Improving university lectures with feedback and consultation
Knol, M.H.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Experiment II:

Effects of intermediate student feedback and collaborative consultation on university professors’ lecturing skills and students’ self-assessed learning
Abstract

Interventions to improve university teaching are implemented on a daily basis, but are seldom investigated rigorously. The aim of the present study is to present an experimental study on the effectiveness of intermediate students’ evaluations of teaching (SETs) with or without consultation (SET consultation) on professors’ lecturing skills and student learning. In total, 9616 students and 75 professors from five different university departments participated in the study. Students rated their professors’ lecturing skills and their own learning process three times during their course with the Instructional Skills Questionnaire (ISQ). The study contained a randomized controlled design with three conditions; a control condition in which professors received the student feedback at the end of the course ($N = 25$), a feedback-only condition in which professors received the student feedback each time shortly after the rated lecture ($N = 24$), and a feedback-plus-consultation condition in which professors received student feedback and collaborative consultation with a consultant after each rated lecture ($N = 26$). Multilevel regression analysis showed significant effects of intermediate feedback plus collaborative consultation on multiple dimensions of lecturing skills and on students’ perceptions on how much they learned from the lectures. Ratings increased most on teaching dimensions that were targeted for improvement during consultation. Intermediate feedback only had no effect on the dependent variables. The implications of these findings for the practice and study of SET consultation are discussed.
Introduction

As accountability of universities on the quality of teaching became more important over the years, faculty development centers have been created, starting in the 1970’s, to support and improve university teaching. Nowadays, interventions to improve university teaching, such as workshops and consultation, are implemented on a daily basis, but their effectiveness is seldom investigated rigorously. Authors of previous reviews on the effects of these interventions all stressed the importance of more experimental research in this field (Levinson-Rose & Menges, 1981; Prebble, Hargraves, Leach, Naidoo, Suddaby, & Zepke, 2004; Stess, Min-Leliveld, Gijbels, & Van Petegem, 2010; Weimer & Lenze, 1997). The aim of the present study is to present an experimental study on the specific effects of intermediate students’ evaluations of teaching (SETs) with or without consultation (SET consultation) on university professors’ lecturing skills and student learning processes.

We focus on intermediate SETs and SET consultation, because both are used frequently to improve university teaching (Knapper & Piccinin, 1999; Penny & Coe, 2004; Prebble et al., 2004). SET consultation has proven to be effective, but the variation in effects is exceedingly large (Menges & Brinko, 1986; Penny & Coe, 2004). The next step in this field of research is therefore to provide more insight into the effectiveness of particular approaches and procedures. Additionally, it is necessary to investigate possible moderating effects of characteristics, such as professors’ age and class size, to find out who benefits most from these interventions. Findings on this matter are important to the development of evidence based effective interventions, aimed at improving university teaching. It is also relevant in ultimate cost-benefit analyses of these types of interventions.

Below we first provide a theoretical framework on the effects of SETs and approaches to SET consultation. Next, we present a study in which the effects of a specific approach to intermediate SETs and SET consultation were investigated with a randomized experimental repeated measures design, with a large variety of professors from different departments. We studied the effects, in terms of changes in student ratings over time as well as changes in students’ perceptions on their learning outcomes. To isolate the effects of feedback and consultation, the design included three conditions: a control condition, a feedback-only condition, and a feedback-plus-consultation condition. We used multilevel regression analysis to take into account the clustering of the data due to random differences between the lectures, the students and the professors.
Research on SETs and SET consultation

Nowadays, students’ evaluations of teaching are collected at the end of a course or term at almost every university. Their main purpose is to provide professors with feedback so they can improve the quality of their teaching. SETs have proven to be reliable and valid under many circumstances (for an overview, see Marsh, 2007b). Although, feedback in general is a powerful learning tool (Hattie & Timperley, 2007), SETs collected at the end of the course or term have little effect on teaching behavior (Kember, Leung & Kwan, 2002, see also Marsh, 2007a; Marsh & Hocevar, 1991b). Considering a period of over 13 years, Marsh and Hocevar (1991b) and Marsh (2007a) showed no improvement in the teaching effectiveness of one hundred ninety-five faculty members according to their student ratings.

The lack of impact of SETs may be explained by the quality and use of the feedback. By quality we mean that feedback should be well timed, specific, and concern changeable behavior (McLaughlin & Pfeifer, 1988). SETs are arguably ill timed and often contain mainly general items (e.g., “Overall, I rate this instructor an excellent teacher”), which mostly serve as a general monitor. In terms of use, Theall and Franklin (2001) found that SETs are often misinterpreted, misused, or simply discarded.

Previous studies have shown positive, but small, effects of providing professors with additional mid-term evaluations, compared to end-of-the term feedback alone (Cohen, 1980; Menges & Binko, 1986). The effects were investigated in terms of an increase in student ratings over time. According to three meta-analyses, these effects of mid-term evaluation were often appreciably greater with additional consultation (Cohen, 1980; Menges & Binko, 1986; Penny & Coe, 2004). Thus, student feedback may still help in the professional development of individual professors, particularly if it is supported by an appropriate process of consultation (Richardson, 2005).

Although SET consultation has proven to be more effective, reviewers have noted that the variation in observed effects is large (Menges & Brinko, 1986; Penny & Coe, 2004). In their meta-analyses on 11 experimental studies, Penny and Coe (2004) found a weighted mean effect size in terms of Cohen’s $d$ of .69 with a 95% confidence interval of .43 to .95. Menges and Brinko (1986) found effect sizes ranging from 0 to 2.5, with an average of 1.1, which again suggests considerable variation in effectiveness.

Penny and Coe (2004) studied the predictors of successful SET consultation. They investigated differences between studies in terms of features, participant characteristics, form of consultation and consultation components. They identified several consultation strategies that appeared to be important (e.g., active involvement of the professor in the learning process, sufficient time for dialogue, and use of high-quality feedback information). Ultimately, they
noted that none of the differences proved to be significant, possibly due to low statistical power given the small number of studies available. They concluded: “Thus the most robust finding may be that more [experimental] research is needed” (p. 236).

With the present experimental study, we investigated the effects of intermediate SETs with and without a collaborative approach to consultation. The rational for this consultation approach is discussed in the following section.

**Approaches to consultation**

Penny and Coe (2004) distinguished diagnostic, advisory, and educational approaches to consultation. The diagnostic approach includes an interpretation of, and advice on, the SETs results by the consultant. The last two are more intensive interventions, as they include at least one additional source of information on teaching behavior (e.g., observation or videotaping), and/or additional educational activities (e.g., seminars and workshops). Penny and Coe’s review contained one study with one small effect of .18 and one study with a medium effect of .46 of the diagnostic approach. The review contained studies with larger effect sizes ranging from .01 to 1.14 of the other two approaches.

The prescriptive model and the collaborative model are the two most common consultation models (Brinko, 1990). In the prescriptive model, the consultant identifies, diagnoses, and solves problems. Penny and Coe’s diagnostic model counts as a prescriptive approach, since it was “identified as the consultation process that simply involved interpretation of ratings, with some discussion and recommendations for improvement” (p. 230). In the collaborative model, the consultant plays a more facilitating role, by encouraging the professor to reflect on their teaching effectiveness, the current situation, and possible alternative teaching strategies. Most experimental studies in Penny and Coe’s meta-analyses were conducted in the 70’s and 80’s. In the past two decades, the collaborative approach to consultation has been adopted more often, usually in non-experimental studies. Overall the results of this approach have been positive (e.g., Dresel & Rindermann, 2011; Piccinin, Cristi & McCoy, 1999; Rindermann, Kohler & Meisenberg, 2007).

These recent findings are important, because they address biasing variables, the importance of using appropriate multilevel procedures and effects in a non-English speaking country for the first time. Whether non-experimental results are due to the intervention or due to alternative explanations remains an open question. For example, in some studies the participating professors approached a teacher training center and were highly motivated to change. Additionally, the effects of a specific approach may be due to a Hawthorne effect (the
attention/social treatment one receives). Dresel and Rindermann (2011) noted the difficulty of conducting research, which is both internally and externally valid. With the present study we aim to augment recent non-experimental findings with up-to-date experimental results for a collaborative approach to consultation.

**The present study**

Students in this study evaluated three lectures (class meetings in which lecturing was the teaching format) during the course. Each time they completed a questionnaire concerning the students’ perception on seven specific dimensions of the professor’s lecturing skills. In addition, the questionnaire contained questions concerning their self-assessed learning outcomes. This procedure was chosen to improve the timing, specificity, and comparability of the student ratings feedback. The student ratings were provided to professors in between the rated lectures, with consultation (experimental condition 2) or without consultation (experimental condition 1). The control condition received all ratings at the end of the course. This allowed us to separate the effects of intermediate feedback from consultation. We further investigated differences between dimensions that were targeted and not targeted during the consultation to investigate whether increases in ratings were either due to a Hawthorne effect, or to the specific approach to consultation.

In addition to the effects in terms of an increase in students’ evaluations of teaching, we added additional questions to the SET instrument to measure changes in students’ perception of their learning outcomes. Based on the literature, Vermunt and Verschaffel (2000) distinguished three domains of the student learning process categorized into cognitive/processing, affective, and regulation functions. When rating the lectures, students were also asked to rate their improvement due to the specific lecture on these three domains. This way, the effects of the interventions were determined in terms of improvement in teaching behavior and improvement of the student learning process, both as perceived by students. Finally, we investigated the moderating effects of professors’ age, professors’ prior teaching quality and class size on both the professor level and the student level dependent variables.

We used multilevel modeling to analyze the data. In the past, this statistical approach was poorly disseminated in terms of user friendly software. At present, many programs are available, in the form of dedicated software (MLwiN; Rasbash, Charlton, Browne, Healy & Cameron, 2009) and procedures (SPSS linear mixed; 2007) and libraries (R nlme; Pinheiro, Bates, DebRoy, Sarkar & the R Development Core Team, 2012). Multilevel regression analysis
allowed us to take the nested structure of the data into account (students were nested within classes and measurement occasions were nested within students). By doing so, we were able to account for the clustering due to random differences between the lectures, the students and the professors.

In summary the present study addresses the following research questions:

1. What are the effects of intermediate SETs with and without collaborative SET consultation, on professors’ lecturing skills, measured by students’ evaluations of teaching, and on students’ perception of their learning outcomes?
2. If effects occur, is there a difference in effect between teaching dimensions which are targeted for improvement during consultation and teaching dimensions which are not targeted?
3. Are the effects on the professor and student level dependent variables moderated by the professors’ age, prior quality of teaching or class size?

**Context information**

The study was conducted at a Dutch university at bachelor level. In the Netherlands, bachelor programs are focused on a specific field of study from day one (no general college courses are taught) and generally take three years. At this university, each course takes eight weeks (varying in workload), and students attend several courses per semester. The selected courses in this study included at least one weekly lecture, with additional meetings during the week in small groups to discuss course assignments with a tutor. The standard lecture time at this university is 90 minutes with a 15-minute break. Final exams take place in the eighth week.

Regular SETs at this university are anonymous and conducted at the end of each course (most often during the final exam). The results are sent to the professor, coordinator, management, and quality control committee.
Method

Participants

Professors
In total, 95 university professors from five departments of one Dutch university met the inclusion criteria. These criteria were: 1) professors were scheduled to teach a minimum of 3 lectures (class meetings in which lecturing was the teaching format) during a course in 2009-2010; 2) the number of enrolled students in the course was at least 25 students; and 3) professors did not follow any other professional development program while participating in this study. From this group, 87 professors agreed to participate. The main reason for not participating was a lack of time. During the study, 12 professors dropped out due to reasons not related to one of the conditions (e.g., illness, rescheduling). This resulted in a final sample of 75 professors (63 male, 12 female, age $M = 46.8$, age $SD = 9.6$) from the departments of Law ($N = 20$), Economics ($N = 24$), Science ($N = 13$), Social and Behavioral Sciences ($N = 13$), and Humanities ($N = 5$). Out of the 225 lectures (three lectures per professor) that were scheduled to be rated by the students, seven lectures were not rated by mistake. This resulted in 73 rated lectures on $T_1$, 74 rated lectures on $T_2$, and 71 rated lectures on $T_3$.

Students
The students rated their professors by completing the Instructional Skills Questionnaire (ISQ, see below) three times during the course. In total the ISQ was completed 14,298 times: 5,353 times in the control condition, 4,602 times in the feedback-only condition, and 4,343 times in the feedback–plus-consultation condition. Student-ID numbers were missing on 1,927 ISQ forms (13.5% of all completed forms). A small number of students attended more than one selected course, and therefore rated different professors in this study. Since students did not know that the professors were participating in an experiment with different conditions, this was not expected to be of any influence. Forms with missing student ID numbers were given a unique student ID number, which resulted in a total of 9,616 unique professor-student combinations.

A mean response rate of 90.2% was observed in 76 randomly selected lectures. The median class sizes, in terms of ISQ forms completed, over the three measurement occasions were 48.5 students ($min = 10$, $max = 365$, $M = 72.3$, $SD = 64.8$), 46 students ($min = 11$, $max = 215$, $M = 66.7$, $SD = 53.1$), and 43 students ($min = 8$, $max = 190$, $M = 57.9$, $SD = 38.2$) in the control, the feedback-only, and feedback–plus-consultation condition, respectively.
There was an expected decrease in attendance over time in all three conditions, as a number of students invariably drop out of courses (N_{ISQ, forms} on T$_1$: 5,900, T$_2$: 4,649, T$_3$: 3,749). In the control condition more students dropped out compared to the two experimental conditions (control versus feedback-only: \( \chi^2(2) = 15.99, p < .05 \), control versus feedback-plus-consultation: \( \chi^2(2) = 13.56, p < .05 \)). The two experimental conditions did not differ \((\alpha = .05)\) in terms of dropout \(\chi^2(2) = 4.83, p = .09\). This difference particularly concerned students who only attended and completed the ISQ once (instead of two or three times). Multilevel-regression analyses on the first measurement occasion (T$_1$) showed significantly \((\alpha = .05)\) higher mean ratings for students who completed the ISQ twice (on T$_1$ and T$_2$, or T$_1$ and T$_3$) \(\beta = .127, SE = .023, t = 5.52\) or three times (on T$_1$, T$_2$ and T$_3$) \(\beta = .116, SE = .023, t = 5.04\) compared to once (on T$_1$). Ratings in the control condition might therefore be slightly biased in terms of more negative on T$_1$, or more positive on T$_2$ and T$_3$, compared to the experimental conditions.

**Consultants**

For this study, five consultants (two male, three female) were trained in SET consultation by the first author. The consultants were experienced faculty members and/or faculty development staff. The collaborative consultation approach, as defined by Brinko (1990), was adapted to this study. Collaborative consultants serve as partners; they encourage their clients to identify, diagnose, and provide solutions to the issues they raise (Brinko, 1990). Therefore, the training of the consultants focused on coaching- and social skills, such as encouraging reflection, and formulation of goals and concrete plans for improvement. Consultants used a consultation protocol (see independent variables) and there were regular meetings between the consultants and first author, to standardize the consultation process.

**Procedure**

The participating professors were assigned to the control condition, feedback-only condition, or the feedback–plus-consultation condition according to a randomized block design. In this design, professors were grouped according to the quality of their teaching (high vs. medium quality) based on previous course evaluations (see section on moderators for grouping information), and their department (departments of Law, Economics, Science, Social and Behavioral Sciences, and Humanities). This resulted in ten groups of professors. Professors of the same department and quality were randomly assigned to one of three conditions to assure equal distributions of these two variables across the three conditions.
Prior to the start of their courses, all professors received procedural instructions by email. Professors sent a standardized email to their students, informing them that they (i.e., the professors) would be participating in a research project on the quality of the lectures at the university. Students were invited to take part by evaluating three lectures during the course. In the final fifteen minutes of the lecture, professors reserved five minutes for an evaluation break. Research assistants distributed the questionnaires and collected them during this break. Students were explicitly instructed to evaluate the current lecture. They were asked to provide their student ID number for research purposes and were assured of anonymity in their evaluations with an extra statement on the ISQ form. The students did not know that their professors were participating in a randomized experiment.

**Independent variables**

**Control condition (N = 25)**

The professors in the control condition received their ISQ results pertaining to the three evaluated lectures at the end of their course. The procedure used with the students was the same as for the experimental conditions. ISQ results contained an overview of number of students, mean student ratings on each item and each dimension, and written answers to the open questions. The three highest and lowest rated items were highlighted.

**Feedback-only condition (experimental condition 1, N = 24)**

Professors in the feedback-only condition received their ISQ results three times, within a week after each evaluated lecture by email. They were free to use the results as they saw fit.

**Feedback–plus-consultation condition (experimental condition 2, N = 26)**

In the feedback–plus-consultation condition, the professor met with a consultant between each evaluated lecture to discuss the ISQ-results. In total there were four meetings: an introductory meeting (prior to the course), two consultation meetings (within three days after the first and after the second evaluated lecture), and a final meeting, after the third evaluated lecture.

*Introductory meeting.* The introduction allowed the consultant and professor to get acquainted. During this meeting, the consultant explained the procedure of feedback and consultation, and the consultation approach.

*Consultation meetings.* The consultation protocol, based on the collaborative approach, involved a five-step procedure: 1) the evaluation of the previous lecture, 2) the evaluation
of the student ratings, 3) the selection of items of the ISQ to improve, 4) the analysis of the 
current situation and problems that explain the selected ratings, and 5) the formulation of 
strategies for improvement.

Because the consultant’s role was to facilitate behavioral change, the professor de-
cided which ISQ items were to be addressed, the identification of areas of improvement, 
and the action plan. The consultants were free to be directive, if they considered that 
expedient, for instance by providing alternative interpretations of the results, alternative 
views when exploring problems in teaching effectiveness, and alternative strategies for 
improvement. Nevertheless, as stipulated in the protocol and in line with the collaborative 
approach, every step of the consultation started and ended with the professor’s views and 
conclusions.

Consultation 2 followed the same protocol as consultation 1. At the beginning of 
consultation 2, the professor reported on his or her experiences in carrying out the previous 
plans made for improvement. The consultant encouraged the professor to reflect on reasons 
for success or failure.

Final meeting. In the final meeting, the professor and consultant again discussed the 
previous lecture and the results of the final student ratings. The consultation ended with an 
evaluation of the program and a discussion of plans for the next course.

Dependent variables

The dependent variables are the scores on the Instructional Skills Questionnaire (ISQ). The 
ISQ was based on the course-evaluation instrument of the University of Amsterdam, the 
Uvalon, developed by Vorst and Van Engelenburg (1992). The Uvalon was based on theo-
ries on effective instruction, and on research on the effects of teaching quality on student 
learning (Frey, Leonard, & Beatty, 1975; De Neve & Janssen, 1982; Janssen & De Neve, 1988; 
Marsh, 1984; Abrami, Apollonia & Cohen, 1990). The Uvalon contains seven dimensions on 
instructional behavior. Its psychometric quality was investigated and confirmed in several 
internal reports from the University of Amsterdam (Vorst & Van Engelenburg, 1992; Verbeek, 

The Uvalon was adapted to a one-lecture instrument, the ISQ, with a selection of 
specific questions on the seven dimensions on instructional behavior. With the ISQ the seven 
dimensions are measured with two positive (indicative) and two negative (contra-indicative) 
worded items on a 7-point likert scale. This resulted a total of 28 items and two open 
questions (“What was good about this lecture?” and “How can this lecture be improved?”).
The psychometric quality of the ISQ was investigated and confirmed with confirmatory factor analyses (see Chapter 2). In more detail, the seven dimensions are:

1. **Structure**: the extent to which the subject matter is handled systematically and in an orderly way. Example item: *The lecture has a clear structure.*
2. **Explication**: the extent to which the instructor explains the subject matter, especially the more complex topics. Example item: *The instructor explains the subject matter clearly.*
3. **Stimulation**: the extent to which the instructor interests students for the subject matter. Example item: *The instructor interests you in the subject matter.*
4. **Validation**: the extent to which the instructor stresses the benefits and the relevance of the subject matter for educational goals or future occupation. Example item: *The instructor indicates the relevance of the subject matter.*
5. **Instruction**: the extent to which the instructor provides instructions about how to study the subject matter. Example item: *The instructor is unclear about which aspects of the subject matter are important* (contra-indicative).
6. **Comprehension**: the extent to which the instructor creates opportunities for questions and remarks regarding the subject matter. Example item: *The instructor encourages students to ask questions about the subject matter.*
7. **Activation**: the extent to which the instructor encourages students to think about and work with the subject matter. Example item: *The instructor involves students in the lecture.*

The dimension score is the student’s mean of the four specific dimension items (the negative worded items are recoded). *Total Instructional Skills* (Total ISQ) is the overall mean score on the ISQ.

Finally, three items were added to the questionnaire to measure the students’ perception of their cognitive, affective, and regulative learning outcomes: “I learned a lot from this lecture” (*Cognition*), “Because of this lecture, I want to learn more about the subject matter” (*Affection*), “Because of this lecture, I now know what I have yet to study” (*Regulation*).

The reliability of the subscales on the professor level were high. Cronbach’s alphas ranged from .88 to .98, with a mean of .93 on T₁, from .92 to .98, with a mean of .94 on T₂, and from .91 to .98, with a mean of .94 on T₃. One reason that these values were quite high is that the professor scores were based on the average test scores of their students. The averages were necessarily subject to less error variance than the student level data.
Missing item responses (3.7%) were imputed with the student's mean of the other three items of that specific dimension. Out of 14,596 forms, 298 forms were excluded; 218 forms remained incomplete after imputation, and 80 forms were marked as extreme outliers. Extreme outliers were detected with the Inter Quartile Range (IQR; distance between the first and the third quartile). For each professor on each measurement occasion separately the IQR was calculated on \textit{Total Instructional Skills}. A form was considered an extreme outlier when the rating was at least two times the IQR lower than the first quartile or two times the IQR higher than the third quartile. This equals a deviance of 3.6 times the standard deviation from the mean.

The final dataset contained 14,298 forms with 527 missing ratings on the student level variable \textit{Cognition}, 237 missing ratings on the variable \textit{Affection}, and 255 missing ratings on the variable \textit{Regulation}.

\textbf{Moderators}

The participating professors made available course evaluation ratings of the same or a similar course that they had given in the previous academic year. The course evaluation instruments and questions differed in formulation and scale. The professor's quality was therefore recoded into two categories, high quality and medium quality professors (there were no notably low quality professors), based on the questions related to the quality of the professor. Professors with a mean rating of 8 or higher on relevant ten-point scale questions or ratings of 4 or higher on five-point scale questions were considered to be high quality professors (\textit{HQ}: coded as 1). Professors with lower ratings, fell in the category medium quality professors (\textit{MQ}: coded as 0). A multilevel \textit{t}-test on baseline mean ratings on \textit{Total Instructional Skills}, measured by the ISQ on the first evaluation occasion, confirmed significant higher ratings for high quality professors compared to medium quality professors (\textit{HQ}: \(\beta = .373, SE = .0798, p < .001\)). The quality of teaching was equally distributed over the conditions (see procedure) when professors were randomly assigned to the control condition (\(N_{\text{HQ}} = 11, N_{\text{MQ}} = 14\)), the feedback-only condition (\(N_{\text{HQ}} = 11, N_{\text{MQ}} = 13\)), or the feedback-plus-consultation condition (\(N_{\text{HQ}} = 12, N_{\text{MQ}} = 14\)). Other demographic information (e.g., age, academic rank, and department) was obtained during recruitment interviews. \textit{Age}, \textit{Quality of Teaching}, and \textit{Class Size} were used as moderators in the multilevel analyses. The moderator \textit{Class Size} represented the number of students who completed the student ratings form on the first measurement occasion (baseline ratings) in each course (\(M = 80.5, SD = 65.3, \text{min} = 13, \text{max} = 365\)). \textit{Age} and \textit{Class Size} were mean-centered at the professor level, rendering the means equal to zero.
Statistical analyses

We used multilevel regression modeling to analyze the data. In so doing, we took into account possible randomness over intercepts and slopes. In addition, we could readily include the variables of interest as moderators of the treatment effects. The dependent variables were scores on Total Instructional Skills, scores on the seven specific teaching dimensions of the ISQ, and scores on the three student level outcome variables. For these dependent variables we conducted the following analysis with MLwiN (Version 2.1; Rasbash, Charlton, Browne, Healy & Cameron, 2009) and R (R Development Core Team, 2008). Appendix I includes the equations for the multilevel models 1 to 7.

Randomization check

To check the randomization, we tested whether the conditions differed with respect to all dependent variables at the baseline measurement occasion. Specifically, for each dependent variable, we fitted two multilevel regression models with students as level 1 variable and professors as level 2 variable. The first model contained an intercept random over professors. The second model contained the additional fixed effect variables Condition (with three conditions). With a deviance test we compared the first and the second model, to analyze the effect of Condition on a 5% significance level. With a deviance test the -2*log-likelihood of one model was compared with the -2*log-likelihood of the other model. If the second model (with the additional variable Condition) did not fit the data better, there was no main effect of Condition, meaning that the baseline ratings were not significantly different for the three conditions. If this was the case for all dependent variables, randomization was successful.

Intra-class correlation

To obtain a measure of clustering, we calculated the intra-class correlations in the intercept only model, which we denote Model 1. Model 1 was fitted on data from all three measurement occasions, with time as level 1 variable, students as level 2 variable and professors as level 3 variable. Model 1 contained an intercept random over professors and students. The variances in ratings were decomposed into the variance of the ratings over time of a given student (\(\sigma^2\)), the variance of the ratings over students of a given professor (\(\tau^2_0\)), and the variance of the ratings over professors (\(\phi^2_0\)). The professor level intra-class correlation (IIC \(_T\)) and the student level intra-class correlation (IIC \(_S\)) were calculated as IIC \(_T\) = \(\phi^2_0/(\sigma^2 + \tau^2_0 + \phi^2_0)\) and IIC \(_S\) = \(\tau^2_0/(\sigma^2 + \tau^2_0 + \phi^2_0)\), respectively. The latter quantifies the degree to which students of a given professor are alike (relative to students of different professors). The former quantifies
the degree to which the professors are alike (in terms of repeated measures by their students relative to the repeated measures of different students).

**Effects of the interventions and correction for random effects**

We analyzed the effects of the interventions on all dependent variables. For each dependent variable, we expanded the intercept-only model (Model 1) with the main effects of *Time* (coded 0,1,2, i.e., we consider the linear effect of time) and *Condition* (Model 2). *Condition* is coded in the feedback-only condition versus the control condition dummy variable (ΔF) and the feedback-plus-consultation condition versus the control condition dummy variable (ΔFC). Next, we expanded Model 2 with the interaction effects *Time*ΔF and *Time*ΔFC (Model 3). With a deviance test we compared Model 2 with Model 3, to analyze the effects of these two interactions. If Model 3 fitted the data better than Model 2 and the parameters of *Time*ΔF and/or *Time*ΔFC were significant on a 5% level, the control condition and the two experimental conditions differed significantly in their ratings over time, hence there is a significant effect of the interventions. Both models contained a random student level intercept and a random professor level intercept.

Next, we extended Model 3 to include a random slope, that is, the effect of time was random at the professor and student level (Model 4). With this procedure we tested if there were individual differences between professors and between students in change of ratings over time. If this model fitted the data better than Model 3, we (re)interpreted the parameters of *Time*ΔF and/or *Time*ΔFC in Model 4, to conclude what the effects are of the interventions with a random slope at the professor and student level taken into account.

Finally, effect sizes were calculated for the effects of the interventions over time. In calculating effect sizes, we followed the rational of basic effect size calculation with single level regression analysis and expanded this rational to the three-level model by adding the random effects of level 2 (students) and level 3 (professors). In addition, we calculated Cohen’s *d* based on the professors mean ratings and standard deviation to be able to compare results with previous findings of studies that did not apply multilevel modeling. We note that Cohen’s *d* likely overestimates the effects, since the nested structure of the data and present random effects are not taken into account. Taking random effects into account often increases the estimates’ standard error (Hox, 2002).

Cohen’s *d* was calculated in two ways. The first Cohen’s *d* was calculated for each condition by dividing it’s mean difference of T₃ and T₁ by its pooled standard deviation (√((SD (T₁,condition)² + SD (T₃,condition)²) / 2). The second Cohen’s *d* was calculated for each experimental condition versus the control condition by dividing the mean difference of T₃
and T₁ of the experimental condition minus the mean difference of T₃ and T₁ of the control condition by its pooled standard deviation \( \sqrt{(SD(T₁_{control})^2 + SD(T₃_{control})^2 + SD(T₁_{condition})^2 + SD(T₃_{condition})^2) / 4} \). Multilevel effect sizes were calculated based on the multilevel modeling output of the final model. Two times the beta of \( Time \) represents the change in mean of the control condition between T₁ and T₃, two times the beta of \( Time*ΔF \) represents the change in mean of the feedback-only condition compared to the control condition between T₁ and T₃, and two times the beta of \( Time*ΔFC \) represents the change in mean of the feedback-plus-consultation condition compared to the control condition between T₁ and T₃. The residual standard deviation \( SD(e_{ij}) \) equals the standard deviation of \( Y_{ij} \) over time of a given student i for a given professor j (assuming homoskedasticity). To standardize the effect, we expressed the effect size as a function of this within student, within professor residual standard deviation. The effect size of the control condition was therefore calculated by dividing two times the beta of \( Time \) by the residual standard deviation \( SD(e_{ij}) \), the effect size of the feedback-only condition compared to the control condition was calculated by dividing two times the beta of \( Time*ΔF \) by the residual standard deviation \( SD(e_{ij}) \), and the effect size of the feedback-plus-consultation condition compared to the control condition was calculated by dividing two times the beta of \( Time*ΔFC \) by the residual standard deviation \( SD(e_{ij}) \). Effect sizes of .2, .5 and .8 were considered as small, medium and large effects respectively (Cohen, 1988).

**Effects of targeted versus non-targeted dimensions**

With Model 5, we analyzed the effect of dimensions that were targeted for improvement during the consultation meetings versus the effect of non-targeted dimensions on each of the seven specific teaching dimensions on each time interval. These additional exploratory analyses were done to link the effects of the feedback-plus-consultation intervention to the specific content of the consultation. In Model 5, professors in the feedback-plus-consultation condition were separated into two groups for each dimension on each time interval based on the consultation reports; a group which targeted the dimension for improvement (Target), meaning that they made concrete plans for improvement, and a group that did not target the dimension (No Target). Condition was therefore recoded into the dummy variables Control-versus-Feedback-only (denoted as \( ΔF \)), Control_versus_Feedback-plus-Consultation_No Target (denoted as \( ΔFC_{NoTarget} \)) and Control_versus_Feedback-plus-Consultation_Target (denoted as \( ΔFC_{Target} \)). Time was recoded for the specific time interval (in case of time interval \( T₁T₂; T₁ = 0 \) and \( T₂ = 1 \), and in case of time interval \( T₂T₃; T₂ = 0 \) and \( T₃ = 1 \)). We did not have enough data to fit a model with these additional parameters plus the parameters for all possible random effects. We therefore limited the random effects to the professor level.
and student level intercept in this model. To compensate, the parameters were tested with a more restricted alpha of .01.

**Moderators**

Finally, we modeled the moderating effects of two professor-level moderators, *Age, Quality of Teaching* and *Class Size*. Let M denote a moderator of interest. First we added the main effect of M (γ₀₀₃) to Model 4 (Model 6). Next, we added the interaction-effects of M*ΔF (γ₀₀₄), M*ΔFC (γ₀₀₅), M*Time (γ₁₀₃), M*Time*ΔF (γ₁₀₄), and M*Time*ΔFC (γ₁₀₅) to Model 6 (Model 7). The interaction effects M*Time*ΔF and M*Time*ΔFC, represent the separate effects of the two interventions for professors with high and low ratings on the specific moderator, compared to the control condition. The significance of the main effects were tested with a deviance test on Model 4 compared to Model 6 on a 5% significance level. The significance of the interaction effects were tested with a deviance test on Model 6 compared to Model 7 on a 5% significance level. Appendix I includes the equations for Model 6 and 7.

**Results**

**Descriptives**

Table 5.1 shows mean ratings, the professor-level standard deviation and the student level standard deviation at each measurement occasion in each condition of all dependent variables. On each dependent variable, mean ratings did not differ with respect to *Condition* at baseline (e.g., *Total Instructional Skills* T₁; β = .029, SE = .055, p = .596), indicating successful randomization.

The professor-level intra-class correlation varied between .06 and .35, with a mean of .19. The student-level intra-class correlations varied between .15 and .38, with a mean of .28. With the subsequent multilevel analyses, we took this clustering into account.

**Effects of the interventions, in the presence of random effects**

Table 5.2 shows the estimates and standard errors of the four models on *Total Instructional Skills*. Deviance tests show that Model 3 (with the two Time*Condition interaction effects) fitted the data significantly better than Model 2 (without the interaction effects) (χ²(2) = 76.9, p < .001). However, Model 4 (with an additional random slope at the professor and student level) fitted the data significantly better than Model 3 (without a random slope) (χ²(2) =
Table 5.1  Mean student ratings and standard deviations of the dependent variables on professor and student level on each measurement occasion in each condition

<table>
<thead>
<tr>
<th></th>
<th>T₁ (within groups)</th>
<th>T₂ (within groups)</th>
<th>T₃ (within groups)</th>
<th>T₁ (within groups)</th>
<th>T₂ (within groups)</th>
<th>T₃ (within groups)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Min</td>
<td>Median</td>
<td>Max</td>
<td>Min</td>
<td>Median</td>
<td>Max</td>
</tr>
<tr>
<td><strong>Control condition</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Professor level variables</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Structure</td>
<td>5.17</td>
<td>0.34</td>
<td>0.64</td>
<td>0.87</td>
<td>1.20</td>
<td>5.16</td>
</tr>
<tr>
<td>Explication</td>
<td>5.36</td>
<td>0.39</td>
<td>0.69</td>
<td>0.86</td>
<td>1.12</td>
<td>5.29</td>
</tr>
<tr>
<td>Stimulation</td>
<td>4.52</td>
<td>0.70</td>
<td>0.83</td>
<td>1.01</td>
<td>1.33</td>
<td>4.56</td>
</tr>
<tr>
<td>Validation</td>
<td>4.80</td>
<td>0.38</td>
<td>0.47</td>
<td>0.88</td>
<td>1.16</td>
<td>4.91</td>
</tr>
<tr>
<td>Instruction</td>
<td>4.72</td>
<td>0.29</td>
<td>0.72</td>
<td>0.87</td>
<td>1.23</td>
<td>4.72</td>
</tr>
<tr>
<td>Comprehension</td>
<td>4.90</td>
<td>0.51</td>
<td>0.65</td>
<td>0.80</td>
<td>1.08</td>
<td>4.89</td>
</tr>
<tr>
<td>Activation</td>
<td>4.62</td>
<td>0.53</td>
<td>0.67</td>
<td>0.90</td>
<td>1.08</td>
<td>4.75</td>
</tr>
<tr>
<td>Total ISQ</td>
<td>4.87</td>
<td>0.31</td>
<td>0.54</td>
<td>0.64</td>
<td>0.81</td>
<td>4.90</td>
</tr>
<tr>
<td>Student level variables</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cognition</td>
<td>4.95</td>
<td>0.40</td>
<td>0.90</td>
<td>1.16</td>
<td>1.47</td>
<td>5.04</td>
</tr>
<tr>
<td>Affection</td>
<td>4.26</td>
<td>0.45</td>
<td>1.12</td>
<td>1.39</td>
<td>1.61</td>
<td>4.34</td>
</tr>
<tr>
<td>Regulation</td>
<td>4.79</td>
<td>0.41</td>
<td>0.88</td>
<td>1.33</td>
<td>1.69</td>
<td>4.96</td>
</tr>
<tr>
<td><strong>Feedback-only condition</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Professor level variables</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Structure</td>
<td>5.21</td>
<td>0.43</td>
<td>0.59</td>
<td>0.88</td>
<td>1.23</td>
<td>5.20</td>
</tr>
<tr>
<td>Explication</td>
<td>5.43</td>
<td>0.52</td>
<td>0.51</td>
<td>0.84</td>
<td>1.23</td>
<td>5.40</td>
</tr>
<tr>
<td>Stimulation</td>
<td>4.83</td>
<td>0.84</td>
<td>0.54</td>
<td>1.01</td>
<td>1.37</td>
<td>4.87</td>
</tr>
<tr>
<td>Validation</td>
<td>4.97</td>
<td>0.35</td>
<td>0.53</td>
<td>0.88</td>
<td>1.27</td>
<td>4.97</td>
</tr>
<tr>
<td>Instruction</td>
<td>4.73</td>
<td>0.33</td>
<td>0.61</td>
<td>0.95</td>
<td>1.24</td>
<td>4.72</td>
</tr>
<tr>
<td>Comprehension</td>
<td>5.03</td>
<td>0.53</td>
<td>0.46</td>
<td>0.83</td>
<td>1.05</td>
<td>4.95</td>
</tr>
<tr>
<td>Activation</td>
<td>4.93</td>
<td>0.67</td>
<td>0.46</td>
<td>0.94</td>
<td>1.20</td>
<td>4.89</td>
</tr>
<tr>
<td>Total ISQ</td>
<td>5.02</td>
<td>0.42</td>
<td>0.30</td>
<td>0.64</td>
<td>0.92</td>
<td>5.00</td>
</tr>
</tbody>
</table>
Table 5.1  
Continued

<table>
<thead>
<tr>
<th></th>
<th>( T_1 )</th>
<th>( T_2 )</th>
<th>( T_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Min</td>
<td>Median</td>
<td>Max</td>
</tr>
<tr>
<td><strong>Feedback-only condition</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Professor level variables</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Structure</td>
<td>5.19</td>
<td>0.44</td>
<td>0.61</td>
</tr>
<tr>
<td>Explication</td>
<td>5.33</td>
<td>0.54</td>
<td>0.58</td>
</tr>
<tr>
<td>Stimulation</td>
<td>4.62</td>
<td>0.75</td>
<td>0.68</td>
</tr>
<tr>
<td>Validation</td>
<td>4.73</td>
<td>0.40</td>
<td>0.65</td>
</tr>
<tr>
<td>Instruction</td>
<td>4.68</td>
<td>0.35</td>
<td>0.70</td>
</tr>
<tr>
<td>Comprehension</td>
<td>5.11</td>
<td>0.61</td>
<td>0.58</td>
</tr>
<tr>
<td>Activation</td>
<td>4.87</td>
<td>0.86</td>
<td>0.47</td>
</tr>
<tr>
<td>Total ISQ</td>
<td>4.93</td>
<td>0.42</td>
<td>0.47</td>
</tr>
<tr>
<td>Student level variables</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cognition</td>
<td>4.96</td>
<td>0.48</td>
<td>0.82</td>
</tr>
<tr>
<td>Affection</td>
<td>4.38</td>
<td>0.49</td>
<td>1.06</td>
</tr>
<tr>
<td>Regulation</td>
<td>4.79</td>
<td>0.38</td>
<td>0.91</td>
</tr>
</tbody>
</table>
Table 5.2  Estimates and standard errors of the four models for Total Instructional Skills

<table>
<thead>
<tr>
<th>Total Instructional Skills</th>
<th>Model 1</th>
<th></th>
<th>Model 2</th>
<th></th>
<th>Model 3</th>
<th></th>
<th>Model 4</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>SE</td>
<td>Estimate</td>
<td>SE</td>
<td>Estimate</td>
<td>SE</td>
<td>Estimate</td>
<td>SE</td>
</tr>
<tr>
<td>Fixed Part</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>4.942</td>
<td>***</td>
<td>0.046</td>
<td></td>
<td>4.845</td>
<td>0.079</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>0.020</td>
<td>**</td>
<td>0.006</td>
<td></td>
<td>-0.033</td>
<td>**</td>
<td>0.010</td>
<td></td>
</tr>
<tr>
<td>ΔFeedback-only</td>
<td>0.119</td>
<td></td>
<td>0.112</td>
<td></td>
<td>0.085</td>
<td>0.113</td>
<td></td>
<td>0.116</td>
</tr>
<tr>
<td>ΔFeedback-plus-Consultation</td>
<td>0.124</td>
<td></td>
<td>0.110</td>
<td></td>
<td>0.024</td>
<td>0.111</td>
<td></td>
<td>0.060</td>
</tr>
<tr>
<td>Time*ΔFeedback-only</td>
<td>0.043</td>
<td>**</td>
<td>0.014</td>
<td></td>
<td>0.011</td>
<td>0.037</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time*ΔFeedback-plus-Consultation</td>
<td>0.126</td>
<td>***</td>
<td>0.014</td>
<td></td>
<td>0.084</td>
<td>*</td>
<td>0.036</td>
<td></td>
</tr>
<tr>
<td>Random Part</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>0.153</td>
<td>***</td>
<td>0.026</td>
<td></td>
<td>0.149</td>
<td>***</td>
<td>0.025</td>
<td></td>
</tr>
<tr>
<td>Level 3: Professor</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 2: Student</td>
<td>0.195</td>
<td>***</td>
<td>0.006</td>
<td></td>
<td>0.195</td>
<td>***</td>
<td>0.006</td>
<td></td>
</tr>
<tr>
<td>Level 1: Time</td>
<td>0.237</td>
<td>***</td>
<td>0.005</td>
<td></td>
<td>0.237</td>
<td>***</td>
<td>0.005</td>
<td></td>
</tr>
<tr>
<td>Intercept-slope covariance</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 3: Professor</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 2: Student</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-2 log likelihood</td>
<td>27686.9</td>
<td></td>
<td>27674.1</td>
<td></td>
<td>27597.2</td>
<td></td>
<td>27265.0</td>
<td></td>
</tr>
</tbody>
</table>

Notes: * p < .05, ** p < .01, *** p < .001, units level 3: 75 professors, units level 2: 9,616 students, units level 1: 14,298 observations, variances of .000 are smaller than .0001.
Experiment II – Effects on professors’ lecturing skills and student learning

Chapter 5

332.2, \( p < .001 \). This signifies that professors and students vary significantly in ratings over time. These variances needed to be taken into account when the effects of the intervention were tested. Model 4 was therefore indicated as the final model. Model 4 also fitted the data best in the analyses of the seven dependent teaching variables and the three student level outcome variables. Table 5.3 shows the estimates and standard errors of all specific dependent variables on Model 4.\(^4\)

In Model 4, the interaction parameters of \( Time^* \Delta F \) were not significant for all dependent variables. The effect for \( Time^* \Delta FC \) were significant for the professor level variables Structure, Validation, Instruction, Total Instructional Skills and the student level variable Cognition (Structure: \( \beta = .092, SE = .043, p = .032 \), Validation: \( \beta = .105, SE = .046, p = .023 \), Instruction: \( \beta = .114, SE = .046, p = .013 \), Total Instructional Skills: \( \beta = .084, SE = .037, p = .023 \), Cognition: \( \beta = .121, SE = .053, p = .022 \)).

Furthermore, ratings on Explication appeared to significantly decrease for the control condition in Model 4 (\( \beta = -.083, SE = .035, p = .018 \)). Ratings of the control condition on the student variable Regulation showed a significant increase (\( \beta = .084, SE = .039, p = .031 \)). In both cases, the two experimental conditions did not significantly differ from the control condition, meaning that they showed an equal pattern of decrease and increase in ratings over time.

The effect sizes calculated with Cohen's \( d \) and calculated with multilevel regression Model 4 output are given in Table 5.4. In terms of Cohen's \( d \), the effects of the feedback-only condition compared to the control condition ranged from -.25 to .22 (mean value of -.06), and the effects of the feedback-plus-consultation condition compared to the control condition ranged from .09 to .60 (mean value of .35). The effects of the feedback-plus-consultation condition on the five variables that were found significant were medium (ranging from .43 to .60). In terms of effect sizes based on the multilevel output, the effects of the feedback-only condition compared to the control condition ranged from -.09 to .16 (mean value of .01), and the effects of the feedback-plus-consultation condition compared to the control condition ranged from .09 to .36 (mean value of .23). The effects on the five variables that were found significant were smaller according to calculation on the multilevel output (ranging from .26 to .36). Overall, effect sizes indicated no effects (one small effect) of the feedback-only condition and small to medium effects of the feedback-plus-consultation condition.

We note that, in Model 3, the effects of the feedback-only condition interaction were significant at the 5% level on four dependent variables (Explanation, Stimulation, Instruction and Total Instructional Skills) and the effects of the feedback-plus-consultation condition were

\(^4\) Tables of detailed results of all models on all dependent variables are available on request.
Table 5.3  Estimates and standard errors of Model 4 for each specific dependent variable

<table>
<thead>
<tr>
<th></th>
<th>Students' perceptions of learning outcomes</th>
<th>ISQ teaching dimensions 1-2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cognition</td>
<td>Affection</td>
</tr>
<tr>
<td>Fixed Part</td>
<td>Estimate</td>
<td>SE</td>
</tr>
<tr>
<td>Intercept</td>
<td>4.974 ***</td>
<td>0.092</td>
</tr>
<tr>
<td>Time</td>
<td>0.000</td>
<td>0.037</td>
</tr>
<tr>
<td>ΔFeedback-only</td>
<td>0.087</td>
<td>0.132</td>
</tr>
<tr>
<td>ΔFeedback-plus-</td>
<td>-0.022</td>
<td>0.129</td>
</tr>
<tr>
<td>Consultation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time*ΔFeedback-only</td>
<td>-0.013</td>
<td>0.054</td>
</tr>
<tr>
<td>Time*ΔFeedback-plus-</td>
<td>0.121 *</td>
<td>0.053</td>
</tr>
<tr>
<td>Consultation</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Random Part

Intercept

| Level 3: Professor  | 0.191 *** | 0.035 | 0.253 *** | 0.046 | 0.104 *** | 0.021 | 0.138 *** | 0.025 | 0.215 *** | 0.037 |
| Level 2: Student    | 0.649 *** | 0.031 | 0.509 *** | 0.026 | 0.450 *** | 0.024 | 0.279 *** | 0.011 | 0.287 *** | 0.011 |
| Level 1: Time       | 0.676 *** | 0.020 | 1.304 *** | 0.026 | 1.236 *** | 0.035 | 0.503 *** | 0.010 | 0.446 *** | 0.009 |

Slope

| Level 3: Professor  | 0.024 *** | 0.006 | 0.021 *** | 0.006 | 0.023 *** | 0.006 | 0.016 *** | 0.004 | 0.024 *** | 0.005 |
| Level 2: Student    | 0.041 *   | 0.016 | 0.000     | 0.000 | 0.021     | 0.026 | 0.000     | 0.000 | 0.000     | 0.000 |

Intercept-slope covariance

| Level 3: Professor  | 0.003     | 0.010 | 0.006     | 0.012 | -0.017    | 0.009 | -0.008    | 0.007 | -0.020 *  | 0.010 |
| Level 2: Student    | -0.036    | 0.018 | 0.000     | 0.000 | -0.044    | 0.028 | 0.000     | 0.000 | 0.000     | 0.000 |

Units level 3: 75
Units level 2: 9235
Units level 1: 13771
Table 5.3  *Continued*

<table>
<thead>
<tr>
<th>ISQ teaching dimensions 3-7</th>
<th>Stimulation</th>
<th>Validation</th>
<th>Instruction</th>
<th>Comprehension</th>
<th>Activation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate</td>
<td>$SE$</td>
<td>Estimate</td>
<td>$SE$</td>
<td>Estimate</td>
<td>$SE$</td>
</tr>
<tr>
<td>Interception</td>
<td>4.531 **</td>
<td>0.150</td>
<td>4.199 **</td>
<td>0.074</td>
<td>4.725 ***</td>
</tr>
<tr>
<td>Time</td>
<td>0.005</td>
<td>0.042</td>
<td>0.026</td>
<td>0.032</td>
<td>-0.040</td>
</tr>
<tr>
<td>ΔFeedback-only</td>
<td>0.290</td>
<td>0.214</td>
<td>0.128</td>
<td>0.106</td>
<td>-0.050</td>
</tr>
<tr>
<td>ΔFeedback-plus-Consultation</td>
<td>0.091</td>
<td>0.210</td>
<td>-0.091</td>
<td>0.104</td>
<td>-0.058</td>
</tr>
<tr>
<td>Time*ΔFeedback-only</td>
<td>0.024</td>
<td>0.060</td>
<td>-0.028</td>
<td>0.046</td>
<td>0.054</td>
</tr>
<tr>
<td>Time*ΔFeedback-plus-Consultation</td>
<td>0.060</td>
<td>0.060</td>
<td>0.105</td>
<td>*</td>
<td>0.046</td>
</tr>
</tbody>
</table>

**Fixed Part**

**Random Part**

*Intercept*

| Level 3: Professor | 0.542 *** | 0.091 | 0.125 *** | 0.023 | 0.094 *** | 0.017 | 0.291 *** | 0.049 | 0.477 *** | 0.080 |
| Level 2: Student | 0.425 *** | 0.015 | 0.315 *** | 0.019 | 0.335 *** | 0.020 | 0.261 *** | 0.010 | 0.227 *** | 0.012 |
| Level 1: Time | 0.613 *** | 0.012 | 0.468 *** | 0.014 | 0.477 *** | 0.014 | 0.418 *** | 0.008 | 0.641 *** | 0.012 |

*Slope*

| Level 3: Professor | 0.035 *** | 0.007 | 0.019 *** | 0.004 | 0.019 *** | 0.004 | 0.022 *** | 0.005 | 0.040 *** | 0.008 |
| Level 2: Student | 0.000 | 0.000 | 0.024 * | 0.011 | 0.040 *** | 0.011 | 0.000 | 0.000 | 0.000 | 0.000 |

*Intercept-slope covariance*

| Level 3: Professor | -0.019 | 0.019 | -0.011 | 0.007 | -0.011 | 0.006 | -0.033 ** | 0.012 | -0.066 *** | 0.020 |
| Level 2: Student | 0.000 | 0.000 | -0.012 | 0.012 | -0.019 | 0.012 | 0.000 | 0.000 | 0.000 | 0.000 |

Units level 3: 75
Units level 2: 9616
Units level 1: 14298

Notes: * $p < .05$, ** $p < .01$, *** $p < .001$, variances of .000 are smaller than .0001.
Table 5.4  Effect sizes calculated with Cohen's $d$ and calculated with multilevel regression Model 4 output

<table>
<thead>
<tr>
<th>Comparison:</th>
<th>Control</th>
<th>Feedback-only</th>
<th>Feedback-plus-Consultation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$(T_3 \text{ vs } T_1)$</td>
<td>$(T_3 \text{ vs } T_1)$ vs Control $(T_3 \text{ vs } T_1)$</td>
<td>$(T_3 \text{ vs } T_1)$ vs Control $(T_3 \text{ vs } T_1)$</td>
</tr>
<tr>
<td>Effect size:</td>
<td>Cohen's $d$</td>
<td>multilevel model</td>
<td>Cohen's $d$</td>
</tr>
<tr>
<td>Calculation:</td>
<td>$\frac{M(T_{3,c}) - M(T_{1,c})}{\text{pooled SD}}$ $2\times \text{Time}$</td>
<td>$\frac{M(T_{3_{FB}}) - M(T_{1_{FB}})}{\text{pooled SD}}$</td>
<td>$(\frac{M(T_{3_{FB}}) - M(T_{1_{FB}})}{\text{pooled SD}})$ $- (\frac{M(T_{3,c}) - M(T_{1,c})}{\text{pooled SD}})$</td>
</tr>
</tbody>
</table>

**Professor level**

| Structure     | -0.12 | -0.07 | -0.19 | -0.11 | 0.01 | 0.29 | 0.43 | 0.26 |
| Explication   | -0.38 | -0.25 | -0.22 | 0.08  | 0.11 | 0.02 | 0.34 | 0.23 |
| Stimulation   | 0.07  | 0.01  | 0.06  | 0.00  | 0.06 | 0.20 | 0.13 | 0.15 |
| Validation    | 0.19  | 0.08  | -0.06 | -0.25 | -0.08 | 0.63 | 0.49 | 0.31 |
| Instruction   | -0.26 | -0.12 | 0.00  | 0.22  | 0.16 | 0.34 | 0.60 | 0.33 |
| Comprehension | -0.07 | 0.02  | -0.09 | -0.02 | -0.03 | 0.22 | 0.30 | 0.20 |
| Activation    | 0.15  | 0.17  | 0.07  | -0.06 | -0.05 | 0.33 | 0.25 | 0.18 |
| Total ISQ     | -0.05 | -0.03 | -0.06 | -0.03 | 0.05 | 0.35 | 0.43 | 0.36 |

**Student level**

| Cognition     | 0.09  | 0.00  | -0.02 | -0.09 | -0.03 | 0.56 | 0.50 | 0.29 |
| Affection     | 0.15  | 0.05  | -0.02 | -0.14 | -0.05 | 0.24 | 0.09 | 0.09 |
| Regulation    | 0.38  | 0.15  | 0.15  | -0.25 | -0.09 | 0.71 | 0.26 | 0.09 |

Notes: The control condition is denoted C, the feedback-only condition is denoted FB, the feedback-plus-consultation condition is denoted FC. Pooled SD for the comparison $T_3 \text{ vs } T_1$ of the control condition was calculated with $\sqrt{((\text{SD}(T_{1_{control}})^2 + \text{SD}(T_{4_{control}})^2)/2)$. Pooled SD for the comparison $T_3 \text{ vs } T_1$ of the experimental conditions versus the control condition was calculated with $\sqrt{((\text{SD}(T_{1_{control}})^2 + \text{SD}(T_{4_{control}})^2 + \text{SD}(T_{1_{experimental}})^2 + \text{SD}(T_{4_{experimental}})^2)/4)}}.$
significant on all seven professor level dependent variables and on two out of three student level variables (Cognition and Affection). The inclusion of a random slope on the professor and student level in Model 4 rendered the effects of both interventions on multiple variables statistically insignificant. This indicates the importance of taking random variations in ratings over time into account when analyzing student ratings data.

**Effects of targeted versus non-targeted dimensions**

On each time interval \((T_1T_2, T_2T_3)\) we investigated differences in effects between dimensions that were targeted or not targeted for improvement during the consultation. On the first time interval \((T_1T_2)\) there was a significant improvement \((p < .01)\) for targeted dimensions in the feedback-plus-consultation condition compared to the improvement made by the control condition on six out of seven teaching dimensions (Structure: \(\beta = .125, SE = .048, p = .009\), Explanation: \(\beta = .223, SE = .059, p < .001\), Stimulation: \(\beta = .155, SE = .048, p = .001\), Validation: \(\beta = .176, SE = .049, p < .001\), Comprehension: \(\beta = .532, SE = .049, p < .001\), Activation: \(\beta = .231, SE = .052, p < .001\), no significant effects for Instruction). Non-targeted dimensions did not improve significantly on any of the seven dimensions compared to the control condition. This was similar to results of the feedback-only condition. Figure 5.1 shows the mean improvement within each condition on each dimension on the first time interval.

![Graph showing mean improvement in ratings](image.png)

**Figure 5.1** Improvement in mean ratings on the first time interval \((T_1T_2)\) for each condition on each ISQ teaching dimension.
We note that professors who did not target the dimension *Comprehension* had ratings on T₁ that were significantly higher than the control condition (Δ*Feedback-plus-Consultation_No-Target*; β = .401, SE = .160, p = .012). On the other six dimensions, there were no significant differences between the conditions on their baseline ratings.

As found in the previous analyses, ratings of the control condition on the dimension *Explication* decreased significantly over time (Time; β = -.121, SE = .025, p < .001). In addition, on this time interval, the control condition significantly increased in ratings on the dimension *Activation* (Time; β = .135, SE = .029, p < .001). Professors who targeted this dimension significantly increased in ratings on top of this increase of the control condition.

On the second time interval (T₂T₃) there were fewer effects of the experimental conditions. Professors who targeted the dimensions *Instruction* and *Activation* increased their ratings significantly compared to the control condition (*Instruction*: β = .173, SE = .050, p < .001, *Activation*: β = .255, SE = .061, p < .001). At the same time, professors who did not target the dimension *Validation* increased their ratings significantly over time compared to the control condition (*Validation*: β = .154, SE = .046, p < .001). Figure 5.2 shows the mean improvement for each condition on each dimension on the second time interval. We note that

---

**Figure 5.2**  Improvement in mean ratings on the second time interval (T₂T₃) for each condition on each ISQ teaching dimension.
on this time interval, ratings of the control condition on Explication decreased significantly again \( (Time; \beta = -0.075, SE = .028, p = .007) \).

In sum, most effects occurred on the first time interval and dimensions that were targeted for improvement during consultation increased more in ratings over time than non-targeted dimensions. This indicates that the effects found on the feedback-plus-consultation condition are due to the consultation rather than due to a Hawthorne effect.

**Moderators Age, Quality of Teaching and Class Size**

We investigated the influence of the professors' Age, Quality of Teaching and Class Size on the treatment effects. These variables were added as main effects (Model 6) and as moderators, by including the interactions with Time, Condition and Time*Condition (Model 7).

Deviance tests between Model 4 and Model 6 indicated that there was a main effect of Age on five out of seven teaching dimensions (i.e., Total Instructional Skills, Structure, Explication, Stimulation, Validation and Instruction), and on the student learning variable Cognition (e.g., Total Instructional Skills: \( \chi^2(1) = 8.495, p = .004 \)). On all of these variables younger professors received higher ratings compared to older colleagues on a 5% level (e.g., Total Instructional Skills: \( \beta = -0.014, SE = .005, p = .005 \)).

There was a main effect of Quality of Teaching on all professor level and student level dependent variables, except for the teaching dimension Comprehension (e.g., Total Instructional Skills: \( \chi^2(1) = 16.621, p < .001 \)). Except for Activation, the parameter of Quality of Teaching was significant on all of these dependent variables (e.g., Total Instructional Skills: \( \beta = .355, SE = .080, p < .001 \)). This indicated that high quality professors received higher ratings by their students (as expected) and students' perceptions of their learning outcomes were higher when they attended lectures taught by these professors.

Finally, deviance tests indicated that there was a main effect of Class Size on all professor and student level dependent variables (e.g., Total Instructional Skills: \( \chi^2(1)= 1092.295, p < .001 \)). The parameter of Class Size was significant on the teaching dimensions Comprehension and Activation and indicated that professors who taught larger classes received lower ratings on these dimensions (i.e., Comprehension: \( \beta = -.003, SE = .001, p = .003 \), Activation: \( \beta = -.004, SE = .001, p < .001 \)).

Deviance tests between Model 6 and 7 indicated that the interaction effects were not significant for Age and Class Size as moderators for all dependent variables (e.g., Age: Total Instructional Skills: \( \chi^2(5) = 4.453, p = .486 \), Class Size: Total Instructional Skills: \( \chi^2(5) = 2.682, p = .749 \)). Thus, the effectiveness of the interventions did not differ for professors from different ages and with different class sizes.
Table 5.5  Estimates and standard errors of Model 7 with moderators Age, Quality of Teaching and Class Size for Total Instructional Skills

<table>
<thead>
<tr>
<th>Total Instructional Skills</th>
<th>Model Moderator</th>
<th>Model Moderator</th>
<th>Model Moderator</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Age</td>
<td>Quality of Teaching</td>
<td>Class Size</td>
</tr>
<tr>
<td></td>
<td>Estimate</td>
<td>SE</td>
<td>Estimate</td>
</tr>
<tr>
<td><strong>Fixed Part</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>4.849</td>
<td>0.073</td>
<td>4.825</td>
</tr>
<tr>
<td>Time</td>
<td>-0.010</td>
<td>0.025</td>
<td>0.059</td>
</tr>
<tr>
<td>ΔFeedback-only</td>
<td>0.200</td>
<td>0.109</td>
<td>-0.048</td>
</tr>
<tr>
<td>ΔFeedback-plus-Consultation</td>
<td>0.054</td>
<td>0.104</td>
<td>-0.149</td>
</tr>
<tr>
<td>Time*ΔFeedback-only</td>
<td>-0.008</td>
<td>0.037</td>
<td>-0.041</td>
</tr>
<tr>
<td>Time*ΔFeedback-plus-Consultation</td>
<td>0.087 *</td>
<td>0.036</td>
<td>0.019</td>
</tr>
<tr>
<td>Moderator</td>
<td>-0.014 *</td>
<td>0.007</td>
<td>0.088</td>
</tr>
<tr>
<td>Moderator*Time</td>
<td>-0.002</td>
<td>0.002</td>
<td>-0.147 **</td>
</tr>
<tr>
<td>Moderator*ΔFeedback-only</td>
<td>-0.005</td>
<td>0.011</td>
<td>0.355</td>
</tr>
<tr>
<td>Moderator*ΔFeedback-plus-Consultation</td>
<td>0.005</td>
<td>0.011</td>
<td>0.449 **</td>
</tr>
<tr>
<td>Moderator<em>ΔFeedback-only</em>Time</td>
<td>0.007</td>
<td>0.004</td>
<td>0.115</td>
</tr>
<tr>
<td>Moderator<em>ΔFeedback-plus-Consultation</em>Time</td>
<td>0.002</td>
<td>0.004</td>
<td>0.145 *</td>
</tr>
</tbody>
</table>
Table 5.5  Continued

<table>
<thead>
<tr>
<th>Total Instructional Skills</th>
<th>Model Moderator Age</th>
<th>Model Moderator Quality of Teaching</th>
<th>Model Moderator Class Size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>SE</td>
<td>Estimate</td>
</tr>
<tr>
<td>Random Part</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 3: Professor</td>
<td>0.126</td>
<td>***</td>
<td>0.101</td>
</tr>
<tr>
<td>Level 2: Student</td>
<td>0.203</td>
<td>***</td>
<td>0.203</td>
</tr>
<tr>
<td>Level 1: Time</td>
<td>0.218</td>
<td>***</td>
<td>0.218</td>
</tr>
<tr>
<td>Slope</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 3: Professor</td>
<td>0.012</td>
<td>***</td>
<td>0.011</td>
</tr>
<tr>
<td>Level 2: Student</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Intercept-slope covariance</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 3: Professor</td>
<td>-0.002</td>
<td>0.005</td>
<td>-0.001</td>
</tr>
<tr>
<td>Level 2: Student</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>-2 log likelihood</td>
<td>27252.080</td>
<td></td>
<td>27232.284</td>
</tr>
<tr>
<td>Deviance with Model 5 (df = 5)</td>
<td>4.452</td>
<td>16.130 **</td>
<td></td>
</tr>
</tbody>
</table>

Notes: *p < .05, ** p < .01, *** p < .001. Age and Quality of Teaching; units level 3: 75 professors, units level 2: 9,616 students, units level 1: 14,298 observations. Class Size; Age and Quality of Teaching; units level 3: 73 professors, units level 2: 9,189 students, units level 1: 13,719 observations. Variances of .000 are smaller than .0001.
When the moderator Quality of Teaching was added, deviance tests showed significant effects of the interactions on Total Instructional Skills ($\chi^2(5) = 16.13, p = .007$), and the teaching dimensions Explication ($\chi^2(5) = 12.73, p = .026$), Stimulation ($\chi^2(5) = 16.09, p = .007$) and Validation ($\chi^2(5) = 19.81, p = .001$). Thus, Quality of Teaching moderated the treatment effects on these variables. On these four variables, medium quality professors in the experimental conditions did not improve their ratings compared to the control condition at all. The effects occurred mainly on high quality professors.

Professors with a high quality of teaching in the feedback-plus-consultation condition show a significant increase in ratings compared to the control condition on Validation ($\beta = .201$, $SE = .087, p = .021$), and Total Instructional Skills ($\beta = .145$, $SE = .068, p = .033$). Professors with a high quality of teaching in the feedback-only condition, show a significant increase in ratings on Stimulation ($\beta = .241$, $SE = .113, p = .033$), compared to the control condition. At the same time, professors with high quality teaching in the control condition significantly decreased in ratings on all four dependent variables (Explication: $\beta = -.155$, $SE = .066, p = .019$, Stimulation: $\beta = -.221, SE = .078, p = .005$, Validation: $\beta = -.148, SE = .061, p = .015$, Total Instructional Skills: $\beta = -.147, SE = .048, p = .002$). This indicates that the effects of the interventions are visible in terms of an increase in ratings as well as the prevention of decrease for high quality professors. We do note that for these four variables the feedback-plus-consultation condition started out significantly higher on baseline ratings compared to the control condition. The two intervention conditions did not differ significantly from each other on baseline ratings. Table 5.5 shows the estimates and standard errors of Model 7 with moderators on Total Instructional Skills.5

Discussion

Instructional development practices are seldom investigated with a rigorous experimental design (Levinson-Rose & Menges, 1981; Prebble, Hargraves, Leach, Naidoo, Suddaby, & Zepke, 2004; Stess, Min-Leliveld, Gijbels, & Van Petegem, 2010; Weimer & Lenze, 1997). The aim of this study was to present the results of an experimental study on the effectiveness of intermediate student evaluations of teaching (SET) with or without collaborative consultation with a consultant (SET consultation) on professors’ lecturing skills and the student learning process (both as assessed by students). With this study, the process of possible improvement during a course on specific teaching behavior and the students’ learning process during lectures was investigated for a wide variety of university professors.

5 Tables of detailed results of Model 6 and 7 on all dependent variables are available on request.
The multilevel regression analyses showed that the intermediate feedback-only intervention had no significant effects on any of the professor and student level dependent variables. Intermediate feedback-plus-consultation was significantly effective on five variables; teaching variables *Structure*, *Validation*, *Instruction*, and *Total Instructional Skills*, and the student variable *Cognition*. In terms of Cohen’s $d$, based on the mean ratings of the professor’s, the effects on these five variables were medium (ranging from .43 to .60). We note that effect sizes calculated on the multilevel output were smaller (ranging from .26 to .36).

There were two time intervals; professors in the experimental conditions received feedback with or without consultation in between measurement occasion one and two, and in between measurement occasion two and three. When dimensions that were targeted during consultation were separated from not target dimension on each time interval we found that most effects occurred on the first time interval. On the first time interval, ratings on six out of seven teaching dimensions improved significantly ($p < .01$) when targeted. Dimensions that were not targeted did not improve. These findings indicate that improvements are due to the collaborative approach, rather than to a Hawthorne effect. On the second time interval, ratings on two targeted dimensions improved (*Instruction* and *Activation*) and ratings on one not-targeted dimension (*Validation*). We conclude that with a collaborative consultation approach, a single consultation session during a course produces desirable effects.

In general, the difference in effectiveness between feedback with or without consultation is consistent with previous findings (Cohen, 1980; Menges & Binko, 1986; Penny & Coe, 2004). Additionally, the effects of the collaborative consultation approach mostly exceed the previous experimental findings on the diagnostic approach (see Penny & Coe, 2004) and support the medium effects found on collaborative approaches used in non-experimental and post-test only studies (e.g., Dresel & Rindermann, 2011; Piccinin, Cristi & McCoy, 1999; Rindermann, Kohler & Meisenberg, 2007).

Exploratory analyses with the moderators Age and Class Size had no influence on the effects of the interventions on any of the dimensions. The effects are the same for younger as for older professors and for professors with smaller and larger classes. Exploratory analyses with the moderator *Quality of Teaching* showed that effects of the interventions partly depend on the professor’s baseline quality of teaching. Results were different between medium quality professors and high quality professors on four dependent variables; *Explication*, *Stimulation*, *Validation*, and *Total Instructional Skills*. On these variables ratings of medium quality professors did not change over time in all three conditions. Ratings of high quality professors improved significantly in the feedback-plus-consultation condition on *Validation* and *Total Instructional Skills*, compared to the control condition. Ratings of high quality
professors in the feedback-only condition significantly increased in ratings on Stimulation. At the same time, high quality professors in the control condition decrease in ratings on Explication, Stimulation, Validation and on Total Instructional Skills. Thus, the effects of the interventions on high quality professors may be viewed as helping them to increase ratings as well as to prevent a decrease in ratings over time. It should be noted that baseline ratings of high quality professors were higher in the feedback-plus-consultation condition compared to the control condition. Effects on high quality professors might therefore be somewhat biased.

In terms of the effects on high and medium quality teaching, the results go against our expectations. We expected professors with a medium quality of teaching to be more susceptible to improvement than their high quality counterparts, due to their lower baseline rating. However, we found that high quality professors benefitted more. Possible explanations are that high quality professors spend more time on their teaching or are more willing or better able to experiment with their teaching behavior within a short time frame. High quality professors tend to be highly reflective on their students’ learning process and their own teaching behavior (McAlpine & Weston, 2000), and might therefore benefit more from student feedback and a collaborative approach to consultation than regular professors do. At the same time, medium quality professors might need more time to successfully implement new teaching behavior (noticed by the students). Marsh and Roche (1993) did not found any effect of intermediate SET consultation, but they did find an interaction effect between professors’ baseline quality of teaching and improvement in ratings at the end of one semester later, indicating that professors who were initially less effective benefitted from intermediate SET consultation over a longer period of time. Also, the current intervention was relatively limited, as it did not include observations, workshops, self-ratings, video recordings, etc. Penny and Coe (2004) found large effects, in terms of Cohen’s $d$, of more extensive interventions. Longer lasting interventions were not feasible in the present study, as the courses were mostly thought once a year and lasted no longer than eight weeks. Still, as Marsh and Roche (1993) did not find effects of intermediate SET consultation, it is worth first to investigate the long term effects of the current procedure and approach to SET consultation in future research. The present findings clearly justify such research.

Professors in this study were members of the current staff of different departments and had not sought out our intervention, or any other, aimed at improving teaching skills. The present effects are therefore not dependent on the professors’ intrinsic motivation. The randomized block design furthermore ensured an equal distribution of professors from different departments and high versus medium teaching quality to each condition. The results therefore generalize to professors at this university in general.
The present study has its limitations. We note that the effects are investigated by means of students’ perceptions of teaching. Although SETs are proven to be valid and reliable in many different settings (see Marsh, 2007), other researchers on evaluation of teaching effectiveness have recommended the use of multiple sources of data to assess teaching quality (Benton & Cashin, 2012). Also, with respect to student learning outcomes, the results were based on students self-reports only. We therefore suggest future research to complement these findings from student ratings with additional measures of teaching effectiveness, such as classroom observations, and of student learning outcomes, such as course grades.

In addition to the findings in the present investigation, this study illustrates the importance of using multilevel regression analyses on student ratings data. For instance, when ignoring the professor and student level randomness in ratings over time, analyses on intermediate feedback-only resulted in significant effects on four dimensions. With these random effects (which were present according to the deviance test), significant effects were absent. It is known that ignoring random effects increases the probability of false positive well beyond the chosen alpha-level due to the underestimation of standard errors of effects (Hox, 2002). In addition, we note that a control condition was essential in this investigation to compare and take random variation in ratings into account. Finally, we note that effect sizes calculated on the multilevel output resulted in smaller effects than Cohen’s $d$ effect sizes calculated on the professors’ mean ratings. Again, this indicates the importance of taking random effects into account.

In summary, when random effects are taken into account, only intermediate student feedback in combination with collaborative consultation actually improved the quality of learning and instruction during lectures. Intermediate feedback only had no significant impact on professors in general. It mainly prevented high quality professors from a decrease in ratings over time. The exceeding effects of SET consultation generalize to professors from a wide variety of departments at this university, despite professors’ age and class size. In terms of scientific relevance, the present study illustrates the importance of using multilevel analysis on student ratings data and complements previous non-experimental findings with experimental results on this approach to consultation. In addition, it complements previous findings with results on students’ perceptions of their learning; students reported to learn more during lectures when professors were provided with intermediate SET consultation, compared to the control condition. With regard to implications for future practice, the results of this study show that mainly the first consultation renders appreciative effects, and targeting dimensions (by means of a collaborative approach to consultation) renders most effects. In short, when feedback is well timed, relevant and specific, and when consultation is collaborative and teacher-centered, these findings indicate that professors and students both benefit.
References


Experiment II – Effects on professors’ lecturing skills and student learning

Chapter 5


Appendix I

Multilevel regression models 1 to 7

Models 1 to 7 were fitted on data from all three measurement occasions, with time as level 1 variable \( t \), students as level 2 variable \( i \) and professors as level 3 variable \( j \). Student rating on dimension \( Y_{tij} \) on occasion \( t \) of student \( i \) in the class of professor \( j \) were modeled as following:

**Model 1.** Model 1 concerns the intercept-only model and comprises the following equations:

\[
\begin{align*}
\text{Level 1:} & \quad Y_{tij} = \beta_{0ij} + e_{tij}, \\
\text{Level 2:} & \quad \beta_{0ij} = \beta_{00j} + u_{0ij}, \\
\text{Level 3:} & \quad \beta_{00j} = \gamma_{000} + v_{00j}.
\end{align*}
\]

By substitution, we obtain the single-equation:

\[
Y_{tij} = \gamma_{000} + v_{00j} + u_{0ij} + e_{tij}.
\]

Here the student rating on dimension \( Y_{tij} \) on occasion \( t \) of student \( i \) in the class of professor \( j \) is modeled by the intercept \( \beta_{0ij} \) and a residual error term \( e_{tij} \). In the second and third level equations (1.2 and 1.3) the intercept \( \beta_{0ij} \) is decomposed by a residual error term for students \( u_{0ij} \) (random intercept on student level), a residual error term for professors \( v_{00j} \) (random intercept on professor level) and a fixed component \( \gamma_{000} \) (the overall mean).

The variances of the three residual error terms are denoted by

\[
\begin{align*}
\text{var} (e_{tij}) &= \sigma^2, \\
\text{var} (u_{0ij}) &= \tau^2_0, \\
\text{var} (v_{00j}) &= \phi^2_0.
\end{align*}
\]

These represent the variance of the ratings over time of a given student \( i \) (\( \sigma^2 \)), the variance of the ratings over students of a given professor \( j \) (\( \tau^2_0 \)), and the variance of the ratings over professors (\( \phi^2_0 \)).

**Model 2.** Model 2 contains the fixed effects of Time (coded 0, 1 and 2) and the conditions and an intercept random over professors and students. Condition is coded in the feedback-only condition versus the control condition dummy variable (\( \Delta F \)) and the feedback-plus-consultation condition versus the control condition dummy variable (\( \Delta FC \)). Model 2 is defined by the equations:

\[
\begin{align*}
\text{Level 1:} & \quad Y_{tij} = \beta_{0ij} + \beta_{ij} \times \text{Time}_{ij} + e_{tij}, \\
\text{Level 2:} & \quad \beta_{0ij} = \beta_{00j} + u_{0ij},
\end{align*}
\]
Experiment II – Effects on professors’ lecturing skills and student learning

Chapter 5

Level 2: \( \beta_{ij} = \beta_{10j} \), \hspace{1cm} (2.3)

Level 3: \( \beta_{00j} = \gamma_{000} + \gamma_{001} \Delta F_j + \gamma_{002} \Delta FC_j + \nu_{00j} \), \hspace{1cm} (2.4)

Level 3: \( \beta_{10j} = \gamma_{100} \). \hspace{1cm} (2.5)

By substitution, we obtain the single-equation:

\[
Y_{tij} = \gamma_{000} + \gamma_{100} Time_{ij} + \gamma_{001} \Delta F_j + \gamma_{002} \Delta FC_j + \nu_{00j} + u_{0ij} + e_{tij} .
\] \hspace{1cm} (2.6)

In 2.2, the intercept \( \beta_{0ij} \) is the sum of the student-level residual \( u_{0ij} \) (random intercept on student level) and a random teacher intercept \( \left( \beta_{00j} \right) \). In equation 2.4, the random teacher intercept is the sum of a fixed effect \( \gamma_{000} \) and the random teacher value \( \nu_{00j} \). The fixed component \( \gamma_{000} \) represents the overall average intercept coefficient for the control condition. The fixed component \( \gamma_{100} \) represents the overall average regression coefficient for \( Time \) (mean slope). The fixed components \( \gamma_{001} \) and \( \gamma_{002} \) represent the main effect of the conditions versus the control condition.

Model 3. In Model 3 we added the Time*Condition interaction effects \( Time \ast \Delta F \) and \( Time \ast \Delta FC \). Model 3 is defined in equations:

Level 1: \( Y_{tij} = \beta_{0ij} + \beta_{1ij} Time_{ij} + e_{tij} \), \hspace{1cm} (3.1)

Level 2: \( \beta_{0ij} = \beta_{00j} + u_{0ij} \), \hspace{1cm} (3.2)

Level 2: \( \beta_{1ij} = \beta_{10j} \), \hspace{1cm} (3.3)

Level 3: \( \beta_{00j} = \gamma_{000} + \gamma_{001} \Delta F_j + \gamma_{002} \Delta FC_j + \nu_{00j} \), \hspace{1cm} (3.4)

Level 3: \( \beta_{10j} = \gamma_{100} + \gamma_{101} \Delta F_j + \gamma_{102} \Delta FC_j \). \hspace{1cm} (3.5)

By substitution, we obtain the single-equation:

\[
Y_{tij} = \gamma_{000} + \gamma_{100} Time_{ij} + \gamma_{001} \Delta F_j + \gamma_{002} \Delta FC_j + \gamma_{101} Time_{ij} \ast \Delta F_j + \gamma_{102} Time_{ij} \ast \Delta FC_j + \nu_{00j} + u_{0ij} + e_{tij} .
\] \hspace{1cm} (3.6)

The parameters are the same as in Model 2. The parameters of \( Time \ast \Delta F \left( \gamma_{101} \right) \) and \( Time \ast \Delta FC \left( \gamma_{102} \right) \) represent the effects of the interventions.

Model 4. In Model 4 we allowed the slope of the ratings over time to be random for the professor and student level. Model 4 is defined through the equations:

Level 1: \( Y_{tij} = \beta_{0ij} + \beta_{1ij} Time_{ij} + e_{tij} \), \hspace{1cm} (4.1)

Level 2: \( \beta_{0ij} = \beta_{00j} + u_{0ij} \), \hspace{1cm} (4.2)
Level 2: \[ \beta_{1ij} = \beta_{10j} + u_{1ij}, \quad (4.3) \]
Level 3: \[ \beta_{00j} = \gamma_{000} + \gamma_{001}\Delta F_j + \gamma_{002}\Delta FC_{j0} + \nu_{00j}, \quad (4.4) \]
Level 3: \[ \beta_{10j} = \gamma_{100} + \gamma_{101}\Delta F_j + \gamma_{102}\Delta FC_{j0} + \nu_{10j}. \quad (4.5) \]

By substitution, we obtain the single-equation:

\[
Y_{tij} = \gamma_{000} + \gamma_{100} Time_{ij} + \gamma_{001}\Delta F_j + \gamma_{002}\Delta FC_{j0} + \\
\gamma_{101} Time_{ij}\cdot\Delta F_j + \gamma_{102} Time_{ij}\cdot\Delta FC_{j0} + u_{1ij} + \nu_{00j} + u_{1ij} + e_{ij}. \quad (4.6)
\]

Again the intercept \( \beta_{0ij} \) is allowed to be random over students and professors by including the random components \( u_{0ij} \) and \( \nu_{0ij} \). In addition, the regression parameter \( \beta_{1ij} \) for Time is allowed to be random over students and professors by including the random effects \( u_{1ij} \) and \( \nu_{10j} \).

The slope variances are denoted by

\[
\text{var}(u_{1ij}) = \tau_1^2, \quad \text{var}(\nu_{10j}) = \varphi_1^2. \quad (4.7)
\]

The intercept-slope covariances are denoted by

\[
\text{cov}(u_{0ij}, u_{1ij}) = \tau_{01}, \quad \text{cov}(\nu_{00j}, \nu_{10j}) = \varphi_{01}. \quad (4.8)
\]

**Model 5.** In Model 5, for each dimension on each specific time interval, we split up the feedback-plus consultation condition in a group that targeted the dimension for improvement (Target) and a group that did not (No Target). Condition was therefore recoded into the dummy variables Control_versus_Feedback-only (denoted as \( \Delta F \)), Control_versus_Feedback-plus-Consultation_No Target (denoted as \( \Delta FC_{\text{NoTarget}} \)) and Control_versus_Feedback-plus-Consultation_Target (denoted as \( \Delta FC_{\text{Target}} \)). Time was recoded for the specific time interval (in case of time interval \( T_1T_2; T_1 = 0 \) and \( T_2 = 1 \) and in case of time interval \( T_2T_3; T_2 = 0 \) and \( T_3 = 1 \)). The random effects were limited to the intercept in this model. Model 5 is defined through the equations:

Level 1: \[ Y_{tij} = \beta_{0ij} + \beta_{1ij} Time_{ij} + e_{ij}, \quad (5.1) \]
Level 2: \[ \beta_{0ij} = \beta_{00j} + u_{0ij}, \quad (5.2) \]
Level 2: \[ \beta_{1ij} = \beta_{10j}, \quad (5.3) \]
Level 3: \[ \beta_{00j} = \gamma_{000} + \gamma_{001}\Delta F_j + \gamma_{002}\Delta FC_{\text{NoTarget}j} + \\
\gamma_{003}\Delta FC_{\text{Target}j} + \nu_{00j} \quad (5.4) \]
Level 3: \[ \beta_{10j} = \gamma_{100} + \gamma_{101}\Delta F_j + \gamma_{102}\Delta FC_{\text{NoTarget}j} + \\
\gamma_{103}\Delta FC_{\text{Target}j} \quad (5.5) \]
By substitution, we obtain the single-equation:

\[
Y_{ij} = \gamma_{000} + \gamma_{100} Time_{ij} + \gamma_{001} \Delta F_j + \\
\gamma_{002} \Delta FC_{NoTarget} + \gamma_{003} \Delta FC_{Target} + \gamma_{101} Time_{ij} \Delta F_j + \\
\gamma_{102} Time_{ij} \Delta FC_{NoTarget} + \gamma_{103} Time_{ij} \Delta FC_{Target} + \\
v_{00j} + u_{0ij} + e_{ij}.
\] (5.6)

The parameters are the same as in Model 2. The fixed component \(\gamma_{100}\) represents the overall average regression coefficient for \(Time\) (mean slope) on the specific time interval. The fixed component \(\gamma_{001}\) represents the main effect of feedback versus the control condition. The fixed components \(\gamma_{002}\) and \(\gamma_{003}\) represent the consultation’s specific main effects of non-targeted dimensions versus the control condition and targeted dimensions versus the control condition.

**Model 6.** In Model 6 we modeled the main effects of three professor-level moderators; \(Age\), \(Quality\ of\ Teaching\), and \(Class\ Size\). Let \(M\) denote a moderator of interest. Its introduction requires the expansion of equation 4.4 and 4.5 as follows:

\[
\beta_{00j} = \gamma_{000} + \gamma_{001} \Delta F_j + \gamma_{002} \Delta FC_j + \gamma_{003} M_j + v_{00j}, \quad (6.1)
\]

\[
\beta_{10j} = \gamma_{100} + \gamma_{101} \Delta F_j + \gamma_{102} \Delta FC_j + \gamma_{103} M_j + v_{10j}. \quad (6.2)
\]

**Model 7.** In Model 7 we added parameters for the moderating interaction effects of \(M^{*}\Delta F\) (\(\gamma_{004}\)), \(M^{*}\Delta FC\) (\(\gamma_{005}\)), \(M^{*}Time\) (\(\gamma_{103}\)), \(M^{*}Time^{*}\Delta F\) (\(\gamma_{104}\)), and \(M^{*}Time^{*}\Delta FC\) (\(\gamma_{105}\)). The interaction effects \(M^{*}Time^{*}\Delta F\) and \(M^{*}Time^{*}\Delta FC\), represent the separate effects of the two interventions for professors with high and low ratings on the specific moderator, compared to the control condition. Equation 6.1 and 6.2 are expanded as follows:

\[
\beta_{00j} = \gamma_{000} + \gamma_{001} \Delta F_j + \gamma_{002} \Delta FC_j + \gamma_{003} M_j + \gamma_{004} \Delta F_j M_j + \\
\gamma_{005} \Delta FC_j M_j + v_{00j}, \quad (7.1)
\]

\[
\beta_{10j} = \gamma_{100} + \gamma_{101} \Delta F_j + \gamma_{102} \Delta FC_j + \gamma_{103} M_j + \gamma_{104} \Delta F_j M_j + \\
\gamma_{105} \Delta FC_j M_j + v_{10j}. \quad (7.2)
\]