trend and by a piecewise random coefficient model with a linear increase between the first and second semester and a linear decline as of the second semester.

Our finding of an initial increase in student ratings over the short period from the first to the second semester replicates the results of two meta-analyses on the short-term effects of student ratings (Cohen, 1980; L'Hommedieu et al., 1990). Where longer-term change in student ratings of instruction is concerned, the present study provides new insights into the way that student ratings influence the performance of college instructors. Only one previous study has analyzed feedback-related change in student ratings over more than two semesters (Stevens & Aleamoni, 1985). Unexpectedly, Stevens and Aleamoni failed to observe an initial increase in student ratings. The findings of the present study, in contrast, suggest that an initial increase in student ratings is followed by a steady decline over time.

From a theoretical perspective, the pattern of change observed is in line with theoretical proposals that regular feedback first leads people to "work harder" and then to "work smarter" (Kluger & DeNisi, 1996; Wood & Locke, 1990). With respect to the impact of student ratings, this means that instructors are motivated to improve their ratings by trying to work harder when first provided with feedback. As a result, their ratings improve considerably. However, teachers might expect much more rapid improvements. Furthermore, their colleagues also improve, meaning that the improvement of a single teacher is not as noticeable, nor is the improvement in their ratings as pronounced. Hence, teachers either lose interest in improving their teaching or experiment with totally new teaching strategies ("working smarter"), with detrimental effects on their performance. This behavior leads to a decline in student ratings.

Strengths and limitations

Several limitations and strengths of the present study are noteworthy. The study's first strength is that it covered the complete teaching staff of an institution that had not previously assessed student ratings of instruction. Therefore, we were able to capture the effect of our intervention without it being confounded with previous rating policies. This is no longer an easy task, as student ratings are now commonplace in most higher education institutions. Moreover, the results are representative for typical student rating interventions in the field. The second strength of our study was that our sample of instructors was diverse in terms of age and experience, meaning that the effects are representative for typical college faculties with respect to these variables.

The first limitation of the study is that it did not include a control group. This decision was made for several reasons. First, considering the evidence from the two meta-analyses, it seemed unethical to deny feedback to a group of teachers. Second, it is often difficult to create control groups in organizations as people in the experimental group tend to communicate about the intervention with people in the control group. Given the lack of a control group, the increase in student ratings from the first to the second semester might arguably be the result of a Hawthorne (placebo) effect and/or other typical threats to the validity of pre–post study designs (Campbell & Stanley, 1963). However, it is rather unlikely that such effects explain the huge effect size differences found between the no-feedback baseline semester and the semester after the first set of feedback, for two reasons. First, the pre–post differences for measurement-only control groups and Hawthorne (placebo) control groups reported in the meta-analytical literature are typically very small for studies conducted in educational and organizational contexts. In a meta-analytical study of effect sizes in research on training, Carlson and Schmidt (1999) reported an average effect size of $d = 0.07$ for pre–post comparisons of control groups receiving no-training. In a meta-analysis of 38 educational experiments, Adair et al. (1989) found a weighted mean effect size of $d = 0.01$ for pre–post comparisons of Hawthorne (placebo) control groups. Second, the two existing meta-analyses (Cohen, 1980; L'Hommedieu et al., 1990) both used studies incorporating control groups, and both found differences between control groups and feedback-only groups.

The second limitation of the present study is the way we operationalized teaching effectiveness. In line with previous research, we used student ratings as a measure of teaching effectiveness. Student ratings of instruction are regarded as valid and useful, albeit imperfect, indicators of teaching effectiveness (d'Apollonia & Abrami, 1997; Greenwald, 1997; Marsh & Roche, 1997, 2000; McKeachie, 1997). Specifically, student ratings of instruction have been found to be related to a variety of criteria of effective teaching as well as teacher-produced student learning (Abrami et al., 1990; d'Apollonia & Abrami, 1997; Marsh & Roche, 1997). Despite these promising findings, some authors have identified biases that reduce the validity of student ratings of instruction, most prominently grading leniency (Greenwald & Gillmore, 1997a), workload (Greenwald & Gillmore, 1997b), the instructor's gender (Basow, 1995), and class characteristics (e.g., Ting, 2000). However, some of these biases have since been

shown to be statistical artifacts or results of valid teaching effects rather than bias (Marsh & Roche, 2000). Thus, the existence and practical relevance of these biases remains a controversial topic in the literature (d'Apollonia & Abrami, 1997; Greenwald, 1997; Marsh & Roche, 1997, 2000). In future research, therefore, it may be worthwhile considering more objective criteria, such as grades or standardized graduate record examinations, as additional criteria of change.

A third limitation of our study is that the generalizability of our findings is restricted to feedback-only interventions based on student ratings. The findings cannot be applied to interventions combining feedback from student ratings with consultation or with pay-for-performance reward systems.

Conclusion

Although many researchers have suggested that feedback from student ratings should be augmented by consultation (Franklin & Theall, 2002; Marsh & Roche, 1993), it is often argued that feedback from student ratings alone is a valuable tool for improving the quality of teaching in higher education. The results of the present study suggest that the positive effects of feedback-only interventions found in two meta-analyses (Cohen, 1980; L'Hommedieu et al., 1990) are sustained only for a short period of time. Over longer periods of time, the effects of regular feedback from student ratings decrease rapidly. Thus, the results of the present investigation cast doubt on the effectiveness of feedback from student ratings as a way to improve college teaching in the long term. Nevertheless, student ratings may still be valuable to higher education organizations in that they can inform administrative decisions and help students to make course choices. Future research should investigate whether the pattern of change found in the present investigation also applies to the effects of interventions that combine feedback with consultation and/or a pay-for-performance reward system. Consultation interventions, in particular, seem to be a much more promising way of achieving long-term improvements. In our theoretical interpretation of the present findings, we argued (drawing on Kluger and DeNisi, 1996) that teachers trying to improve their ratings after repeated unsatisfactory feedback might start to experiment with new teaching methods. It is at this point that consultation interventions might be able to provide motivated teachers with the necessarily support and knowledge to effect long-term change.

Acknowledgements

We would like to thank Jessica Lang, Annette Kluge, Jan Schilling, and two anonymous reviewers for their helpful comments on an earlier version of this article, and Paul D. Bliese for answering questions on random coefficient modeling and data aggregation. Further thanks go to Susannah Goss for improving the language of this article.

Notes

1. Note that the linear term in the quadratic model indicates that ratings increased at a significant linear rate in the first semester and does not indicate that there was a significant linear trend over all four semesters. For a detailed account of how to interpret lower-order polynominal trends in random coefficient models and other growth curve models, see Biesanz et al. (2004).
2. Model comparisons using restricted maximum likelihood estimation (REML) are only meaningful when the fixed effects of the models are identical (Pinheiro & Bates, 2000). Full maximum likelihood estimation should be used to compare models with different fixed effects. As the two models had an equal number of degrees of freedom and were not nested, a χ^2-test could not be applied.

Appendix: Items used in the study

1. What grade would you give the instructor?
2. What grade would you give this class?
3. The instructor was not particularly interested in the students' progress. (R)
4. The instructor's attitude toward the students was cold and unpersonal. (R)
5. The instructor seemed to see teaching as a duty and a routine activity. (R)
6. The instructor was clearly only interested in getting through the material. (R)
7. It was easy to follow the material covered in the course.
8. Too much material was covered in the course. (R)
9. The pace was too fast. (R)
10. You had to put in a lot of extra work to keep up with the course. (R)
11. The course was often confusing because it seemed to lack structure, and it was easy to get lost. (R)
12. The instructor presented the material in a clear and understandable manner.

13. The instructor planned and delivered the course well.
14. The course was clearly structured.

Note: Original German versions of the items may be found in Diehl (2002). Items 1 and 2 are from the global subscale, items 3–6 from the rapport subscale, items 7–10 from the difficulty subscale, and items 10–14 from the teaching skill subscale. R = items scored reversely to form the overall index.

References

Abrami, P.C. & d'Apollonia, S. (1991). Multidimensional students' evaluations of teaching effectiveness? Generalizability of "N = 1" research: Comment on Marsh (1991). *Journal of Educational Psychology* 83: 411–415.

Abrami, P.C., d'Apollonia, S. & Cohen, P.A. (1990). Validity of student ratings of instruction: What we know and what we do not. *Journal of Educational Psychology* 82: 219–231.

Adair, J.G., Sharpe, D. & Huynh, C.L. (1989). Hawthorne control procedures in educational experiments: A reconsideration of their use and effectiveness. *Review of Educational Research* 59: 215–228.

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In B.N. Petrov and F. Csaki (eds), *Second international symposium on information theory*, pp. 267–281. Akademiai Kiado: Budapest, Hungary.

Armstrong, S.J. (1998). Are student ratings of instruction useful? *American Psychologist* 53: 1223–1224.

Basow, S.A. (1995). Student evaluations of college professors: When gender matters. *Journal of Educational Psychology* 87: 656–665.

Biesanz, J.C., Deeb-Sossa, N., Papadakis, A.A., Bollen, K.A. & Curran, P.J. (2004). The role of coding time in estimating and interpreting growth curve models. *Psychological Methods* 9: 30–52.

Bliese, P.D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In K.J. Klein and S.W.J. Kozlowski (eds), *Multilevel theory research, and methods in organizations: Foundations, extensions, and new directions*, pp. 349–381. Jossey-Bass: San Francisco, CA.

Bliese, P.D. & Jex, S.M. (2002). Incorporating a multilevel perspective into occupational stress research: Theoretical, methodological, and practical implications. *Journal of Occupational Health Psychology* 7: 265–276.

Bliese, P.D. & Ployhart, R.E. (2002). Growth modeling using random coefficient models: Model building, testing, and illustrations. *Organizational Research Methods* 5: 362–387.

Bryk, A.S. & Raudenbush, S.W. (1987). Applications of hierarchical linear models to assessing change. *Psychological Bulletin* 101: 147–158.

Campbell, D.T. & Stanley, J.C. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Carlson, K.D. & Schmidt, F.L. (1999). Impact of experimental design on effect size: Findings from the research literature on training. *Journal of Applied Psychology* 84: 851–862.

Carter, R.E. (1989). Comparison of criteria for academic promotion of medical-school and university-based psychologists. *Professional Psychology: Research and Practice* 20: 400–403.

Cashin, W.E. & Downey, R.G. (1992). Using global student rating items for summative evaluation. *Journal of Educational Psychology* 84: 563–572.

Cohen, J., Cohen, P., West, S.G. & Aiken, L.S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences,* 3rd ed. Mahwah, NJ: Erlbaum.

Cohen, P.A. (1980). Effectiveness of student-rating feedback for improving college instruction: A meta-analysis. *Research in Higher Education* 13: 321–341.

Coleman, J. & McKeachie, W.J. (1981). Effects of instructor/course evaluations on student course selection. *Journal of Educational Psychology* 73: 224–226.

Cronbach, L.J. & Furby, L. (1970). How we should measure "change": Or should we?. *Psychological Bulletin* 74: 68–80.

d'Apollonia, S. & Abrami, P.C. (1997). Navigating student ratings of instruction. *American Psychologist* 52: 1198–1208.

DeShon, R.P., Ployhart, R.E. & Sacco, J.M. (1998). The estimation of reliability in longitudinal models. *International Journal of Behavior and Development* 22: 493–515.

Diehl, J.M. (2002) VBVOR–VBREF. Fragebögen zur studentischen Evaluation von Hochschulveranstaltungen – Manual. [VBVOR – VBREF. Questionnaires for students' evaluations of college courses–Manual]. Retrieved on March 17, 2005 from http://www.psychol.uni-giessen.de/dl/det/diehl/2368/.

Diehl, J.M. (2003). Normierung zweier Fragebögen zur studentischen Beurteilung von Vorlesungen und Seminaren [Student evaluations of lectures and seminars: Norms for two recently developed questionnaires]. *Psychologie in Erziehung und Unterricht* 50: 27–42.

Diehl, J.M. & Kohr, H.-U. (1977). Entwicklung eines Fragebogens zur Beurteilung von Hochschulveranstaltungen im Fach Psychologie [Development of a psychology course evaluation questionnaire]. *Psychologie in Erziehung und Unterricht* 24: 61–75.

Firebaugh, G. (1978). A rule for inferring individual-level relationships from aggregate data. *American Sociological Review* 43: 557–572.

Franklin, J. & Theall, M. (2002). Faculty thinking about the design and evaluation of instruction. In N. Hativa and P. Goodyear (eds), *Teacher thinking beliefs and knowledge in higher education*. Kluwer: Dordrecht, The Netherlands.

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

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

Greenwald, A.G. & Gillmore, G.M. (1997b). No pain, no gain? The importance of measuring course workload in student ratings of instruction. *Journal of Educational Psychology* 89: 743–751.

Greenwald, A.G. & Gillmore, G.M. (1998). How useful are student ratings? Reactions to comments on the current issues section. *American Psychologist* 53: 1228–1229.

Guzzo, R.A., Jette, R.D. & Katzell, R.A. (1985). The effects of psychologically based intervention programs on worker productivity: A meta-analysis. *Personnel Psychology* 38: 275–291.

Hernández-Lloreda, M.V., Colmenares, F. & Martínez-Arias, R. (2004). Application of piecewise hierarchical linear growth modeling to the study of continuity in behavioral development of baboons (*Papio hamadryas*). *Journal of Comparative Psychology* 118: 316–324.

Hofmann, D.A., Jacobs, R. & Baratta, J.E. (1993). Dynamic criteria and the measurement of change. *Journal of Applied Psychology* 78: 194–204.

Howell, A.J. & Symbaluk, D.G. (2001). Published student ratings of instruction: Revealing and reconciling the views of students and faculty. *Journal of Educational Psychology* 93: 790–796.

James, L.R., Demaree, R.Q. & Wolf, G. (1984). Estimating withingroup interrater reliability with and without response bias. *Journal of Applied Psychology* 69: 85–98.

Klein, K.J., Dansereau, F. & Hall, R.J. (1994). Levels issues in theory development, data collection, and analysis. *Academy of Management Review* 19: 195–229.

Kluger, A.N. & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin* 119: 254–284.

L'Hommedieu, R., 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.

Longford, N. (1993). *Random coefficient models*. Oxford: Oxford University Press.

Marsh, H.W. (1991). Multidimensional students' evaluations of teaching effectiveness: A test of alternative higher-order structures. *Journal of Educational Psychology* 83: 285–296.

Marsh, H.W. (1994). Comments to: "Review of the dimensionality of student ratings of instruction: I. Introductory remarks. II. Aggregation of factor studies. III. A meta-analysis of the factor studies". *Instructional Evaluation and Faculty Development* 14: 13–19.

Marsh, H.W. & Hocevar, D. (1991). The multidimensionality of students' evaluations of teaching effectiveness: The generality of factor structures across academic discipline, instructor level, and course level. *Teaching and Teacher Education* 7: 9–18.

Marsh, H.W. & Roche, L.A. (1993). The use of student evaluations and an individually structured intervention to enhance university teaching effectiveness. *American Educational Research Journal* 30: 217–251.

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. & Roche, L.A. (2000). Effects of grading leniency and low workload on students' evaluations of teaching: Popular myth, bias, validity, or innocent bystanders? *Journal of Educational Psychology* 92: 202–228.

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

Pinheiro, J.C. & Bates, D.M. (2000). *Mixed-effects models in S and S-PLUS*. New York: Springer.

R Development Core Team (2005). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Raftery, A.E. (1995). Bayesian model selection in social research. *Sociological Methodology* 25: 111–196.

Raudenbush, S.W. & Bryk, A.S. (2002). *Hierarchical linear models: Applications and data analysis methods,* 2nd ed. Thousand Oaks, CA: Sage.

Rousseau, D. (1985). Issues of level in organizational research: Multi-level and cross-level perspectives. In L.L. Cummings and B.M. Staw (eds), *Research in organizational behavior* (Vol. 7, pp. 1–37). JAI Press: Greenwich, CT.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* 6: 461–464.

Stevens, J.J. & Aleamoni, L.M. (1985). The use of evaluative feedback for instructional improvement: A longitudinal perspective. *Instructional Science* 13: 285–304.

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

Wilhelm, W.B. (2004). The relative influence of published teaching evaluations and other instructor attributes on course choice. *Journal of Marketing Education* 26: 17–30.

Wood, R.E. & Locke, E.A. (1990). Goal setting and strategy effects on complex tasks. In L.L. Cummings and B.M. Staw (eds), *Research in organizational behavior* (Vol. 12, pp. 73–109). JAI Press: Greenwich, CT.