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

Citation for published version (APA):
Knol, M. H. (2013). Improving university lectures with feedback and consultation

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 I (pilot study):

Effects of intermediate student feedback and collaborative consultation on university professors’ lecturing skills

This chapter is an adapted version of:
Knol, M.H., In ’t Veld, R., Vorst, H.C.M., Van Driel, J.H., & Mellenbergh, G.J. (accepted). Experimental Effects of Student Evaluations Coupled with Collaborative Consultation on University Professors’ Instructional Skills, Research in Higher Education.
Abstract

This experimental study concerned the effects of repeated students’ evaluations of teaching coupled with collaborative consultation on professors’ instructional skills. Twenty-five psychology professors from a Dutch university were randomly assigned to either a control condition or an experimental condition. During their course, students evaluated them four times immediately after a lecture (class meeting in which lecturing was the teaching format) by completing the Instructional Skills Questionnaire (ISQ). Within two or three days after each rated lecture, the professors in the experimental condition were informed of the ISQ-results and received consultation. Each consultation, three in total, resulted in a plan to improve their teaching for the next lectures. Controls received neither their ISQ-results nor consultation during their course. Multilevel regression analyses showed significant differences in ISQ-ratings in the experimental condition compared to the control condition, specifically on the instructional dimensions Explication, Comprehension and Activation. In addition, the impact of each of the three consultations plus differences between targeted versus non targeted dimensions were analyzed. This study complements recent non-experimental research on a collaborative consultation approach with experimental results in order to provide evidence-based guidelines for faculty development practices.
Introduction

Interventions employed to improve university teaching include (1) workshops, seminars and programs; (2) consultation; (3) instructional improvement grants; (4) resources such as newsletters, manuals or sourcebooks; and (5) colleagues helping colleagues (Weimer & Lenze, 1997). Previous reviews of the effects of these interventions all stress the importance of more research in this field employing more rigorous designs (Levinson-Rose & Menges, 1981; Prebble et al., 2004; Stess, Min-Leliveld, Gijbels, & Van Petegem, 2010; Weimer & Lenze, 1997). The aim of the present study is therefore to present an experimental study, in which we focus specifically on the effects of individual consultation based on student evaluation of teaching (SET consultation). Individual peer or expert consultation is one of the most commonly used interventions in faculty development and support (Knapper & Piccinin, 1999; Penny & Coe, 2004; Prebble et al., 2004). Based on the available research, Lenze (1996) identified consultation as an instructional development strategy preferable to the other approaches stated above. It has shown to increase considerably the impact of student ratings on teaching practices (Menges & Brinko, 1986; Penny & Coe, 2004; Weimer & Lenze, 1997). On the other hand, the variation in effect size and in actual procedure (implementation) is large (Cohen, 1980; Menges & Brinko, 1986; Penny & Coe, 2004). As SET consultation is a widely used, but relatively expensive intervention, we need to know more about the effects of specific models and procedures in order to guide current faculty development practices in the field.

In terms of consultation models, we investigated a collaborative approach to consultation on the instructional skills of professors at a Dutch university. In terms of consultation procedures, we studied the effects of SET consultation during a course, instead of at the end of the course, to assess the role of timing of the feedback and intervention. Students in this study rated four specific lectures (class meetings in which lecturing was the teaching format) during the course, to improve the specificity, comparability, and quality of the feedback. We studied the specific effects of one intermediate consultation and the additional effects of a second and third intermediate consultation on student ratings. We investigated these issues in a two group experimental design and used multi-level regression analyses to take into account random effects. The professors at the time of their participation had not approached by, or were not involved with, a teacher-training center with the aim of improving their teaching. This implies that the effects were investigated among professors in general, rather than professors, who were particularly motivated to change. First we provide an overview of previous research on the effects of SETs and SET consultation and a theoretical framework for a collaborative approach to consultation.
Research on SETs and SET consultation

At many universities, course evaluations based on students' evaluations of teaching are common practice. One of the main purposes of collecting SETs is to provide professors with feedback so they can improve their teaching practices. As with all learning processes, feedback is considered one of the most powerful tools to achieve progress (Hattie & Timperley, 2007). SETs provide a unique perspective on teaching practices, and have proven to be valid and reliable in many different settings (Marsh, 2007b).

Despite the effort that goes into collecting SETs for every course at the end of every term, SETs do not often improve teaching practice (Kember, Leung & Kwan, 2002). 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, as judged by their student ratings. This implies that simply collecting student ratings does not automatically help to improve the quality of teaching (see also Richardson, 2005).

The lack of a positive effect of course evaluation systems on teaching practices may be understood in the light of the basic rules of effective feedback. Specifically, feedback should be well-timed, specific, reliable, and should concern changeable behavior (McLaughlin & Pfeifer, 1988). Evaluations, provided at the end of term, are arguably ill timed, as they do not provide professors with an immediate opportunity to benefit from this feedback. Furthermore, course evaluations often contain mainly unspecific items (e.g., “rate your professor”), which do not provide concrete feedback and serve merely as a general monitor of teaching quality.

Still, well-timed qualitative feedback is insufficient. Multiple studies and meta-analysis have shown more improvement with mid-term feedback compared to end-of-the-term feedback on student ratings, but the effects are small according to both short and long term analyses (Cohen, 1980; Lang & Kersting, 2007; Menges & Brinko, 1986). Besides the quality and timing of the feedback, the fundamental validity issue in student evaluations concerns the interpretation and use of the data (McKeachie, 1997). Theall and Franklin (2001) found that faculty often misinterpret, misunderstand, or misuse SETs, and that consequently SETs seldom contribute to actual improvements in teaching. When SETs are augmented with consultation (SET consultation), the effects are considerably larger compared to (mid-term) feedback alone (Cohen, 1980; Menges & Brinko, 1986; Penny & Coe, 2004). According to Penny and Coe's (2004) meta-analysis based on 11 studies, SET consultation resulted in a weighted mean effect size of .69. Menges and Brinko (1986) found an even larger average effect size of 1.1.

However, the variation in effects and procedures is large. The confidence interval in the study by Penny and Coe ranged from .43 to .95, which suggests considerable variation
in the effectiveness of SET consultation. Furthermore, results of Marsh and Roche’s study (1993), who found no effect of mid-term SET consultation, were excluded. The range in effectiveness was also noted by Menges and Brinko (1986). They found effect sizes ranging from 0 to 2.5.

Penny and Coe (2004) attempted to identify the factors that contributed to successful SET consultation. They could not find clear, statistically significant differences amongst consultation approaches, due to the small number of experimental studies available. They concluded: “... the most robust finding may be that more research is needed.” (p. 236). “Considerably more research on the effects of consultative feedback in settings other than North America is sorely needed. The sample for our meta-analysis was too small to provide adequate statistical power to demonstrate clearly the effectiveness of consultation or to identify moderator variables. ..... Although our review uncovered some strategies that may be important for consultative feedback, there is need for research that directly assesses the effects of these strategies”. (p. 248)

Two additional issues underline the importance of more experimental research in this field. First, the large majority of studies in Penny and Coe’s meta-analysis were published in the period 1975-1986. Exceptions are Marsh and Roche’s study in 1993 and one study by Hampton and Reiser, which was published in 2004. The results of these two studies on the effects of SET consultation are inconsistent. Hampton and Reiser did find effects of mid-term consultation, while Marsh and Roche only found effects of end-of-the-term consultation and not for mid-term consultation. Additional experimental research is required to identify the specific effects of specific procedures, and ultimately to formulate guidelines for faculty development practices in current university settings.

Second, most studies in the 1975-1986 period could not take the multilevel structure of the data into account by means of multilevel regression modeling. The current statistical software now allows us to analyze student ratings data while taking into account variance on different levels. The failure to take into account the multilevel structure often results in incorrect conclusions (Snijders & Bosker, 1999).

In the past decade, a few non-experimental studies on SET consultation practices were published (e.g., Rindermann, Kohler & Meisenberg, 2007; Dresel & Rindermann, 2011; Piccinin, Cristi & McCoy, 1999). These studies showed positive results, but varied in procedures. Rindermann, Kohler and Meisenberg (2007) used one mid-term collaborative consult in a private school for speech therapy, and found a medium effect on their total instructor scale with all faculty (16 in total) included in the analyses (without the three best faculty they found a large effect). Piccinin, Cristi and McCoy (1999) used data from
Experiment I – Effects on professors’ lecturing skills

Chapter 3

54

participants, who had approached a teacher training centre to improve their teaching. Faculty members were assigned to three different interventions based on their needs; SET consultation, SET consultation plus observation, and SET consultation plus observation and student consultation. Results showed positive, but different, patterns of increase in ratings over time in each group. Dresel and Rindermann (2011) provided 12 German faculty members (teaching 98 courses over a period of 2 years) with SET consultation in the first year, and found moderate to large effects. In this study, however the intervention lasted a full day. The study demonstrated the generalizability of SET consultation effects to other courses. Using multilevel analyses they controlled for potential bias and unfairness variables on the professor, course and student level.

The results of these recent non-experimental studies are important, because they have good external validity, they address biasing variables and effects in non-English speaking countries for the first time. Dresel and Rindermann (2011) also illustrated the importance of using appropriate multilevel procedures. Whether non-experimental results are due to the intervention remains an open question. Non-experimental designs suffer from potential selection bias and are often open to alternative explanations of the results (like the Hawthorne effect). Dresel and Rindermann (2011) underline the difficulty of conducting research, which is both internally and externally valid. Therefore, we stress that different studies with different characteristics need to be conducted. With this study we aim to augment recent non-experimental studies with up-to-date experimental results. In addition, we aim to provide more knowledge on the effects of amount of consultation.

Theoretical framework for collaborative consultation

Penny and Coe (2004) defined instructional consultation as a structured, collaborative, problem-solving process that uses information about teaching performance as a basis for discussion about improving teaching practice. They concluded from the literature that consultation for teaching improvement should be voluntary, individualized, confidential, reflective, and carried out for formative purposes, not for summative evaluation.

Penny and Coe categorized experimental studies by the approach to consultation used. They distinguished a diagnostic ($N = 2$), an advisory ($N = 6$) and an educational ($N = 3$) approach. They found a medium effect size for the first approach and larger effect sizes for the last two, which involved more extensive interventions. Specifically, they included at least one other source of information on teaching behavior, e.g., observation or videotaping, and/or additional educational activities, such as seminars and workshops.
When it comes to approaches to consultation, the two most common models are the prescriptive model and the collaborative model (Brinko, 1990). In the prescriptive model, the consultant identifies, diagnoses, and solves problems. In a collaborative approach the consultant serves as a facilitator, by encouraging faculty to reflect on the current situation, teaching effectiveness, and possible alternative teaching strategies to achieve his or her goals. The diagnostic approach, as described by Penny and Coe, is a prescriptive approach.

Psychological theories on behavior and behavioral change help identify conditions for effective consultation. Here we focus specifically on aspects of Eagly and Chaiken’s attitude-behavior theory (1998), which includes the theory of reasoned action (Aizen & Fishbein, 1980), theory of planned behavior (Ajzen, 1991), and theories on self-efficacy (Bandura, 1977, 1997). Empirical support for these theories is reported in several studies (Madden, Ellen, & Ajzen, 1992; Sheppard, Hartwick, & Warshaw, 1988; Van den Putte, 1993). Figure 3.1 depicts a combination of these theories, including the most relevant variables for our present purposes.

We consider these theories as they provide us with a frame of reference on consultation from a teacher-centered perspective. According to these theories, behavior starts with an intention. Intentions relating to (new) behavior are based on the professor's personal attitude (either positive or negative), and the professor's self-efficacy concerning this (new) behavior. A positive or negative attitude depends on the expected outcomes of this behavior, and on the value of these expected outcomes. Finally, these values depend on personal goals (Eagly & Chaiken, 1998).

By considering the attitude-behavioral theories in the teaching context, we identified the conditions that should be met in the consultation process to facilitate effective change in teaching practice based on SETs. According to these theories, planned improvement in teaching is more likely to be successful if the professor’s expected outcomes are consistent with the professor’s values and instructional goals, and if the professor feels sufficiently

![Figure 3.1](image_url)  
**Figure 3.1** Combination of attitude-behavior models, conditions for behavioral change.
confident to achieve the desired outcomes. For example, given a low rating on 'engaging students during the lecture', a professor will choose to interact with the students if he or she believes that this will indeed result in a more active participation of the students, and if the professor values this outcome in the process of teaching. In addition, the professor should feel confident enough about actually engaging his or her students in this fashion. On the other hand, a professor may view interaction with students during the lecture as a diversion that serves no purpose given his or her teaching objectives. Based on these theories, we postulate that consultation is effective if

a) the professor's current behavior (according to student ratings) is explored,
b) the professor's values and goals are explored,
c) discrepancies between the current effects of the professor's behavior and his or her instructional values and goals form the starting point for discussion and plans for improvement,
d) plans for behavioral change are considered important by the professor, and
e) the professor feels confident about executing the plans for improvement.

This implies that, in order to achieve lasting behavioral change, the professor should accept the outcome of each step of SET consultation, i.e., 1) the interpretation of the student ratings, 2) the selection of specific ratings to improve, 3) the diagnosis and analysis on the cause of selected low ratings, 4) possible strategies for improvement, and 5) the selection of final strategies. In the present study, we used a teacher-centered consultation protocol, in which each step of the consultation started with the professor's opinion and ended with the professor's conclusion. The protocol is consistent with a collaborative approach to consultation.

In summary this study addresses the following research questions:

1. What are the effects of a collaborative approach to SET consultation, provided during a course, on seven specific dimensions of college professors' lecturing skills?
2. What are the effects of one consultation, and the additional effects of a second and third consultation during a course on college professors' lecturing skills?
3. If effects occur, is there a difference in effect between dimensions which are targeted for improvement during consultation and dimensions which are not targeted, which would imply that the effects can be related to either the feedback or the consultation?
Context information

The current experiment was conducted at a Dutch University. In the Netherlands, bachelor programs generally take three years and are focused on a specific field of study from day one (no general college courses are taught). This study concerned the third (final) year of the bachelor program of psychology. At this university, students select a specific field of interest within psychology (e.g., social psychology, clinical psychology, etc.) in their third year. Courses take eight weeks (varying in workload) and students attend six to eight courses per semester. Regular SETs are anonymous and collected at the end of each course (most often during the exam). The results are used as feedback for professors, coordinators, management and quality control committees. In order to get tenure, it is mandatory to attain a teaching certificate. Obtaining positive SETs is one of the criteria for attaining such a certificate.

Method

Participants

Professors

At the department of psychology of a Dutch university, we selected 27 eight-week courses from the second semester of the third and final bachelor year (the equivalent of the senior year). Each course included one weekly lecture (class meeting in which lecturing was the teaching format), with a minimum of four lectures given by the same professor (27 professors in total). A standard lecture at this university takes 90 minutes with a 15-minute break after 45 minutes. Therefore course level and teaching format were comparable. The selected courses were all designed for psychology majors. Of the 27 professors, 25 agreed to participate (14 males, 11 females), and were randomly assigned to the conditions. All participating professors had a PhD, and were full-time ranked assistant, associate, or full professors.

Students

The students in the participating courses completed the Instructional Skills Questionnaire (ISQ), the instrument used to rate the professors on four measurement occasions. In total the ISQ was completed 1,954 times, with 1,225 forms containing a student ID number. There were 604 unique professor - student ID combinations. The 729 forms with a missing student ID number were given a unique ID number, which resulted in a total of 1,333 unique professor-
student combinations. Some students attended more than one course and therefore rated both professors in the control and experimental condition (N students attaining one course = 342, two courses = 90, three courses = 23, four courses = 2, and five courses = 1, according to the available student ID numbers). Since students did not know that the professors were participating in an experiment, this was not expected to be of any influence.

The mean class size was 19.7 students (median = 14, SD = 13.08, min = 6, max = 62). The mean class sizes in the control condition and experimental condition were 13.8 (median = 11.5, SD = 7.5, min = 6, max = 42) and 25.7 (median = 25, SD = 14.8, min = 8, max = 62), respectively.

Consultants

The first and second authors served as consultants to assure full insight in the practical and procedural aspects of the feedback and consultation protocol. The collaborative consultation approach, as defined by Brinko (1990), was adapted for this study. According to Brinko, collaborative consultants function as partners: they encourage their clients to identify, diagnose, and provide solutions to the issues they raise. The consultants were trained in coaching- and social skills, including creating a safe learning environment, structured consultation, encouraging reflection, and formulation of concrete plans.

Design

The experiment was a randomized two-group design with four repeated measures. Twenty-five professors were randomly assigned to either the control condition (N = 13) or the experimental condition (N = 12). In both conditions, students evaluated four 90-minute lectures of their professors during the course with the ISQ, at four measurement occasions. The ISQ measures seven dimensions of the professors’ lecturing skills, specifically Structure, Explication, Stimulation, Validation, Instruction, Comprehension, and Activation. Professors in the experimental condition received consultation based on their ISQ ratings within two or three days after each rated lecture. In total, three consultation sessions took place between the ratings. Professors in the control condition received all ISQ ratings at the end of the course.

Dependent variables

Feedback should be specific, multidimensional, reliable, and concern changeable behavior. In terms of specificity, we used the ISQ, which contained items that covered the seven
dimensions relating to the professor’s lecturing behavior. These are based on previous practice and research on student evaluation. Marsh and colleagues developed a well studied course evaluation instrument (SEEQ), containing nine dimensions; Instructors Enthusiasm, Organization/Clarity, Group Interaction, Individual Rapport, Breadth of Coverage, Learning/Value, Examination/Grading, Assignments/Readings, Workload/Difficulty (Marsh, 1984, 1987; Marsh & Hocevar, 1991b). The first five dimensions mentioned concern the teaching behavior. De Neve and Janssen (1982) developed a questionnaire for evaluation of lectures (Evalec) containing five specific dimensions: Validating, Stimulating, Interacting, Directing, and Structuring behavior. Based on the literature and on these two instruments, Vorst and Van Engelenburg (1994) developed a course evaluation instrument (UvAlon) for a Dutch university containing the general dimensions (Learning/Value, Entry Level, Time Invested/Workload, Difficulty, Literature, and Examination) and seven specific dimensions for teaching behavior (Structure, Explication, Stimulation, Validation, Instruction, Conversation and Interaction). The psychometric quality of this instrument was investigated and confirmed in different studies (Vorst & Van Engelenburg, 1994; SCO Kohnstamm Institute, 2002, 2005). We adapted this instrument to a one-lecture instrument with a selection of the twenty-eight specific questions on instructional behavior. This resulted in the Instructional Skills Questionnaire (ISQ), which comprises 28 items, and measures the seven dimensions identified by Vorst and Van Engelenburg. We renamed the dimensions Conversation and Interaction to Comprehension and Activation, respectively, to more accurately convey their meaning. Each of the seven dimensions is measured with four items, two positive (indicative) and two negative (contra-indicative) worded items. The response format of the ISQ items is a 5-point Likert-scale. In more detail, the seven ISQ dimensions are:

1. **Structure**: the extent to which the subject matter is handled systematically and in an orderly way. Example item: *The subject matter is presented incoherently* (contra-indicative).

2. **Explication**: the extent to which the professor explains the subject matter, especially the more complex topics. Example item: *The instructor gives clarifying examples*.

3. **Stimulation**: the extent to which the professor interests students for the subject matter. Example item: *The lecture is boring* (contra-indicative).

4. **Validation**: the extent to which the professor stresses the benefits and the relevance of the subject matter for educational goals or future occupation. Example item: *The instructor shows why the subject matter is important*. 
5. **Instruction**: the extent to which the professor provides instructions about how to study the subject matter. Example item: *It is clear what the instructor requires of me.*

6. **Comprehension**: the extent to which the professor creates opportunities for questions and remarks regarding the subject matter. Example item: *The instructor provides insufficient occasion to ask questions.*

7. **Activation**: the extent to which the professor encourages students to think about and work with the subject matter. Example item: *Students are encouraged to think along during the lecture.*

The dimension score is the student’s mean of the four specific dimension items. The total mean score is used as an estimate of *Total Instructional Skills.*

Missing item responses (.3%) were imputed with the student’s mean of the other three items of that specific dimension. Correlations between the seven domains range from .10 to .66, with a median of .30. Cronbach’s alpha on all professors mean scores on the first measurement occasion range from .66 to .95 with a mean of .82.

**Procedure**

All psychology professors at this Dutch university, who taught an undergraduate third year course, designed for psychology majors, in the second semester with at least four lectures (i.e., class meetings in which lecturing was the teaching format), were invited to voluntarily participate in the study. Professors first received an email with project information, and a request to meet with one of the researchers. At the subsequent meeting, the researcher then explained the procedure. Professors, who agreed to participate, were randomly assigned to either the control condition or experimental condition. There was no open access to previous SETs of the professors. Therefore, the director of the undergraduate school of the psychology department confirmed that the distribution of teaching quality was comparable in the two conditions based on previous SETs of the professors. A multilevel t-test on baseline mean ratings on *Total Instructional Skills*, measured by the ISQ on the first evaluation occasion, confirmed no differences in teaching quality between professors in the control condition and experimental condition. Then, prior to their course, all professors received procedural instructions by email.

During the first lecture of each course, one of the researchers invited the students to participate in a research project by completing the ISQ four times during the course (at the end of the first, third, fifth and seventh lecture). Students were instructed to focus on the
current lecture, while completing the questionnaire. They were asked to provide their student ID number. The researcher guaranteed that professors would only receive anonymous ratings. In addition, students were assured of anonymity in their evaluations with a similar statement on the ISQ form. To further ensure anonymity, not the professor but one of the students in each course collected the completed ISQ forms on each evaluation occasion and brought them to the researchers. The students did not know that their professors were participating in a randomized experiment involving SET consultation.

**Control condition**

Professors in the control condition received the ISQ ratings pertaining to all four measurement occasions at the end of their course. The procedure followed with the students was the same as for the experimental condition.

**Experimental condition**

In the experimental condition (i.e., SET consultation condition), each professor met with a consultant after every rated lecture, for a total of three consultations, to discuss the ISQ ratings. The professor and the consultant also met prior to the study (the introductory meeting) and after the final lecture for a final evaluation.

*Introductory meeting*. The introduction allowed the consultant and professor to get acquainted. The consultant explained the project, and the procedure of feedback and consultation, and topics, such as the collaborative consultation approach, additional responsibilities of the professor, and conditions that were deemed necessary for effective feedback (e.g., commitment in terms of time and motivation to achieve improvement).

*Consultation*. The consultation protocol was based on the collaborative approach. The consultant was responsible for the consultation process, by following each step of the protocol. The consultant’s role was to facilitate behavioral change. The professor had ultimate control over the content of the consultation, i.e., the professor determined which items of the student ratings questionnaire were addressed, the formulation of areas of improvement, and action plan. Still, consultants were free to be directive at any stages of the consultation, e.g., by providing alternative interpretations of student ratings, alternative views when exploring problems in teaching effectiveness, and alternative strategies for improvement. Nevertheless, according to the theory of planned behavior (Ajzen, 1991), it is important that the professors recognize and identify with the newly formulated views on teaching and plans for improvement. Therefore we set out to organize the professor-consultant-professor approach such that every step of the consultation started and ended with the professor’s opinion and conclusions.
Consultation 1. The consultations involved a five-step procedure. The steps are 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.

The consultation started with discussing how the professor experienced the lecture. Then the consultant explained the different ISQ dimensions. The consultant provided the professor with a profile based on the mean ratings on every dimension, and on the specific item scores. The consultant assisted with the interpretation of the results. The professor then undertook to link the results to his or her own experience and goals during the class. Results that were in any way surprising, unexpected, or unsatisfactory to the professor were discussed. The professor then selected the questionnaire items that he or she identified as being open to improvement. The consultant encouraged further reflection on the selected questionnaire items, and discussed (new) set goals, line of thought, possible internal conflicts, and practical difficulties. Once the desired goals, and current problems in achieving these goals were explicated, the consultant encouraged the professor to think about possible plans for improvement. If necessary, the consultant also provided suggestions. Eventually the professor decided on the final concrete plan of action. Finally, the consultant asked whether the professor had enough time to prepare, and whether he or she considered the plans to be sufficiently realistic, feasible, and relevant to pursue.

Consultations 2 and 3. Consultations 2 and 3 followed the same procedure as consultation 1, except they started out with discussing previous plans. At the beginning of consultation 2 and 3, the professor reported on his or her experiences in implementing the previously made plans. The consultant encouraged the professor to reflect on reasons for success or failure.

Final session. In the final session, the professor and consultant again discussed the previous lecture and the results of the final student ratings. They finished the consultation with an evaluation of the program and plans for the future.

Statistical analyses

The data analysis required multilevel regression modeling, because measurement occasions are nested within students, and students are nested within professors (Snijders & Bosker, 1999). Also, we wanted to investigate differences in ratings between professors whilst taking variation between students into account. Finally, individual professors and students might vary in ratings at the first measurement occasion (intercept variance) and in their increase or decrease of ratings over time (slope variance). With multilevel regression analyses we are
able to accommodate random intercept and slope variation while analyzing the fixed (mean) effects of the intervention.

The effects of SET consultation were analyzed on each of the seven dimensions (Structure, Explication, Stimulation, Validation, Instruction, Comprehension, Activation) and on the Total Instructional Skills score. We analyzed each of these dependent variables, by fitting multilevel models to the data with time as level 1 variable ($t$), students as level 2 variable ($i$) and professors as level 3 variable ($j$). With the first model, we analyzed the intra-class correlation for the professor level and the student level, the proportion of the total variance that is due to differences between professors and due to differences between students. The second and third models were used to analyze whether the slope for the professor and student level was indeed random. If this was the case, we needed to take this random slope variance into account in analyzing the effects of the intervention by adding it to the next model. The fourth model was used to analyze the effect of SET consultation. With the fifth model we analyzed the specific effect of the first consultation and the additional effects of the second and third consultation on the dependent variable Total Instructional Skills. If the fifth model showed an effect of SET consultation on a specific time interval ($T_1T_2$, $T_2T_3$ and/or $T_3T_4$), the sixth model was used to analyze specific effects of targeted interventions versus non-targeted interventions on this specific time interval for each dimension.

**Intra-class correlation**

The first model, the intercept-only model, contained a random intercept for professors and students (Model 1). Model 1 is defined through the equations:

Level 1:  
$$ Y_{ij} = \beta_{0ij} + e_{ij} \quad (1.1) $$

Level 2:  
$$ \beta_{0ij} = \beta_{00j} + u_{0ij} \quad (1.2) $$

Level 3:  
$$ \beta_{00j} = \gamma_{000} + \nu_{00j} \quad (1.3) $$

Here the student rating on dimension $Y_{ij}$ on occasion $t$ of student $i$ in the class of professor $j$ is modeled by the intercept $\beta_{0ij}$ and a residual error term $e_{ij}$. In this model the intercept varied between students and between professors. Thus, in the second and third level equation (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 $\nu_{00j}$ (random intercept on professor level) and a fixed effect parameter $\gamma_{000}$ (the overall mean). The variances of the three residual error terms are denoted by

$$ \text{var} (e_{ij}) = \sigma^2, \quad \text{var} (u_{0ij}) = \tau_0^2, \quad \text{var} (\nu_{00j}) = \phi_0^2 \quad (1.4) $$
This model was used to calculate the intra-class correlation for the professor level by dividing the variance of the professor level \((\phi_0^2)\) by the total variance \((\sigma^2 + \tau_0^2 + \phi_0^2)\). For the proportion of the total variance that is due to differences between students we divided the variance of the student level \((\tau_a^2)\) by the total variance \((\sigma^2 + \tau_0^2 + \phi_0^2)\).

**Random and fixed effects**

In the second model the linear main effects of *Time* (occasion 1, 2, 3 and 4, coded as 0, 1, 2 and 3) and the main effect of *Condition* (control group = 0 and experimental condition = 1) were added (Model 2). Model 2 is defined by the equations:

\[
\text{Level 1: } Y_{ij} = \beta_{0ij} + \beta_1 \text{Time}_{ij} + \beta_2 \text{Condition}_j + e_{ij} \quad (2.1)
\]
\[
\text{Level 2: } \beta_{0ij} = \beta_{00j} + u_{0ij} \quad (2.2)
\]
\[
\text{Level 3: } \beta_{00j} = \gamma_{000} + v_{00j} \quad (2.3)
\]

Here \(\beta_{0ij}\) is the intercept, \(\beta_1\) is the fixed effect parameter for *Time*, \(\beta_2\) is the fixed effect parameter for *Condition* and \(e_{ij}\) is the residual error term. Again the intercept \(\beta_{0ij}\) is allowed to be random over professors and students by the decomposition into one fixed component \((\gamma_{000})\) and two random components \((u_{0ij}\) and \(v_{00j})\) in de second and third level equations (2.2 and 2.3).

In the third model, we allowed the slope of professors and students to be random (Model 3). In both conditions, some professors may display systematic variation over time on the ratings. Similarly, students within classes display systematic variation in their ratings over time. Model 3 accommodated this possible variation. It is important to establish whether these random effects are significant, because their presence should be taken into account in studying the effects of the intervention. By comparing models 2 and 3 with a deviance test\(^2\), we evaluated whether it was necessary to retain a random slope. Model 3 (with random slope variances) is defined through the equations:

\[
\text{Level 1: } Y_{ij} = \beta_{0ij} + \beta_1 \text{Time}_{ij} + \beta_2 \text{Condition}_j + e_{ij} \quad (3.1)
\]
\[
\text{Level 2: } \beta_{0ij} = \beta_{00j} + u_{0ij} \quad (3.2)
\]
\[
\text{Level 2: } \beta_1 \text{Time}_{ij} = \beta_{10j} + u_{1ij} \quad (3.3)
\]
\[
\text{Level 3: } \beta_{00j} = \gamma_{000} + v_{00j} \quad (3.4)
\]
\[
\text{Level 3: } \beta_{10j} = \gamma_{100} + v_{10j} \quad (3.5)
\]

\(^2\) The deviance test is the likelihood ratio test to compare models; the \(-2*\log\)-likelihood of one model is compared with the \(-2*\log\)-likelihood of the other model. The difference has a chi-square distribution with degrees of freedom equal to the difference in the number of parameters estimated in the models being compared.
Here $\beta_{0ij}$ is the intercept, $\beta_1$ is the random effect parameter for Time, $\beta_2$ is the fixed effect parameter for Condition and $e_{ij}$ is the residual error term. Again the intercept $\beta_{0ij}$ is allowed to be random over students and professors by including the random components $u_{0ij}$ and $v_{00j}$.

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 $v_{10j}$. The fixed component $\gamma_{00j}$ represents the overall average regression coefficient for Time (mean slope).

The slope variances are denoted by

$$\text{var}(u_{1ij}) = \tau_1^2, \text{var}(v_{10j}) = \phi_1^2$$

(3.6)

The intercept-slope covariances are denoted by

$$\text{cov}(u_{0ij}, u_{1ij}) = \tau_{01}, \text{cov}(v_{00j}, v_{10j}) = \phi_{01}$$

(3.7)

**Effects of SET consultation**

With the fourth model we analyzed the effect of the intervention by adding the interaction effect $\text{Time} \times \text{Condition}$ (Model 4). Model 4 is defined through the equations:

**Level 1:**

$$Y_{ij} = \beta_{0ij} + \beta_{1ij} \text{Time}_{ij} + \beta_2 \text{Condition}_j + \beta_3 \text{Time} \times \text{Condition}_{ij} + e_{ij}$$

(4.1)

**Level 2:**

$$\beta_{0ij} = \beta_{00j} + u_{0ij}$$

(4.2)

**Level 2:**

$$\beta_{1ij} = \beta_{10j} + u_{1ij}$$

(4.3)

**Level 3:**

$$\beta_{00j} = \gamma_{000} + v_{00j}$$

(4.4)

**Level 3:**

$$\beta_{10j} = \gamma_{100} + v_{10j}$$

(4.5)

The parameters are the same as in Model 3. The additional fixed effect parameter for $\text{Time} \times \text{Condition}$ ($\beta_3$) represents the effect of SET consultation: if Model 4 fits the data better than the previous model according to a deviance test and the parameter of $\text{Time} \times \text{Condition}$ is significant, the control condition and experimental condition differ significantly in their ratings over time.

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 the control condition by dividing its mean difference of $T_4$ and $T_1$ by its pooled standard deviation ($\sqrt{((SD(T_{1,\text{condition}})^2 + SD(T_{4,\text{condition}})^2)) / 2}$). The second Cohen’s $d$ was calculated for the experimental condition versus the control condition by dividing the mean difference of $T_4$ and $T_1$ of the experimental condition minus the mean difference of $T_4$ and $T_1$ of the control condition by its pooled standard deviation ($\sqrt{((SD(T_{1,\text{control}})^2 + SD(T_{4,\text{control}})^2 + SD(T_{1,\text{exp}})^2 + SD(T_{4,\text{exp}})^2) / 4}$). Multilevel effect sizes were calculated based on the multilevel modeling output of the fourth model. The beta of $Time$ represents the change in mean of the control condition on each measurement occasion (three times the beta of $Time$ represents the change in mean of the control condition from $T_1$ to $T_4$). The beta of $Time*Condition$ represents the change in mean of the experimental condition compared to the control condition on each measurement occasion (three times the beta of $Time*Condition$ represents the relative change in mean from $T_1$ to $T_4$). 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 teacher residual standard deviation. The effect size of the control condition was therefore calculated by dividing three times the beta of $Time$ by the residual standard deviation $SD(e_{ij})$, and the effect size of the experimental condition compared to the control condition was calculated by dividing three times the beta of $Time*Condition$ 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 on each time-interval and on targeted versus non-targeted dimensions**

With the fifth model we specifically analyzed the effect of the first consultation and the additional effects of the second and third consultation on the dependent variable *Total Instructional Skills* (Model 5). We recoded $Time$ into the dummy variables $T_1T_2$, $T_2T_3$, and $T_3T_4$, representing the comparison of time 1 with 2, 2 with 3 and 3 with 4, respectively. 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 intercept in this model. Model 5 is defined by the equations:

$$Level \ 1:\quad Y_{ij} = \beta_{0ij} + \beta_1 \text{Condition}_j + \beta_2 T_1T_2_{ij} + \beta_3 T_2T_3_{ij} + \beta_4 T_3T_4_{ij} + \beta_5 T_1T_2*\text{Condition}_{ij} + \beta_6 T_2T_3*\text{Condition}_{ij} + \beta_7 T_3T_4*\text{Condition}_{ij} + e_{ij} \quad (5.1)$$
The parameter $\beta_{0ij}$ represents the intercept. The intercept is random over professors and students. The fixed effect parameter $\beta_1$ represents the main effect of Condition. The fixed effect parameters $\beta_2$, $\beta_3$ and $\beta_4$ for $T_1T_2$, $T_2T_3$ and $T_3T_4$ represent the contrasts. The fixed interaction effect parameter for $T_1T_2$*Condition$_{ij}$ ($\beta_5$) represents the effect of the first SET consultation. The fixed interaction effect parameters for $T_2T_3$*Condition$_{ij}$ ($\beta_6$) and $T_3T_4$*Condition$_{ij}$ ($\beta_7$) represent the additional effects of the second and third SET consultation.

If there was an effect of SET consultation on a specific time interval, we specifically analyzed the effect of targeted dimensions versus non-targeted dimensions on each dependent dimension on that time interval (Model 6). These additional exploratory analyses were done to link the effects of the intervention either to the feedback or the consultation. In Model 6, professors in the experimental condition were separated into two groups for each dimension on the specific time interval; a group which targeted the dimension for improvement (Target) and a group that did not target the dimension (No Target). Condition was therefore recoded into the dummy variables Control-versus-No Target and Control-versus-Target. We restricted the analyses to the time intervals associated with an effect in Model 5 to limit the number of tests on the data. In addition, to prevent a Type I error, these effects were tested with a more conservative alpha of .01. Time was recoded for the specific time interval (in case of time interval $T_1T_2$: $T_1 = 0$ and $T_2 = 1$, in case of time interval $T_2T_3$: $T_2 = 0$ and $T_3 = 1$, in case of time interval $T_2T_3$: $T_2 = 0$ and $T_3 = 1$). Again, we limited the random effects to the intercept in this model. Model 6 is defined by the equations:

\begin{align*}
\text{Level 1:} & \quad Y_{tij} = \beta_{0ij} + \beta_1 \text{Time}_{tij} + \beta_2 \text{Control-versus-NoTarget}_j \\
& \quad + \beta_3 \text{Control-versus-Target}_j + \beta_4 \text{Time}*\text{Control-versus-NoTarget}_{ij} \\
& \quad + \beta_5 \text{Time}*\text{Control-versus-Target}_{ij} + e_{tij} \
\text{Level 2:} & \quad \beta_{0ij} = \beta_{00j} + u_{0ij} \
\text{Level 3:} & \quad \beta_{00j} = y_{000} + v_{00j}
\end{align*}

The parameter $\beta_{0ij}$ represents the intercept. The intercept is random over professors and students. The fixed effect parameter $\beta_1$ represents the main effect of Time on the specific time interval. The fixed effect parameters $\beta_2$ and $\beta_3$ represent the contrasts Control-versus-No Target and Control-versus-Target. The fixed interaction effect parameter for $\text{Time}*\text{Control-versus-NoTarget}_{ij}$ ($\beta_4$) and $\text{Time}*\text{Control-versus-Target}_{ij}$ ($\beta_5$) represent the effects of SET consultation for non-targeted dimensions and targeted dimensions compared to the control condition.
Table 3.1 shows the number of participating professors, the number of ISQ forms completed, mean ISQ scores, the standard deviation of the professors and the standard deviation of the students within classes in the two conditions on each measurement occasion. Mean ratings and standard deviations of the professors on each dimension are shown in Figures 3.2a-h. At baseline (occasion 1), multilevel t-test revealed that there were no significant mean differences between the conditions on each dimension. The experimental and control condition were therefore comparable on teaching skills at baseline. The control and experimental conditions did not differ with respect to the inevitable student drop out ($\chi^2[3] = 5.834, p = 0.12$).

Figure 3.2a-2h  Mean ratings on the eight dependent variables in the experimental condition ($N = 12$) and control condition ($N = 13$) on the four measurement occasion.
Based on the intercept-only multilevel regression model (Model 1), the intra-class correlation for the professor level for all dependent variables varied between .09 and .42, with a mean of .21. The intra-class correlation for the student level varied between .26 and .51, with a mean of .38. Since a mean of 21% of the variance is due to differences between professors and a mean of 38% of the variance is due to differences between students within professors, the use of multilevel regression modeling is indicated.

We performed deviance tests between the first four models on all eight dependent variables, to determine which model fitted the data best. Table 3.2 shows the deviance tests between the first four models on Total Instructional Skills

---

Tables of detailed results on the seven specific dimensions are available on request.
Table 3.1  Number of participants, mean ISQ scores and standard deviations on professor level and student level on measurement occasion 1 to 4 for the control condition and experimental condition

<table>
<thead>
<tr>
<th></th>
<th>Measurement occasion 1</th>
<th></th>
<th></th>
<th>Measurement occasion 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Professor level</td>
<td>Student level</td>
<td>Professor level</td>
<td>Student level</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(between classes)</td>
<td>(within classes)</td>
<td>(between classes)</td>
<td>(within classes)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>N  M  SD</td>
<td>N  SD Min  Median  Max</td>
<td>N  M  SD</td>
<td>N  SD Min  Median  Max</td>
<td></td>
</tr>
<tr>
<td>Control group</td>
<td>13  231</td>
<td></td>
<td>13  186</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Structure</td>
<td>3.65 0.23</td>
<td>0.24 0.48 0.91</td>
<td>3.65 0.35</td>
<td>0.20 0.45 0.74</td>
<td></td>
</tr>
<tr>
<td>Explication</td>
<td>3.77 0.17</td>
<td>0.12 0.47 0.77</td>
<td>3.69 0.12</td>
<td>0.26 0.45 0.77</td>
<td></td>
</tr>
<tr>
<td>Stimulation</td>
<td>3.81 0.44</td>
<td>0.30 0.51 0.99</td>
<td>3.66 0.49</td>
<td>0.44 0.58 0.95</td>
<td></td>
</tr>
<tr>
<td>Validation</td>
<td>3.61 0.25</td>
<td>0.24 0.48 0.72</td>
<td>3.56 0.23</td>
<td>0.17 0.51 0.72</td>
<td></td>
</tr>
<tr>
<td>Instruction</td>
<td>3.50 0.21</td>
<td>0.24 0.45 0.66</td>
<td>3.41 0.17</td>
<td>0.35 0.56 0.69</td>
<td></td>
</tr>
<tr>
<td>Comprehension</td>
<td>3.93 0.23</td>
<td>0.29 0.42 0.59</td>
<td>3.85 0.28</td>
<td>0.25 0.50 0.63</td>
<td></td>
</tr>
<tr>
<td>Activation</td>
<td>3.77 0.43</td>
<td>0.40 0.56 0.74</td>
<td>3.68 0.49</td>
<td>0.27 0.56 0.79</td>
<td></td>
</tr>
<tr>
<td>Total Instruction Skills</td>
<td><strong>3.72</strong> 0.20</td>
<td><strong>0.19</strong> 0.31 0.47</td>
<td><strong>3.64</strong> 0.18</td>
<td><strong>0.21</strong> 0.34 0.55</td>
<td></td>
</tr>
<tr>
<td>Experimental group</td>
<td>12  378</td>
<td></td>
<td>12  352</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Structure</td>
<td>3.77 0.25</td>
<td>0.30 0.52 0.61</td>
<td>3.87 0.17</td>
<td>0.36 0.45 0.63</td>
<td></td>
</tr>
<tr>
<td>Explication</td>
<td>3.84 0.17</td>
<td>0.36 0.49 0.64</td>
<td>3.83 0.21</td>
<td>0.39 0.47 0.71</td>
<td></td>
</tr>
<tr>
<td>Stimulation</td>
<td>3.69 0.31</td>
<td>0.32 0.61 0.76</td>
<td>3.62 0.29</td>
<td>0.22 0.63 0.78</td>
<td></td>
</tr>
<tr>
<td>Validation</td>
<td>3.65 0.27</td>
<td>0.29 0.48 0.66</td>
<td>3.61 0.22</td>
<td>0.39 0.55 0.82</td>
<td></td>
</tr>
<tr>
<td>Instruction</td>
<td>3.43 0.19</td>
<td>0.33 0.53 0.68</td>
<td>3.53 0.19</td>
<td>0.35 0.54 0.71</td>
<td></td>
</tr>
<tr>
<td>Comprehension</td>
<td>3.81 0.50</td>
<td>0.31 0.48 0.71</td>
<td>3.91 0.39</td>
<td>0.35 0.51 0.67</td>
<td></td>
</tr>
<tr>
<td>Activation</td>
<td>3.37 0.64</td>
<td>0.40 0.58 0.70</td>
<td>3.51 0.65</td>
<td>0.39 0.56 0.75</td>
<td></td>
</tr>
<tr>
<td>Total Instruction Skills</td>
<td><strong>3.65</strong> 0.22</td>
<td><strong>0.21</strong> 0.36 0.46</td>
<td><strong>3.70</strong> 0.24</td>
<td><strong>0.27</strong> 0.38 0.50</td>
<td></td>
</tr>
</tbody>
</table>
Table 3.1  

<table>
<thead>
<tr>
<th></th>
<th>Measurement occasion 3</th>
<th>Measurement occasion 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Professor level (between classes)</td>
<td>Student level (within classes)</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>M</td>
</tr>
<tr>
<td>Control group</td>
<td>13</td>
<td>157</td>
</tr>
<tr>
<td>Structure</td>
<td>3.56</td>
<td>0.27</td>
</tr>
<tr>
<td>Explication</td>
<td>3.65</td>
<td>0.20</td>
</tr>
<tr>
<td>Stimulation</td>
<td>3.63</td>
<td>0.43</td>
</tr>
<tr>
<td>Validation</td>
<td>3.64</td>
<td>0.33</td>
</tr>
<tr>
<td>Instruction</td>
<td>3.39</td>
<td>0.26</td>
</tr>
<tr>
<td>Comprehension</td>
<td>3.82</td>
<td>0.28</td>
</tr>
<tr>
<td>Activation</td>
<td>3.66</td>
<td>0.49</td>
</tr>
<tr>
<td>Total Instructional Skills</td>
<td>3.61</td>
<td>0.20</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Measurement occasion 3</th>
<th>Measurement occasion 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Professor level (between classes)</td>
<td>Student level (within classes)</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>M</td>
</tr>
<tr>
<td>Experimental group</td>
<td>11</td>
<td>225</td>
</tr>
<tr>
<td>Structure</td>
<td>3.89</td>
<td>0.25</td>
</tr>
<tr>
<td>Explication</td>
<td>3.80</td>
<td>0.23</td>
</tr>
<tr>
<td>Stimulation</td>
<td>3.59</td>
<td>0.34</td>
</tr>
<tr>
<td>Validation</td>
<td>3.62</td>
<td>0.16</td>
</tr>
<tr>
<td>Instruction</td>
<td>3.56</td>
<td>0.28</td>
</tr>
<tr>
<td>Comprehension</td>
<td>3.96</td>
<td>0.39</td>
</tr>
<tr>
<td>Activation</td>
<td>3.64</td>
<td>0.59</td>
</tr>
<tr>
<td>Total Instructional Skills</td>
<td>3.71</td>
<td>0.22</td>
</tr>
</tbody>
</table>

Note: student level N refers to the number of evaluation forms completed at measurement occasion one, two, three and four.
For all of the dependent variables the deviance tests between the second and third multilevel regression models showed that the third model (with the explanatory variables Time and Condition and a random intercept and slope) fitted the data significantly better than the second model (without a random slope). This significance of the random slope shows that professors and students vary significantly in the change of ratings during a course. We retained these random effects when we investigate the effect of the intervention with Model 4.

Model 4, which included Time*Condition as the effect of the intervention, fitted the data significantly better than Model 3 for four out of eight dependent variables: Explication, Comprehension, Activation and Total Instructional Skills. Model 3 (without the interaction parameter for Time*Condition) was therefore the final model for the remaining variables; Structure, Stimulation, Validation, and Instruction. Table 3.3 shows the estimates and standard errors of the explanatory variables under Model 3 for these dependent variables. Table 3.4 shows the estimates and standard errors of the explanatory variables under Model 4 on the dependent variables Explication, Comprehension, Activation and Total Instructional Skills. In short, professors in the experimental condition significantly improved their instructional skills on three specific dimensions and on their total rating score with the specified feedback and consultation protocol compared to the control condition. Table 3.5 shows effects sizes calculated with Cohen's d and calculated with multilevel Model 4 output for the dependent variables on which significant effects of the intervention were found. The effect size was large on Total Instructional Skills (Cohen's d = .84, multilevel effect size =.88) and medium to large on the three specific dimensions (Cohen's d ranged from .39 to .86, multilevel effect sizes ranged from .42 to .73). Notably, the effect sizes on the specific dimensions calculated with

<table>
<thead>
<tr>
<th>Table 3.2</th>
<th>Deviance tests between the four models for Total Instructional Skills</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-2 log likelihood</td>
</tr>
<tr>
<td>Model 1 (random intercept only)</td>
<td>1655.1</td>
</tr>
<tr>
<td>Model 2 (random intercept + Time + Condition)</td>
<td>1653.4</td>
</tr>
<tr>
<td>Model 3 (random intercept + random slope + Time + Condition)</td>
<td>1599.8</td>
</tr>
<tr>
<td>Model 4 (random intercept + random slope + Time + Condition + Time*Condition)</td>
<td>1591.7</td>
</tr>
</tbody>
</table>
Table 3.3  Estimates and standard errors of the explanatory variables under Model 3 for the dependent variables *Structure, Stimulation, Validation* and *Instruction*

<table>
<thead>
<tr>
<th></th>
<th>Structure Model 3</th>
<th>Stimulation Model 3</th>
<th>Validation Model 3</th>
<th>Instruction Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate ( SE )</td>
<td>Estimate ( SE )</td>
<td>Estimate ( SE )</td>
<td>Estimate ( SE )</td>
</tr>
<tr>
<td><strong>Fixed Part</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_0$</td>
<td>Constant</td>
<td>3.360 ( 0.069 )</td>
<td>3.762 ( 0.104 )</td>
<td>3.588 ( 0.071 )</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>Time</td>
<td>-0.001 ( 0.021 )</td>
<td>-0.044 ( 0.017 )</td>
<td>* -0.005 ( 0.016 )</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>Condition</td>
<td>0.189 ( 0.096 )</td>
<td>-0.089 ( 0.148 )</td>
<td>0.066 ( 0.095 )</td>
</tr>
<tr>
<td><strong>Random Part</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \varphi_{01} )</td>
<td>Level 3: Professor</td>
<td>0.051 ( 0.017 ) *</td>
<td>0.122 ( 0.039 ) *</td>
<td>0.060 ( 0.020 ) *</td>
</tr>
<tr>
<td>( \tau_{01} )</td>
<td>Level 2: Student</td>
<td>0.136 ( 0.017 ) *</td>
<td>0.263 ( 0.025 ) *</td>
<td>0.131 ( 0.017 ) *</td>
</tr>
<tr>
<td>( \sigma^2 )</td>
<td>Level 1: Time</td>
<td>0.141 ( 0.010 ) *</td>
<td>0.164 ( 0.012 ) *</td>
<td>0.138 ( 0.010 ) *</td>
</tr>
<tr>
<td><strong>Slope variance</strong></td>
<td>( \varphi_{1} )</td>
<td>Level 3: Professor</td>
<td>0.008 ( 0.003 ) *</td>
<td>0.003 ( 0.002 )</td>
</tr>
<tr>
<td>( \tau_{1} )</td>
<td>Level 2: Student</td>
<td>0.002 ( 0.004 )</td>
<td>0.011 ( 0.005 ) *</td>
<td>0.007 ( 0.004 )</td>
</tr>
<tr>
<td><strong>Intercept-slope covariance</strong></td>
<td>( \varphi_{01} )</td>
<td>Level 3: Professor</td>
<td>-0.004 ( 0.005 )</td>
<td>0.002 ( 0.006 )</td>
</tr>
<tr>
<td></td>
<td>Level 2: Student</td>
<td>0.000 ( 0.007 )</td>
<td>-0.014 ( 0.010 )</td>
<td>0.000 ( 0.007 )</td>
</tr>
</tbody>
</table>

Notes: * $p < .05$. Units level 3: 25 professors, units level 2: 1,333 students, units level 1: 1,954 questionnaires.
Table 3.4  Estimates and standard errors of the explanatory variables under Model 4 for the dependent variables Total Instructional Skills, Explication, Comprehension, and Activation

<table>
<thead>
<tr>
<th></th>
<th>Instructional Skills Model 4</th>
<th>Explication Model 4</th>
<th>Comprehension Model 4</th>
<th>Activation Model 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate ( SE )</td>
<td>Estimate ( SE )</td>
<td>Estimate ( SE )</td>
<td>Estimate ( SE )</td>
</tr>
<tr>
<td>Fixed Part</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_0$ Constant</td>
<td>3.710 ( 0.060 )</td>
<td>3.763 ( 0.055 )</td>
<td>3.922 ( 0.101 )</td>
<td>3.748 ( 0.146 )</td>
</tr>
<tr>
<td>$\beta_1$ Time</td>
<td>-0.040 ( 0.015 ) *</td>
<td>-0.057 ( 0.019 ) *</td>
<td>-0.040 ( 0.024 )</td>
<td>-0.026 ( 0.035 )</td>
</tr>
<tr>
<td>$\beta_2$ Condition</td>
<td>-0.045 ( 0.086 )</td>
<td>0.074 ( 0.076 )</td>
<td>-0.086 ( 0.144 )</td>
<td>-0.340 ( 0.209 )</td>
</tr>
<tr>
<td>$\beta_3$ Time*Condition</td>
<td>0.065 ( 0.020 ) *</td>
<td>0.053 ( 0.025 ) *</td>
<td>0.086 ( 0.033 ) *</td>
<td>0.103 ( 0.049 ) *</td>
</tr>
<tr>
<td>Random Part</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept variance</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\varphi_0^2$ Level 3: Professor</td>
<td>0.041 ( 0.013 ) *</td>
<td>0.025 ( 0.010 ) *</td>
<td>0.118 ( 0.036 ) *</td>
<td>0.258 ( 0.077 ) *</td>
</tr>
<tr>
<td>$\tau_0^2$ Level 2: Student</td>
<td>0.091 ( 0.008 ) *</td>
<td>0.153 ( 0.018 ) *</td>
<td>0.114 ( 0.011 ) *</td>
<td>0.175 ( 0.022 ) *</td>
</tr>
<tr>
<td>$\sigma^2$ Level 1:Time</td>
<td>0.047 ( 0.003 ) *</td>
<td>0.141 ( 0.010 ) *</td>
<td>0.158 ( 0.009 ) *</td>
<td>0.175 ( 0.012 ) *</td>
</tr>
<tr>
<td>Slope variance</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\varphi_0^2$ Level 3: Professor</td>
<td>0.001 ( 0.001 )</td>
<td>0.001 ( 0.001 )</td>
<td>0.004 ( 0.002 ) *</td>
<td>0.011 ( 0.004 ) *</td>
</tr>
<tr>
<td>$\tau_0^2$ Level 2: Student</td>
<td>0.005 ( 0.002 ) *</td>
<td>0.008 ( 0.004 ) *</td>
<td>0.000 ( 0.000 )</td>
<td>0.005 ( 0.005 )</td>
</tr>
<tr>
<td>Intercept-slope covariance</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\varphi_{01}$ Level 3: Professor</td>
<td>-0.002 ( 0.002 )</td>
<td>-0.001 ( 0.002 )</td>
<td>-0.009 ( 0.006 )</td>
<td>-0.007 ( 0.013 )</td>
</tr>
<tr>
<td>$\tau_{01}$ Level 2: Student</td>
<td>0.002 ( 0.003 )</td>
<td>-0.003 ( 0.007 )</td>
<td>0.000 ( 0.000 )</td>
<td>-0.001 ( 0.009 )</td>
</tr>
</tbody>
</table>

Notes: * p < .05. Units level 3: 25 professors, units level 2: 1,333 students, units level 1: 1,954 questionnaires.
Table 3.5  Effect sizes calculated with Cohen’s $d$ and calculated with Model 4 output for the dependent variables Explication, Comprehension, Activation, and Total Instructional Skills

<table>
<thead>
<tr>
<th>Condition</th>
<th>Control condition</th>
<th>Experimental condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comparison</td>
<td>(T₄ vs T₁)</td>
<td>(T₄ vs T₁) vs Control (T₄ vs T₁)</td>
</tr>
<tr>
<td>Effect size</td>
<td>Cohen’s $d$</td>
<td>Multilevel model</td>
</tr>
<tr>
<td>Calculation</td>
<td>$M(T₄_C) - M(T₁_C)$</td>
<td>$3^\times$Time pooled SD</td>
</tr>
<tr>
<td></td>
<td>$SD (e_{ij})$</td>
<td>$SD (e_{ij})$</td>
</tr>
<tr>
<td>Explication</td>
<td>-0.77</td>
<td>-0.46</td>
</tr>
<tr>
<td>Comprehension</td>
<td>-0.41</td>
<td>-0.31</td>
</tr>
<tr>
<td>Activation</td>
<td>-0.03</td>
<td>-0.18</td>
</tr>
<tr>
<td><strong>Total Instructional Skills</strong></td>
<td><strong>-0.49</strong></td>
<td><strong>-0.55</strong></td>
</tr>
</tbody>
</table>

Notes: the control condition is denoted C, the experimental condition is denoted E. Pooled $SD$ for the comparison T₄ vs T₁ of the control condition was calculated with $\sqrt{(SD(T₁_C)^2 + SD(T₄_C)^2)/2}$. Pooled $SD$ for the comparison T₄ vs T₁ of the experimental condition versus the control condition was calculated with $\sqrt{(SD(T₁_C)^2 + SD(T₄_C)^2 + SD(T₁_E)^2 + SD(T₄_E)^2)/4}$.

Model 4 output deviated from the effect sizes calculated with Cohen’s $d$, due to the random effects that were taken into account in the multilevel analyses.

Class sizes in the control condition were smaller than in the experimental condition on the first measurement occasion (mean control condition versus mean experimental condition; $t(18.8) = -2.5, p = .02$). Class Size (mean-centered) was therefore added as a covariate to Model 4 (Model 4b) in analyzing the effects of the intervention. Class Size did not influence the results found with Model 4. The increase in AIC and BIC with Model 4b did indicate an overfit (with additional parameters the AIC and BIC fit statistics normally decrease).

With the fifth model, we specifically analyzed the effect of the first consultation and the additional effects of the second and third consultation on the dependent variable Total Instructional Skills. Table 3.6 shows the estimates and standard errors of the explanatory variables under Model 5.

Results showed a significant effect of the first consultation (parameter $\beta_5$) and no additional effects of the second and third consultation (parameters $\beta_6$ and $\beta_7$). The parameter of $T₁T₂ (\beta_2)$ indicates that the control condition decreased significantly in ratings between the first and second measurement occasion. Compared to the control condition, the experimental condition significantly increased in ratings on the same time interval (parameter $\beta_5$). The
ratings of the control condition decreased further between the second and third measurement occasion, but not significantly (parameter $\beta_3$), and remains stable between the third and fourth measurement occasion (parameter $\beta_4$). Compared to the control condition, ratings of the experimental condition increased after the second consultation, but not significantly (parameter $\beta_6$), and remained stable after the third consultation (parameter $\beta_7$).

Because there was a significant effect of SET consultation on the first time interval $(T_1T_2)$, the differences in effects of targeted and non-targeted dimensions was analyzed on this time interval with the sixth model with a more conservative alpha of .01 for each of the seven specific dimensions. Results of these exploratory analyses showed a significant increase in ratings when they were targeted compared to the control condition on four dimensions: Structure, Instruction, Comprehension and Activation (Structure: $\beta = .22$, $SE = .07$, $p = .002$; Instruction: $\beta = .23$, $SE = .07$, $p = .001$; Comprehension: $\beta = .40$, $SE = .06$, $p < .001$; Activation: $\beta = .29$, $SE = .07$, $p < .001$). When dimensions were not targeted there was still an effect on one dimension: Instruction ($\beta = .21$, $SE = .07$, $p = .004$). Furthermore, control condition ratings on Stimulation ($\beta = -.15$, $SE = .05$, $p = .008$) decreased significantly on this time interval. Ratings of the targeted dimensions started lower on baseline ratings compared to

<table>
<thead>
<tr>
<th>Table 3.6</th>
<th>Estimates and standard errors of the explanatory variables under Model 5 for the dependent variable Total Instructional Skills</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fixed Part</strong></td>
<td><strong>Instructional Skills</strong> Model 5</td>
</tr>
<tr>
<td>$\beta_0$</td>
<td>Constant</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>Condition</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>T1T2</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>T2T3</td>
</tr>
<tr>
<td>$\beta_4$</td>
<td>T3T4</td>
</tr>
<tr>
<td>$\beta_5$</td>
<td>T1T2*Condition</td>
</tr>
<tr>
<td>$\beta_6$</td>
<td>T2T3*Condition</td>
</tr>
<tr>
<td>$\beta_7$</td>
<td>T3T4*Condition</td>
</tr>
<tr>
<td><strong>Random Part</strong></td>
<td><strong>Intercept variance</strong></td>
</tr>
<tr>
<td>$\varphi_0^2$</td>
<td>Level 3: Professor</td>
</tr>
<tr>
<td>$\tau_0^2$</td>
<td>Level 2: Student</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>Level 1: Time</td>
</tr>
</tbody>
</table>

Notes: * $p < .05$. Units level 3: 25 professors, units level 2: 1,333 students, units level 1: 1,954 questionnaires.
Experiment I – Effects on professors’ lecturing skills

Chapter 3

non-targeted dimensions for almost all dimensions, but in no case significantly lower with an alpha of .01 (on the dimension Structure ratings were lower with an alpha of .05).

Discussion

Student evaluations of teaching (SETs) collected at the end of the course often do not help improve professors’ instructional skills (Kember, Leung & Kwan, 2002). This may be due to bad timing, lack of specificity, and ineffective use of the feedback. This study addressed the effects of intermediate student evaluations of teaching followed by collaborative consultation on professors’ instructional skills, compared to a control group. We collected student feedback on seven dimensions of instructional skills at the end of four single lectures during the course (class meetings in which lecturing was the teaching format). In so doing we ensured the feedback was optimally timed and highly specific. Professors in the experimental condition met with a consultant within two or three days after each evaluated lecture to formulate an appropriate action plan for the following lectures based on the feedback. By repeating this procedure of feedback and collaborative consultation during the course, we evaluated the effects of each SET consultation on the teaching skills. On the time-intervals on which the intervention had a significant impact, the effects of targeted dimensions compared to non-targeted dimensions were further investigated.

At baseline, professors in the experimental and control condition were comparable on teaching skills. The courses were taught in the same semester, the students were at the same academic level (third and final bachelor year), and the teaching format was the same (i.e., lectures). Also, on the seven specific dependent teaching dimensions and on Total Instructional Skills, the two conditions did not differ at baseline.

The professors, who participated in the experimental condition, showed a significant increase in Total Instructional Skills compared to the control condition. More specifically, we found significant effects of the intervention on the instructional dimensions Explication, Comprehension and Activation. In our analyses of the ratings on all seven dimensions, we included significant intercept and slope variances between professors and between students within classes. The effects of the intervention are therefore significant despite the differences between individual professors on their baseline rating, and despite differences between professors in how much they randomly change in ratings over time. The effects of the intervention are also significant, despite the differences between students within classes.

Time-interval analyses showed that, of the three consultations during the course, only the first consultation resulted in a significant effect on the professors’ Total Instructional
Skills ratings. The ratings in the experimental condition did increase (relative to the control condition) after the second consultation, but this increase was not statistically significant. The third consultation did not result in an increase in ratings. Thus, only the first SET consultation had a significant impact on student ratings of their professors. Further analyses on this time interval (T₁ vs. T₂) showed a significant increase for four dimensions when they were targeted for improvement (Structure, Instruction, Comprehension and Activation) and for one dimension (Instruction) when dimensions were not targeted for improvement. In the control condition, ratings on one dimension (Stimulation) decreased significantly during this time interval.

We note that the differences observed were small from an absolute perspective. This was due to the small effective scale; although we used a 5-point Likert scale, the actual range of professors’ mean ratings was much smaller. The baseline mean ratings of professors on Total Instructional Skills varied from 3.32 to 4.07, with a high baseline mean of 3.7 and a small standard deviation of .22. The relatively small range and variance are comparable to those of previous experimental studies on SET consultation (e.g., Marsh & Roche, 1993; Hampton & Reiser, 2004).

The absolute size of the differences between experimental and control condition is therefore rather small, but relative to the standard deviation it is substantial. The effect size, in terms of Cohen’s d, on Total Instructional Skills was large (d = .84) and medium to large on the specific dimensions on which we found an effect (ranging from .39 to .86). Penny and Coe (2004) found that mainly studies with an advisory or educational approach showed large effects. As interventions, these approaches are more elaborate than the collaborative intervention used in this study. Considering the costs of the current and more extensive interventions, the approach to SET consultation used in this study seems to be valuable. Mainly the first intermediate SET consultation results in appreciable effects. The second and third intermediate SET consult appear to be less beneficial.

Notably, ratings of professors in the control condition showed a decrease over time. One explanation could be that professors may have given their best in the first few lectures, and resorted to routine later in the course. The results on specific dimensions show a decrease on Explication and Stimulation in the control condition. Both depend on using lively examples and diverse stimulating and clarifying ways of presenting the subject matter, which requires extra time to prepare. While ratings of professors in the control group decreased on Explication, ratings of professors in the experimental condition increased on this dimension. The effects of the intervention are therefore visible in terms of an increase in ratings as well as the prevention of decrease.
Furthermore, we note that there was a difference in class sizes between the conditions. Professors in the experimental condition taught significantly larger groups of students than professors in the control condition did. When we controlled for differences in class sizes it did not influence the results, but the increase in fit indices indicated an over fit, meaning that there was not enough data to fit such a complex model. This means that the difference between the conditions may still have had an effect on the ratings. For example, this may explain why professors in the experimental condition started out lower on the dimension Activation, since it is more difficult to interact with larger groups of students. Nevertheless, during the intervention, these professors seemed to catch up, since the ratings on Activation in the experimental condition increased significantly over time, and reached the same level of ratings on Activation as the control condition on the third measurement occasion. The Activation ratings of the control condition remained stable over time.

In terms of scientific relevance, the present study complements recent non-experimental findings (e.g., Dresel & Rindermann, 2011; Rindermann, Kohler, & Meisenberg, 2007) with positive experimental results for a collaborative approach to SET consultation. Marsh and Roche (1993) were one of the few to have found no effects of mid-term collaborative SET consultation. They noted that the mid-term feedback may have been less effective compared to end-of-the-term feedback as the course evaluation instrument used (SEEQ) contained inappropriate items (e.g., items relating to assignments and examinations) to evaluate mid-term teaching effectiveness. The authors suggested that this may have undermined the confidence in the intervention for the mid-term group. They did find positive effects of (more appropriate) end-of-the-term SET consultation on ratings collected one semester later. The generalizability of mid-term SETs to end-of-the-term SETs have been questioned by others as well (L’Hommedieu et al., 1990). The SET instrument in the current study measured specific teaching dimensions relevant and appropriate to intermediate evaluation and comparable with ratings collected at the subsequent measurement occasions. We recommend researchers in this field as well as faculty developers to be considerate of the relevance, specificity and comparability of the student feedback used in (research on) SET consultation.

An additional important feature of this study is the use of multilevel regression analyses to take into account systematic variation between professors and students in their ratings on teaching effectiveness at baseline and over time, when investigating the effects of the intervention. Compared to Cohen’s $d$ effect sizes (calculated on the teacher mean ratings), the effect sizes calculated on the multilevel data showed a similar large effect on Total Instructional Skills (.88), but deviant effects on the specific dimensions. We stress the importance of
taking random effects into account in future research, as they were of significant influence on analyses on each dependent teaching variable investigated in this study.

Although the results of this study showed positive experimental effects of SET consultation on different teaching skills, three limitations to this study deserve attention when interpreting these findings. They also imply suggestions for further research. First, the results represent the combined effects of intermediate feedback and consultation. In order to determine whether the specific results are due to the student feedback or due to the consultation, future studies on collaborative consultation should differentiate in conditions with one or the other. Here, ratings on targeted dimensions evidently showed more increase than ratings on non-targeted dimensions, which provides an argument that the collaborative consultation made a difference rather than the feedback alone. These results agree with Marsh and Roche's findings (1993) showing that targeted dimension were associated with significantly more improvement than non-targeted dimensions. In addition, previous studies have consistently found small effects of mid-term feedback only and larger effects of feedback plus consultation (Cohen, 1980; Lang & Kersting, 2007; Menges & Brinko, 1986; Penny & Coe, 2004), thus suggesting that the currently found effects are due to the consultation. Nevertheless, we suggest that the present results are complemented with further research on this approach to SET consultation with a more complex design.

Second, the sample size of twenty-five professors is relatively small. Consequently, it was not possible to investigate differences in effects due to potentially relevant teaching and course characteristics (e.g., faculty gender, rank, age, experience, prior teaching quality and class size). New faculty may for example respond differently to the intervention than faculty of older age, who are more experienced, but also – possibly – more set in their ways. The small professor sample size is therefore a limitation of this study. Notably, previous studies have found positive effects of SET consultation with teaching assistants (e.g., Hampton & Reiser, 2004: 37 teaching assistants) as well as full-time ranked faculty (e.g., Dresel & Rindermann, 2011: 12 faculty). Additionally, in their meta-analysis, Penny and Coe (2004) found no differences in effects of SET consultation between teaching assistants and full-time faculty. But, as noted earlier, the number of experimental studies available is limited. Particularly studies with large and diverse samples. With regard to prior teaching effectiveness, Marsh and Roche (1993) found that professors who were initially less effective benefited more from the received mid-term plus end-of-the-term SET consultation than the initially more effective professors. Their study is one of the few studies with a large sample in which this issue was addressed. Considering the costs of faculty development practices and given the current findings and the corresponding equal effects found on more extensive interventions, research on this matter is important.
With more knowledge on the moderating effects of potentially relevant teaching and course characteristics on the current and more extensive interventions, faculty developers can target and combine the optimal interventions, corresponding to prior aims for improvement.

Finally, some planned improvement may require several lectures to implement successfully, and major changes in the course or lectures cannot always be achieved during the current course. Piccinin and colleagues (1999) found a delayed effect, in terms of an increase on course ratings, one to three years after the initial SET consultation. In these cases, the feedback and consultation may not have an immediate effect on teaching behavior, but may still have an impact on the professor’s perception of his or her teaching, goals, attitudes, self-efficacy, and teaching strategies. Related to this is the notion that researchers on faculty evaluation often recommend the use of multiple sources of data to assess teaching quality (Benton & Cashin, 2012). Here, we are limited to assessment by student ratings. In future studies, it would be useful to include additional outcome variables to gain insight in the full impact of this intervention.

In summary, with regard to implications for future research, we conclude that the present results justify further research on this approach to SET consultation on a larger scale with a more complex design. The results are promising, but more experimental research in this field is necessary to solidify these findings. The use of multilevel regression analyses in future investigations in this field is highly recommended. With regard to implications for future practice, the results of this study indicate that SET consultation is potentially equal effective as more intensive interventions, when feedback is well timed, relevant and specific, and when consultation is collaborative and teacher-centered. Under these conditions, we observed that consultation does not need to be repeated often during a course in order to have a significant impact.
References


