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.
Summary and discussion
As accountability concerning the quality of teaching at universities has become more important over the years, so have instructional development practices to support and improve the teaching of university professors. At the same time, the effectiveness of instructional development practices in the field is seldom investigated thoroughly. Reviewers have consistently called for more experimental research on various levels of evaluation (Levinson-Rose & Menges, 1981; Prebble et al., 2004; Steinert et al., 2006; Stes, Min-Leliveld, Gijbels, & Van Petegem, 2010; Weimer & Lenze, 1997). This dissertation concerns the effectiveness of providing university professors with intermediate students’ evaluations of teaching (SETs) on individual lectures, with or without collaborative consultation with a consultant (SET consultation).

The aim of this dissertation was to investigate the effectiveness of these two types of interventions in terms of Guskey’s (2000) first, second, fourth, and fifth level of evaluation, i.e., in terms of professors’ self-reported satisfaction with the interventions (level 1: satisfaction), professors’ self-reported learning concerning lecturing (level 2: learning), professors’ lecturing skills, as measured by students’ evaluations of lecturing (level 4: behavior), and students’ self-assessed learning outcomes (level 5: student learning). Additional aims of this dissertation were: 1) to investigate the psychometric quality of the SET instrument (the Instructional Skills Questionnaire, ISQ) used to provide professors with feedback, and to evaluate the improvement in professors’ lecturing skills on seven specific teaching dimensions; 2) to investigate the specific effects on student ratings of a first intermediate consultation, and the additional effects of a second and third intermediate consultation; 3) to investigate the differences in effects on teaching dimensions which were and were not targeted for improvement by the professor during consultation; 4) to investigate the moderating effects on each level of evaluation of specific professor and course characteristics (i.e., professors’ age, professors’ prior quality of teaching, and class size).

The dissertation includes two experimental studies. The first experiment concerned a pilot study at the University of Amsterdam, with 25 (assistant, associate, or full) Psychology professors, 1,333 students, and 2 consultants. Professors were randomly assigned to either the experimental condition with intermediate feedback-plus-consultation or the control condition with neither feedback nor consultation. The second experiment concerned a larger study at the same university, with 75 professors from a wide variety of departments, 9,616 students, and 5 consultants. Professors were randomly assigned to one of three conditions, an intermediate feedback-plus-consultation condition, an intermediate feedback-only condition, or a control condition with neither feedback nor consultation. Where appropriate, the data were analyzed using multi-level modeling to take into account random differences between students and professors.
In this final chapter, I discuss the findings in this dissertation and draw my conclusions on the implications of these findings to research and practices in this field.

**Main findings**

**Quality of the Instructional Skills Questionnaire**

Chapter 2 concerned an investigation of the psychometric quality of the instrument (the ISQ), used to measure seven dimensions of the professors’ lecturing skills, as assessed by students. The analyses were based on 14,298 ISQ forms administered in the second experiment on three measurement occasions.

The conceptualization of teaching behavior in terms of the seven ISQ dimensions (Structure, Explication, Stimulation, Validation, Instruction, Comprehension and Activation) was based on the dimensions previously proposed in the literature (Marsh & Hocevar, 1991b; De Neve & Janssen, 1982; Vorst & Van Engelenburg, 1992), and on Feldman’s categories of effective teaching behavior (Feldman, 2007). The professor level reliabilities of the seven dimensions were found to be good. In addition, confirmatory two-level factor analysis confirmed a seven dimensional factor structure on professor level on each measurement occasion.

Furthermore, the factor structure at the student level was analyzed with exploratory and confirmatory factor analysis. Results showed that students differed in their perception of classroom interaction and in their perception of the clarity, interest and importance of the subject matter. Finally, multilevel regression analyses revealed that specific teacher level factors and student level factors significantly predicted students’ perception of their learning outcomes. These results supported the proposed theoretical framework concerning the relationship between the ISQ teaching dimensions and the student learning process, thus providing evidence of the validity of the instrument. In addition, these findings showed that professors have a direct influence on how useful a lecture actually is, in terms of students’ perceptions of their learning outcomes.

In sum, the content validity, internal structure, construct validity, and reliability of the ISQ teaching dimensions were confirmed in this chapter. Thus, I conclude that the instrument provided reliable and valid ratings on professors’ lecturing skills from a wide variety of departments at this university, on multiple measurement occasions. In the context of this dissertation, these findings are relevant as they support the reliability and validity of the findings on the fourth level of evaluation (behavioral level), and hence the quality of the intermediate feedback provided to professors in the two experimental interventions.
Effects of the interventions on evaluation level one and two

Chapter 4 addressed the effects, relative to the control condition, of intermediate feedback only and intermediate feedback plus consultation in terms of Guskey’s evaluation level one and two, i.e., professors’ self-reported satisfaction with the interventions, and professors’ self-reported learning on lecturing due to the interventions.

With respect to the first level of evaluation (satisfaction), professors in all three conditions were positive about the lecture evaluations, and stated that they would recommend them to their colleagues, particularly to their junior colleagues. In terms of mean ratings, professors in the feedback-plus-consultation condition were most satisfied with the lecture evaluations and the intervention they received. They reported to be satisfied with the consultation itself as well. They stated that they would recommend consultation to both junior and senior colleagues. In terms of comparisons with professors in the control condition, they considered the lecture evaluations to be significantly more useful to improve their teaching. The feedback-only condition did not differ from the control condition with respect to perceived usefulness of the lecture evaluations, even though professors in the feedback-only condition received the feedback between the rated lectures, while professors in the control condition received the feedback at the end of the course.

On the second level of evaluation (self-reported learning), results showed significant differences between the conditions ($p < .01$) on twenty-five out of forty-four outcome variables. Comparisons between each experimental condition versus the control condition showed that nearly all significant differences were due to differences between the feedback-plus-consultation condition and the control condition. In this comparison, significant differences with a large effect (Cohen’s $d > .80$) were found on twenty dependent variables. Most effects concerned variables related to professors self-reported gained knowledge and an increased focus of attention to various teaching phases and teaching dimensions. In addition, professors in the feedback-plus-consultation condition made more plans for improvement of teaching and learned more from the program that they followed, compared to the professors in the other two conditions. In contrast, professors in the feedback-only condition differed significantly from their colleagues in the control condition on only one dependent variable: they gained more knowledge on how students experienced their lectures. In short, according to professors themselves, intermediate feedback-plus-consultation had a significant impact on the learning level (level 2), while intermediate feedback-only had little effect.
Effects of the interventions on evaluation level four and five

The same difference in impact was found with respect to evaluation levels four and five, i.e., teaching behavior (Chapter 3 and 5) and student learning (Chapter 5), both as perceived by the students. With respect to the behavioral level (level 4), the feedback-only condition had no significant impact on any of the seven teaching dimensions, compared to the control condition. In both the first and the second experiment, there was a significant effect of the feedback-plus-consultation condition on total mean student ratings of professors’ lecturing behavior (Total Instructional Skills), compared to the control condition. In the first experiment, there were additional significant effects on the teaching dimensions Explication, Comprehension, and Activation, in the feedback-plus-consultation condition, in comparison to the control condition. In the second experiment, there were additional significant effects on Structure, Validation, and Instruction, in the feedback-plus-consultation condition, compared to the control condition. In terms of Cohen’s $d$ and effect sizes based on the multilevel output, the effects found in the first experiment were medium to large, and those in the second experiment were small to medium.

In the second experiment, the effects on students’ perceptions of their learning outcomes were investigated. In the feedback-plus-consultation condition, students’ ratings expressing how much they reported to have learned from the lecture (variable Cognition) increased significantly over time, compared to the students’ ratings in control condition. In terms of Cohen’s $d$, the effect size was medium. The multilevel effect size was small (I will discuss these differences in effect sizes in the section on the scientific contribution of the findings). Again, feedback-only had no significant impact on this level of evaluation.

Generalization of the findings

The sample of the second experiment included professors from a wide variety of departments of the University of Amsterdam, who differed in age, rank, experience, course level, and class size. Given the diversity of the sample and the randomized block design (controlling for prior quality of teaching and department), the results should generalize to professors at this university.

Additionally, the influence of professors’ age, prior quality of teaching, and class size on the findings was investigated with exploratory analyses. On the learning level of evaluation (level 2), there was one significant main effect of professors’ age on the dependent variable improved skills on teaching, indicating that in all three conditions younger professors reported more improvement than older professors. On the fourth and fifth levels of evaluation, main effects of professors’ age, prior quality of teaching, and class size were found on various teaching and learning dimensions. These findings indicated that, according to student
ratings, in all three conditions older professors were less effective in their teaching, high quality teachers were more effective in their teaching, and professors teaching larger classes were less effective in their teaching (particularly on dimensions, which involve interaction with students; i.e., Comprehension and Activation).

Furthermore, results showed that professors’ age and class size did not influence the effects of the interventions on professor and student ratings at any of the four levels of evaluation. Thus, the effects found generalize to professors from various ages and with a wide variety of class sizes. The third moderator, professors’ prior quality of teaching, did influence the results. This influence is discussed in the next section.

**Summary**

In summary, feedback-plus-consultation had a considerable impact on all four levels of evaluation, in comparison to the control condition. In contrast, feedback-only had little to no significant impact on the four levels, compared to the control condition. The present results are consistent with findings of reviews on the general effects of intermediate feedback-only and feedback-plus-consultation on student ratings (Cohen, 1980; Menges & Brinko, 1986; Penny & Coe, 2004). Specifically, findings in these reviews indicated that the effects of intermediate feedback-only on students’ total mean ratings are generally small, while the effects of feedback-plus-consultation are medium to large (i.e., in terms of Cohen’s $d$). Importantly, the present results complement these previous findings by providing insight into the detailed impact of intermediate feedback and consultation on four levels of evaluation. Furthermore, these findings shed light on the process of achieving results on the highest levels of evaluation, particularly when the results on other additional exploratory analyses are taken into account. I elaborate on this in the next section.

**The process of improving teaching effectiveness**

The process of achieving results with the two interventions, in terms of an increase in student ratings, is quite demanding, as it comprises the following stages. First, professors have to be willing and to make time to act immediately on the intermediate feedback (and consultation). Second, professors have to interpret the ratings carefully, reflect on their current teaching behavior, and come up with new strategies to improve their teaching (if that is indicated). Third, new planned teaching behavior needs to be implemented and executed successfully. Fourth, the professors’ efforts as a whole should have effects, that is, result in an increase in subsequent student ratings on the professors’ lecturing skills and on students’ self-perceived learning.
The present findings on the limited impact of feedback-only on the learning level (level 2) suggest that the process in this condition already stagnates at the first or second stage. Reflecting on these first two stages of the process, Theall and Franklin (2001) found that student ratings are often misinterpreted, misused, or not used at all. Additionally, McKeachie (1997) pointed out that when professors perceive the ratings to be low, this may have a negative effect on their motivation. Arthur (2009) investigated professors’ responses to negative student feedback and distinguished four possible reactions: shame (It’s my fault and I can’t do anything about it), blame (It’s their fault and I can’t do anything about it), tame (It’s about them, but I can respond to their needs) and reframe (It’s to do with me, but I can learn and develop as a result). Only ‘tame’ and ‘reframe’ result in positive changes to teaching behavior. These findings in the literature provide an explanation for the limited effects found in the feedback-only condition; when student ratings are misinterpreted, misused, demotivating, or not used at all, limited effects occur on the learning level of evaluation. McKeachie (1997) therefore concluded that part of the validity of student ratings is in its use. Even though the results in this dissertation may only be generalized to the professors at the University of Amsterdam, these findings suggest that the efforts undertaken at many universities to provide professors with feedback to improve their teaching require supplemental support.

Considering the process of achieving results in the feedback-plus-consultation condition, the impact of the intervention was large on the learning level (level 2) and smaller on the behavioral level (level 4) (as assessed by students). Consistently, fewer significant effects were found on professors’ self-assessed improvement in skills on the teaching dimensions (items starting with “I became better at...”). As most effects were found in the areas of increased knowledge, focus of attention, and plans made for improvement, the process of achieving results on the behavioral level seems to stagnate in implementing and executing new planned teaching behavior successfully.

One explanation of these findings is that the interventions and measurements of the effects took place in a relatively short period of time (generally courses lasted eight weeks). Some planned improvement (like activating students during the lecture) may require several lectures to implement successfully, and major changes (like reducing the amount of subject matter discussed in the lectures) in the course or lectures cannot always be achieved during the current course. Guskey (2000) noted that the most worthwhile changes in education require time for adaption, adjustment, and refinement. Some findings in other studies support this statement. For example, Piccinin, Cristi and McCoy (1999) found a delayed effect, in terms of an increase of course ratings, one to three years after the initial SET consultation.
Additionally, the exploratory analyses in this study on the moderating effects of professors’ prior quality of teaching revealed that high quality professors made more improvement on the behavioral level (level 4), according to students, compared to medium quality professors. At the learning level (level 2), high and medium quality professors did not differ in terms of self-reported effects (which were large at this level).

Again, this sheds light on the process of achieving improvements. Apparently, high quality professors are able to successfully implement and execute new planned teaching behavior within the time span of the course. According to McAlpine and Weston (2000), high quality professors tend to be highly reflective on their students’ learning process and their own teaching behavior. They may therefore be more skilled in experimenting with their teaching behavior successfully within a relatively short time frame. Medium quality professors might need more time to successfully improve their teaching effectiveness, as perceived by students. Marsh and Roche (1993) support this suggestion. These authors found no effect of mid-term SET consultation on the first and the second semester ratings, but they did find an interaction effect between professors’ baseline quality of teaching and improvement in ratings at the end of the second semester. Their findings indicated that professors, who were initially less effective, benefited from intermediate SET consultation in the long run. Considering these previous findings and the current effects found on professors’ self-reported learning, further research on the long term effects of the current approach to SET consultation is justified and necessary.

An alternative explanation of the limited results on the behavioral level (level 4) is that the findings presented thus far on the effects of consultation on student ratings were somewhat biased. During the consultation meetings professors targeted only a few teaching dimensions for improvement. The effects of consultation presented thus far on each teaching dimension are based on ratings of professors, who did and did not target the specific dimension. Therefore, in additional exploratory analyses, the effects on targeted dimensions were separated from effects on non-targeted dimensions (on the fourth level of evaluation). Compared to the findings above, in both experiments, results showed additional significant effects ($p < .01$) on teaching dimensions that were targeted for improvement. In the second experiment significant effects of targeted dimensions were found on six out of seven teaching dimensions in the first time interval. In contrast, non-targeted dimensions did not improve significantly on any of the seven dimensions on this time interval, compared to the control condition. These findings are consistent with previous findings by Marsh and Roche (1993), and reveal a more comprehensive impact of feedback-plus-consultation on the behavior level. Furthermore, these findings indicate that the effects are due to the consultation approach, rather than to a Hawthorne effect (the attention/social treatment one receives).
Finally, I note that exploratory analyses on each time-interval revealed that in both experiments the effects of feedback-plus-consultation mainly occurred in the first time-interval (due to the first consultation meeting). In the pilot study, two additional consultation meetings took place in between the rated lectures. In the second experiment, one additional consultation took place in between the rated lectures. These additional consultations had fewer effects or no effects at all. I therefore conclude that only the first intermediate consult results in appreciable effects.

Summary

In summary, the confirmatory and exploratory analyses in this dissertation provide insight in the process of improving teaching effectiveness. In the feedback-only condition, the process tended to stagnate in the early stages, resulting in only one significant effect (of forty-four dependent variables) on the learning level, and no effects on the behavior level, or student learning level (levels 2, 4 and 5). In the feedback-plus-consultation condition, the process tended to stagnate later on in the process, with the specific consequence of large significant effects on twenty out of forty-four dependent variables in the learning level (level 2), and small to medium significant effects on four out of eight dependent variables at the behavior level, and one out of three dependent variables at the student learning level (level 4 and 5). During the course, high quality professors (and their students) benefited most from the intervention with consultation. Medium quality professors may benefit more over a longer period of time, but this requires further research. Finally, targeting dimensions (with a collaborative approach to consultation) displayed significant effects on more dependent variables at the behavior level, particularly in the first consultation meeting. After the first consultation meeting, additional consultation meetings during the course appear to have little effect and may well be superfluous.

Limitations

As with all research, the investigations in this dissertation have their limitations. Despite the efforts undertaken to ensure the validity of the findings, some possible threats to validity remain. I address these in this section.

First, I note that the effects on professors’ learning and students’ learning (level 2 and 5) were based on professors’ and students’ self-reports only. No objective measures, like a summative assessment of learning, were used. Self-reports are open to socially desirable responding. Nonetheless, professors’ and students’ self-reported ratings differed between the
outcome variables, and there were significant differences in self-reported ratings between conditions, and differences within conditions on various outcome variables. This variation provides some support for the internal validity of the findings. Additionally, the differences in effects between the two interventions were consistent with findings on other levels of evaluation.

Second, the effects on professors’ lecturing behavior were evaluated by means of student ratings. Notwithstanding, the efforts undertaken to ensure valid and reliable findings, one may still ask to what extent the student ratings reflected the full effects of the interventions on professors’ lecturing behavior. Even though students were instructed to evaluate the specific lecture they had just attended, some students might still have based their evaluation on a general impression they had of their professor, due to evaluation fatigue, haste or maturation. L’Hommedieu and colleagues (1990) discussed previous findings on the stability of student ratings collected at different times in the instructional sequence. They quoted Rotem and Glasman (1979) who wrote “such stability, however, should also raise questions with regard to student’s sensitivity to changes that may occur during the interval” (p. 506). In this light, actual changes during the course, due to the interventions, may not fully be detected with student ratings only. In practice, evaluation fatigue is less likely to occur with more occasional use of these interventions. Furthermore, results have shown that repeated SET consultation during a course is not necessary. Only the first SET consultation resulted in appreciable effects. To prevent maturation of the rater (i.e., forming estimates of teacher effectiveness on the basis of their early impressions, see L’Hommedieu et al., 1990), I suggest that SET consultation takes place in the beginning of the course. This is also important considering potential student dropout during the course. This brings me to the final issue to address.

In each condition, courses incurred student dropout. In the second experiment, analyses showed that students, who completed the ISQ twice or three times, rated their professors significantly higher on the first measurement occasion, compared to students who rated the professor only once (i.e., on the first measurement occasion). Thus, as students drop out during the course, ratings might stay high artificially. Analyses, of the data of students, who completed the ISQ on the first measurement occasion, showed that more of these students drop out in the control condition, compared to the experimental conditions. Ratings in the control condition might therefore be slightly biased, in terms of more positive on the second and third measurement occasion, compared to the experimental conditions. I note that this difference between the control condition and the experimental conditions might also be due to the interventions. The resulting improvement in teaching, attributable to the interventions,
may have decreased dropout. At least, several findings in this dissertation indicate that the interventions also prevented a decrease in ratings over time.

In sum, these limitations complicate the assessment of the full impact of the two interventions, as compared to the control condition. When student ratings stay high artificially over time due to student dropout, or when students do not detect all changes, the analyses of intermediate student ratings may result in modest effects. Other researchers, who evaluated teaching effectiveness, recommended the use of multiple sources of data to assess teaching quality (Benton & Cashin, 2012). I therefore suggest future research to complement these findings from student ratings with additional measures of teaching effectiveness, such as classroom observations.

**Scientific contribution of the findings**

One of the main contributions of this dissertation is that it complements previous non-experimental findings on the effectiveness of collaborative consultation with experimental results. Reviewers have addressed important limitations of previous studies on the effects of both interventions, such as the use of small and/or selected samples, lack of a control condition, lack of random assignment, and/or control for moderating variables, lack of thorough investigation on the psychometric quality of the instruments used, and investigation limited to only one level of evaluation (Levinson-Rose & Menges, 1981; Prebble et al., 2004; Steinert et al., 2006; Stes, Min-Leliveld, Gijbels, & Van Petegem, 2010; Weimer & Lenze, 1997). In addition, l’Hommedieu and colleagues (1990) provided multiple recommendations for research in this field, such as consideration for the appropriate unit of analysis and use of comparable measures. In this dissertation these limitations and recommendations have been taken into account.

A second important scientific contribution of this dissertation is the use of multilevel analyses on the student ratings data. Chapter 2 provided an illustration of the use of exploratory and confirmatory factor analyses on a student ratings instrument on both the professor level and the student level. Additionally, it provided new insights into the classroom dynamics that characterize university lectures. Finally, it validated an instrument to evaluate single lectures and/or investigate differences between professors, as well as differences between students within classes.

The investigations with multilevel analyses on the effects the interventions showed that significant random intercept and slope effects were present at both professor and student level, meaning that mean ratings of professors differed significantly at baseline, and mean
ratings of individual professors varied significantly between lectures. Additionally, ratings of students within classes varied significantly at baseline and between lectures over time. Chapter 5 illustrates the importance of taking these random effects into account; without a random slope on the professor and student level, effects of the feedback-only condition were significant on four dependent teaching variables, and effects of the feedback-plus-consultation condition were significant on all seven dependent teaching variables plus two learning outcome variables. The inclusion of a random slope rendered all effects of the feedback-only condition insignificant. The effects of the feedback-plus-consultation condition remained significant on four out of seven teaching variables and on one learning outcome variable. The effects of consultation were established in the presence of the random differences between professors and students.

Effect sizes based on the multilevel output were often smaller than effect sizes calculated with Cohen's $d$. In the past, this multilevel analysis was poorly disseminated in terms of user-friendly software. As previous findings are therefore often solely based on Cohen's $d$ effect sizes and ANOVA or single level regression analyses, possibly the effects found in these previous studies are somewhat overestimated. Therefore, I urge future studies to make use of multilevel analyses on student ratings data.

In summary, the present dissertation complements previous results with experimental findings, adds new findings, provides a new reliable and valid student ratings instrument, and illustrates the use of modern statistical approaches to investigate the internal structure of the instrument and effects on student ratings data.

**Practical implications of the findings:**

To use or not to use intermediate student feedback with or without consultation in instructional development practices?

At first sight, the answer to this question is clear: intermediate feedback only had little to no significant impact on the four levels of evaluation investigated in this dissertation, compared to the control condition. On the other hand, combining intermediate feedback with consultation had a considerable impact on all four levels of evaluation, compared to the control condition. Professors in the feedback-plus-consultation condition found the lecture evaluations more useful to the improvement of their teaching, compared to the other conditions, and they recommended SET consultation to both junior and senior colleagues. They reported to have learned more on various teaching dimensions and teaching phases. Their students perceived improvement on various teaching dimensions, and reported to have learned more during
the lectures. Thus, at first sight, it is a clear no to intermediate feedback only and a clear yes to combining feedback with consultation.

However a closer look at the intermediate feedback only condition revealed that professors in this condition appreciated the lecture evaluations, and would recommend this form of feedback to their junior colleagues. Also, it did increase professors’ knowledge on how students perceived their lectures, and it enabled high quality professors to maintain high ratings during the course. Thus, intermediate feedback based on a specific questionnaire like the ISQ may be useful to inform professors how students perceive their lectures. However, given the effort on the part of the students, one may question the cost-effectiveness of continuously providing intermediate feedback in addition to end-of-the-course evaluations.

With respect to feedback coupled with consultation, Penny and Coe (2004) detected larger effects on the behavior level (level 4) with more extensive interventions. Like Penny and Coe, I find the use of SET consultation recommendable, but advise the use of additional sources of feedback, such as classroom observation or videotaping (also to observe the full impact of the intervention).

Given the impact of SET consultation on the learning level (level 2), the intervention is also useful as a supplement to other instructional development activities (such as seminars and workshops on teaching matters). In their review, Stes and colleagues (2009) found that, compared to a collective course in isolation, a collective course combined with an alternative form of instructional development often had more impact on the teacher behavioral level. The findings by Stes and colleagues indicate that seminars and workshops often suffer from a lack of ‘transfer of training’ (see Baldwin & Ford, 1988) to professors own teaching practices. Intermediate SET consultation, with a collaborative approach to consultation, encourages reflection ‘in action’ and ‘on action’ (see Schön, 1987) and may help overcome this issue. Based on previous effect studies in the literature, Lenze (1996) identified consultation as an instructional development strategy preferable to other approaches, such as workshops, grants for instructional improvement, advice from colleagues, and provision of resource materials. Based on differences in content of these strategies, I contend that SET consultation is useful in conjunction with other strategies. For example, workshops and resource materials help educate professors on pedagogical principles that serve to facilitate the students’ learning process (Prebble et al., 2004).

In terms of procedures, more than one consultation session during the course appeared to be redundant. This is relevant to ultimate cost-benefit analyses. The first consultation had most (and appreciable) effects, particularly when dimensions were targeted for improvement with a collaborative approach to consultation. The effects of a collaborative consultation
approach exceed the effects found in previous studies on a more diagnostic consultation approach (see Penny & Coe, 2004), in which consultants are the ones who interpret the student ratings and provide recommendations for improvement. A follow-up SET consultation in the next course or semester might be important to maintain effects. Stes and colleagues (2009) found that instructional development interventions spread out over time have more positive behavioral outcomes than one-time events.

Thus, should all professors be provided with at least one intermediate SET consultation? Considering the impact on all levels of evaluation, the answer was yes, surely in combination with other forms of formative assessments and/or educational activities. Considering the costs, however, instructional interventions are often provided mainly to professors, who appear to be less effective in their teaching. I believe this is a mistake. As I stated in the introduction of this dissertation, the educational training of university professors is extremely limited compared to those of their colleagues in primary and secondary education. At the same time, increased importance of accountability on the quality of teaching at universities puts pressure on the university’s administration and on individual professors (who are expected to be a professional researcher, as well as a professional teacher). This calls for a shift towards a more supportive teaching culture at universities. A supportive teaching culture is not only necessary to improve teaching effectiveness (if required), but also to maintain excellence in teaching and to promote faculty motivation (Feldman & Paulsen, 1999). In this context, merely providing professors with student ratings of their teaching effectiveness, even intermediate student ratings, is clearly not sufficient. A basic teacher evaluation system should at least be accompanied by some sort of support system, such as a system with information on how to interpret the ratings and possible strategies for improvement, and the possibility for peers to easily exchange effective teaching strategies. This would be a first small step towards a supportive teaching culture.

In the Netherlands, the recent instatement of a teaching certificate (the BKO) for all teaching staff members at universities is a second, substantially larger, step towards such a culture. In addition, Feldman and Paulsen (1999) identified eight characteristics of universities and colleges, which reflect a supportive teaching culture. Among these are: a) an administrative commitment and support by giving high visibility and support to instructional activities and sufficient rewards to effective professors; b) professors’ involvement, shared values, and a sense of ownership in planning and implementing activities that encourage instructional excellence and improvement; c) a faculty development program or campus teaching center; and d) frequent interaction, collaboration, and community among faculty to improve teaching effectiveness, increase intellectual stimulation, and reduce the degree of isolation associated with traditional teaching at universities.
Some of the consultants in the present investigation were trained professors. If costs are an issue in implementing SET consultation by an external consultant, it makes sense to train peers to do the job as part of their own professional development. Handal (1999) makes a striking comparison with traditions in academic culture related to quality assurance and professional development in research:

“Criticizing other researchers’ reports and publications is an accepted activity. It is carried out by means of comprehensive refereeing procedures in the case of scientific and professional publications and conferences. Another ritualized example is the thesis defense, a key element in the evaluation and approval of graduate degrees… Providing criticism is one of the skills that scholars within the university system must develop to gain recognition as competent members of the academic profession. …[When providing criticism] we usually learn a lot ourselves. We get new ideas, become acquainted with fresh research, and are made aware of different perspectives and methods… I believe that we lack corresponding traditions in academic culture when it comes to teaching. Educators engage relatively rarely in systematic appraisal of their colleagues’ teaching… University teaching is more or less the private property of the individual instructor, and any commentary could be construed as meddling.” (Handal, 1999, p. 59-65).

Professors’ teaching activities demand their own professional development and quality assurance, like professionalism in professors’ research activities, with similar standards and traditions. From this perspective, training professors to serve as collaborative consultants would not only be beneficial in terms of costs, it would also serve a supportive and professional teaching culture.

In addition, the present findings show that high quality professors (as well as their students) benefit from SET consultation as well. I contend that, if faculty development practices, such as (expert or peer) consultation and extensive feedback, are reserved for relatively ineffective professors, the professionalism and responsibility of university professors as educators is highly underestimated. Thus, I suggest that feedback combined with consultation is made available to professors regardless of their teaching effectiveness.

To end, the following famous quote is often used in the educational field:

“Who dares to teach, must never cease to learn” (John Cotton Dana).

Let me close this dissertation by complementing this quote:

“Who dares to expect professors to teach, must never cease to support them in their learning”.
References


