Improving University Lectures with Feedback and Consultation

Mariska Knol

Improving University Lectures with Feedback and Consultation

Mariska Knol

Cover	Mariska Knol
Layout	Renate Siebes, Proefschrift.nu
Printed by	Ipskamp Drukkers B.V.
ISBN	978-94-6191-683-9

Improving University Lectures with Feedback and Consultation

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. D.C. van den Boom
ten overstaan van een door het college voor promoties
ingestelde commissie,
in het openbaar te verdedigen in de Aula der Universiteit
op vrijdag 19 april 2013, te 13.30 uur

door

Mariska Henriëtte Knol

geboren te Eindhoven

Promotiecommissie

Promotor	prof. dr. H.L.J. van der Maas
	prof. dr. J.H. van Driel
Copromotor	prof. dr. G.J. Mellenbergh
Overige leden	prof. dr. D. Borsboom
	prof. dr. J.F.M.J. van Hout
	prof. dr. F.J. Oort
	prof. dr. M.L.L. Volman
	prof. dr. M. Brekelmans
	prof. dr. D. Beijaard

Faculteit der Maatschappij- en Gedragswetenschappen

Contents

1

Introduction

Faculty development is hot, research on faculty development is not

Accountability at universities, towards the end of promoting excellence in research and education, has become more and more important over the years. Enders, De Boer and Weyer (2012) provided an overview of the current situation of universities in the Netherlands. They state that at present roughly two-third of the budget of Dutch universities derives from the government. Universities are seriously dependent on this income stream that comes with strings attached as regards funding by student numbers. Dutch universities are required to have an internal and external evaluation system for both teaching and research. In addition, other internal and external accountability requirements, e.g., quality assurance schemes, accreditation, and performance monitoring, have been introduced over the years (Enders, De Boer & Weyer, 2012).

To excel in research, university professors (meaning part-time and full-time assistant, associate and full professors) start with extensive training through PhD projects (in the Netherlands, a PhD position is considered a proper salaried job). At the same time, the educational training of university professors is extremely limited. This state of affairs is similar to that in other countries. In an opinion piece in the International Journal for Academic Development, Baume (2006) stated that teaching in higher education may be one of the last non-professions. Unlike their colleagues in elementary and secondary education, most professors have little formal training in teaching. Mostly, they rely on their own past experience as teachers, and on the examples set by their professors when they themselves were students. University teaching therefore resembles on-the-job training, which often takes place in isolation, with little help, and no resources (Knapper & Piccinin, 1999). In the Netherlands, teachers are required to follow four years of teacher education to be able to teach in primary or secondary education. Master graduates are able to teach at secondary education after one additional year of teacher training. In contrast, to be able to teach at universities, one basically needs only to be a research expert on the subject matter.

At all educational levels, educational researchers agree that teaching not only requires subject matter knowledge, but also pedagogical content knowledge (knowledge on how to teach the subject matter, see Shulman, 1986) and teaching skills that distinguish teachers from subject matter specialists. Ramsden (2002) defined the activity of teaching in higher education to include the objectives of the curriculum, the pedagogies of conveying the knowledge that these objectives embody, the assessment of students, and the evaluation of the effectiveness of the teaching.

Thus, to be a subject matter specialist does not guarantee good teaching. Of course, some subject matter specialists at universities do excel in the quality of their teaching. In an

extensive meta-analysis based on 58 articles, Hattie and Marsh (1996) showed no correlation between the quality of research and the quality of teaching of university professors. Regardless discipline, professors come in all different flavors; good at teaching and bad at research, good at teaching and good at research, bad at teaching and good at research, and bad at both. Related to this, Handal (1999) speaks of "dual professionalism", and Baume (2006) noted: "I hope that, soon, teaching in higher education will be recognized for its difficulty, its importance and the great extent of its responsibility" (p. 58).

In response to lack in formal preparation for teaching, faculty development centers have been created, starting in the 1970s, to support and improve university teaching. In addition, the presidents of all Dutch universities signed an agreement in 2008 on the instatement of a basic quality of teaching certificate for university professors. The certificate concerns new standards of teaching, and is recognized at all fourteen Dutch universities (VSNU, 2008). In other Western countries, similar initiatives have been taken, e.g., the UK introduced a teaching qualification which meets new standards of teaching, and needs to be obtained by all new teaching staff from 2006 (DfES, 2003, cited in Baume, 2006).

In addition to these developments, studies on effective interventions to improve the quality of university teaching have become equally important. Unfortunately to date, research funding on (effective interventions in) higher education has been limited. For example, the Dutch Programme Council for Educational Research (PROO) funds educational research on primary education, general secondary education, (pre-) vocational education and teacher training institutes (training primary and secondary education teachers). However, research on higher education is excluded (PROO, 2012). As a result, the effectiveness of faculty development practices is seldom investigated thoroughly. In other western countries, research grants have been allocated to this matter, but, compared to other educational fields, this field of research is largely neglected. In each review of the literature on the effectiveness of faculty development practices, authors stressed the importance of more research, and specifically more experimental research in this field (Levinson-Rose & Menges, 1981; Prebble et al., 2004; Steinert et al., 2006; Stes, Min-Leliveld, Gijbels, & Van Petegem, 2010; Weimer & Lenze, 1997). Each of these reviews concluded that many studies in this field are characterized by important limitations. Some of the most important limitations are the following.

First, studies are often limited to small and/or selected samples. For example, participants are often professors who approached faculty development centers with the aim to improve their teaching, and were therefore often well motivated to change from the start. In most cases, results show positive effects, but whether the effects are due to the intervention, the initial motivation, or simply time on task remains unknown. Second, the evaluations

are generally limited to measures of satisfaction of the participants. Third, participants are often not randomly assigned to various interventions, leaving the specific and relative effects unclear. Fourth, effect studies often lack a control condition, leaving the relative effects compared to no intervention unclear. Fifth, studies often lack thorough research on the psychometric quality of the instruments used to evaluate improvements.

In summary, with the increasing acknowledgement of (and investment in) university teaching, additional experimental research on faculty development practices on multiple levels of evaluation has become indispensible. Below, the various levels of evaluation are discussed.

Considering levels of evaluation

Kirkpatrick (1994) distinguished four levels of evaluation of training programs in business and industry: reaction, learning, behavior, and results. The reaction level of evaluation concerns participants' satisfaction with the program. The learning level concerns the knowledge, attitudes, and skills that participants acquire as a result of the program. The behavioral level concerns participants' behavioral changes on the job, due to the program, and the result level concerns the effects of the program on the organization. Guskey (2000) adapted Kirkpatrick's evaluation model to the educational field, specifically to evaluate the professional development of teachers. Guskey's five-level evaluation model comprises participants' reactions (level 1), participant's learning (level 2), organizational support and change (level 3), participant's application of new knowledge and skills (level 4), and student learning outcomes (level 5). Both models imply a hierarchic arrangement of levels, from simple to more complex, whereas each higher level builds on the preceding levels (Guskey, 2000). One example of more extensive quasi-experimental research, containing pre- and post-tests, control group comparison, and multiple levels of evaluation, are the studies of Stes and colleagues (2010, 2011). They found significant effects of their instructional development program on the teachers' approaches to teaching (which relates to level 2 and 4) (Stes, Coertjens, & Van Petegem, 2010), but limited effects on students' approaches to their study (which relates to an alternative result on the student outcome level 5) (Stes, De Maeyer, Gijbels, & Van Petegem, 2011). Such studies stress the importance of evaluation on multiple levels of effects. With more knowledge on the actual impact on various levels, faculty development centers can target and combine the optimal interventions, corresponding to prior aims for improvement.

The present aims

The first aim in the present dissertation is to overcome the limitations in previous research on faculty development interventions, as stated above, in an investigation of the effects of two specific faculty development interventions on university professors' lectures (class meetings in which lecturing is the teaching format). Although lecturing is not the most popular teaching format used in education, I chose to focus on this format since it still constitutes a substantial, and often indispensible, part of regular teaching practices at universities (Lammers & Murphy, 2002). In addition, the format of lecturing is largely comparable over different departments, and so provides me with the opportunity to gather a substantial amount of data to investigate the effects of these interventions on professors' teaching behavior and students' self-reported learning.

The dissertation includes two experimental studies; a pilot study with twenty-five participants from a single department at the University of Amsterdam, and a larger study with seventy-five participants from a wide variety of departments at the same university. The participants were professors, who had not approached a faculty development center for support. Both studies included a control condition. In each experiment, professors were randomly assigned to the experimental and control conditions.

The second aim is to investigate the impact of the two interventions on Guskey's levels one, two, four, and five, that is, the effects on professors' self-reported satisfaction with the interventions (level 1), professors' self reported learning (level 2), professors' use of new knowledge and skills, measured by students' evaluations of lecturing (level 4), and students' self-assessed learning outcomes (level 5).

Below I present the rationale concerning the two faculty development interventions considered here and the main research question investigated in this dissertation. Next, I provide an outline of the dissertation. I end this dissertation with an overview and a discussion of the results.

Investigating student feedback and consultation

In the present research project, the objective was to investigate thoroughly the effects of student feedback provided to professors on their lectures, with and without additional individual consultations with professors. Aside from formal training programs and workshops, individual peer or expert consultation is one of the most commonly used interventions in faculty development, specifically instructional development with respect to small and larger classroom teaching (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 other approaches, such as workshops, grants for instructional improvement, advice from colleagues, and provision of resource materials. A common consultation procedure is for the consultant to clarify teaching goals, to encourage reflection about aims and methods, combined with some sort of feedback, to facilitate discussion on improvement strategies, and sometimes to conduct follow-up evaluation (Knapper & Piccinin, 1999; Penny & Coe, 2004). Feedback is gathered either through students' evaluations of teaching (e.g., Rindermann, Kohler & Meisenberg, 2007), or by more extensive means, such as classroom observations (e.g., Wilson, 1986; Piccinin, Cristy & McCoy, 1999), videotaping (e.g., Rozeman & Kerwin, 1991), or arranged student focus groups (e.g., Piccinin, Cristy & McCoy, 1999; Coffman, 1998). In this dissertation, I focus on combining students' evaluations of teaching (SETs) with individual consultation (SET consultation).

SETs are considered a potentially useful source of feedback (Prebble et al., 2004), and are valued as a formative feedback instrument by faculty members and faculty developers (Baxter, 1991; Schmelkin, Spencer & Gellman, 1997; Penny & Coe, 2004). Nowadays, collecting SETs at the end of the term or course has become common practice at universities worldwide. Unfortunately, despite the effort and despite its main purpose to provide faculty with feedback, collecting student feedback at the end of the term or course per se has little to no effect on teaching behavior (Hendry & Dean, 2002; Kember, Leung & Kwan, 2002; Marsh, 2007a). Providing professors with intermediate student feedback has some effect, in terms of an increase in SET ratings (Cohen, 1980; Menges & Brinko, 1989). Subsequently, augmenting SETs with individual consultation (SET consultation) has proven to be considerably more effective in various studies and meta-analyses (e.g. Cohen, 1980; Hampton & Reiser, 2004; Menges & Brinko, 1986; Marsh & Roche, 1993; Penny & Coe, 2004; Piccinin, Cristi & McCoy, 1999; Rindermann, Kohler & Meisenberg, 2007; Dresel & Rindermann, 2011). Consultation is considered an improvement to end of the term evaluations, since the latter often come too late to be of use, and generally come without any practical suggestions or support for change and improvement (Knapper & Piccinin, 1999; Penny & Coe, 2004).

I stated earlier that reviewers call for more in-depth research on the effects of faculty development interventions in general. In specific reviews on the effects of the two interventions investigated, intermediate SETs and SET consultation, this is also the case. One important finding, noted by these reviewers, is that the variation in effects of SET consultation is large (Menges & Brinko, 1986; Penny & Coe, 2004). The next step in this field of research is therefore to provide more insight into the effectiveness of particular approaches and procedures. In a meta-analysis, Penny and Coe (2004) studied the predictors of successful SET consultation, but

were limited in their research due to the small number of experimental studies. They noted: "Thus, the most robust finding may be that more research is needed" (p.236). In addition, they called for more research on SET consultation in settings other than North America, and replication of studies of various approaches to consultation. Furthermore, l'Hommedieu, Menges and Brinko (1990) provided a critical assessment of the limited effects of student feedback only. They focused on important methodological issues in previous research, and stated that the literature is hampered by pervasive threats to the internal and external validity of research findings. They stressed the need for further research to consolidate these findings. Some of their recommendations concern more adequate research on the instruments used, larger samples, sampling across subject areas and teacher characteristics, pre-tests, systematically assigned subjects and/or statistically controlling for moderating variables, studies of large lecture classes, consideration for the appropriate unit of analysis, and use of comparable measures (mid-term evaluation does not necessarily compare to end-of-the-term evaluation). In this dissertation these recommendations are taken into account.

The interventions investigated in this dissertation involve providing professors with SETs on their lectures with or without individual consultation during the course they are teaching. In the first experimental (pilot) study the aim was to investigate the effects of intermediate SETs with consultation. In the second experimental study the aim was to separate the effects of feedback and consultation by investigating intermediate SETs with and without consultation.

The main research question addressed in this dissertation is:

> What are the effects of intermediate student feedback with and without consultation on professors' self-reported satisfaction with the interventions (level 1), professors' self reported learning on lecturing (level 2), professors' lecturing skills, measured by students' evaluations of lecturing (level 4), and students' self-assessed learning outcomes (level 5)?

Based on the previous literature, I hypothesized that the effects of intermediate feedback on these levels of evaluation were small and the effects of intermediate feedback with consultation on these levels were medium to large.

The effects on student ratings data were investigated with multilevel regression analysis. This statistical approach accounts for the clustering in the data due to systematic differences between the lectures, the students, and the professors. The aim was to complement previous findings with new analyses, using this modern statistical approach.

Additionally, the effects on the teaching dimensions that were targeted for improvement during consultation were separated from the effects on non-targeted dimensions, to indicate whether the effects were due to the selected consultation approach or due to a Hawthorne

effect (i.e. due to the attention/social treatment one receives). Finally, the moderating effects of specific professor and course characteristics (i.e., professors' age, professors' prior quality of teaching and class size) were investigated. In doing so, I hope to contribute to the current body of knowledge in this field.

Outline of this dissertation

The chapters in this dissertation consist of accepted or submitted journal articles. Chapter 2 concerns an investigation on the psychometric quality of the instrument used to evaluate lectures, as assessed by students (the instrument measures various dimensions of lecturing skills), by means of confirmatory multilevel factor analysis. The chapter contains analyses on the construct validity, internal structure, and reliability of these teaching dimensions. These analyses are based on data collected in the second experiment of this dissertation. Furthermore, this chapter provides a theoretical framework on the relationship between the professors' lecturing behavior and the students' learning process. Test of these relationships are included in this chapter.

Chapter 3 provides a theoretical framework on the approach to consultation used in this study (collaborative consultation) based on theories on behavioral change. In addition, chapter 2 concerns a first experimental (pilot) study with twenty-five psychology professors, who were randomly assigned to either the experimental condition with SET consultation or the control condition with neither feedback nor consultation. The effects are studied in terms of changes in the professors SET results during the course (Guskey's level 4: changes in behavior, according to students). Chapter 2 (investigating the instrument) and three (piloting the procedure and approach to intermediate SET consultation) are considered to be a preparation for the investigations conducted in chapter 4 and 5.

Chapter 4 concerns a second experiment with seventy-five professors from a wide variety of departments, who were randomly assigned to one of three conditions, a feedback-only condition, a feedback-plus-consultation condition, or a control condition. This chapter concerns the effects of the second experiment on Guskey's level of evaluation 1 and 2; the effects on professors' self-reported satisfaction with the interventions, and professors' self reported learning.

Chapter 5 concerns the effects of the second experiment in terms of changes in the professors' SET results during the course (Guskey's level 4), and the effects in terms of changes in students' self-assessed learning outcomes (Guskey's level 5). The dissertation closes with a final discussion on the effectiveness of the chosen approach to intermediate feedback and intermediate feedback plus consultation.

References

Baume, D. (2006). Towards the end of the last non-professions? *International Journal for Academic Development, 11,* 57-60.

Baxter, E.P. (1991). The TEVAL experience, 1983-88: the impact of a student evaluation of teaching scheme on university teachers. *Studies in Higher Education, 16,* 151-179.

Coffman, S.J. (1998). Small group instructional evaluation across disciplines. *College Teaching, 46,* 106-111.

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

DfES (2003). *The future of higher education.* London: Department for Education and Skills.

Dresel, M., & Rindermann, H. (2011). Counseling university instructors based on student evaluations of their teaching effectiveness: a multilevel test of its effectiveness under consideration of bias and unfairness variables. *Research in Higher Education, 52,* 717-737.

Enders, J., De Boer, H., & Weyer, E. (2012). Regulatory autonomy and performance: the reform of higher education re-visited. *Higher Education*, DOI 10.1007/s10734-012-9578-4

Guskey, T.R. (2000). *Evaluating professional development.* Thousand Oaks, CA: Corwin Press.

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

Handal, G. (1999). Consultation using critical friends. In Knapper, C. & Piccinin, S. (Eds.), *Using Consultants to Improve Teaching. New Directions for Teaching and Learning, 79,* 59-70. San Francisco, CA: Jossey-Bass.

Hattie, J., & Marsh, H.W. (1996). The relationship between research and teaching—a meta-analysis. *Review of Educational Research, 66,* 507-542.

Hendry, G.D., & Dean, S.J. (2002). Accountability, evaluation of teaching and expertise in higher education. *International Journal of Academic Development, 7,* 75-82.

Kember, D., Leung, D.Y.P., & Kwan, K.P. (2002). Does the use of student feedback questionnaires improve the overall quality of teaching? *Assessment & Evaluation in Higher Education, 27,* 411-425.

Kirckpatrick, D.L. (1994). *Evaluating training programs: The four levels.* San Francisco, CA: Berret-Koehler Publishers.

Knapper, C., & Piccinin, S. (1999). Consultation about teaching: An overview. In Knapper, C. and Piccinin, S. (Eds), *Using consultants to improve teaching. New Directions for Teaching and Learning, 79,* 3-8. San Francisco, CA: Jossey-Bass.

Lammers, W.J., & Murphy, J.J. (2002). A profile of teaching teachniques used in the university classroom. *Active Learning in Higher Education, 3,* 54-67.

Lenze, L.F. (1996). Instructional development: What works? *National Education Association, Office of Higher Education Update, 2,* 1-4.

Levinson-Rose, J., & Menges, R.J. (1981). Improving college teaching: A critical review of research. *Review of Educational Research, 51,* 403-434.

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

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

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

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

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

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

Prebble, T., Hargraves, H., Leach, L., Naidoo, K., Suddaby, G., & Zepke, N. (2004). *Impact of student support services and academic development programmes on student outcomes in undergraduate tertiary study: A synthesis of the research.* Report to the Ministry of Education, Massey University College of Education.

PROO (2012). *Programma Onderwijsonderzoek 2012-2015.* http://www.nwo.nl/files.nsf/pages/NWOP_8RHJNZ/$file/Programmaonderwijs2012tm2015.pdf. Accessed January 2013.

Ramsden, P. (2002). *Learning to teach in higher education*. London: Routledge.

Rindermann, H., Kohler, J., & Meisenberg, G. (2007). Quality of instruction improved by evaluation and consultation of instructors. *International Journal for Academic Development, 12,* 73-85.

Rozeman, J.E., & Kerwin, M.A. (1991). Evaluating the effectiveness of a teaching consultation program on changing student ratings of teaching. *The Journal of Staff, Program & Organizational Development, 9,* 223-230.

Schmelkin, L.P., Spencer, K.J., & Gellman, E.S. (1997). Faculty perspectives on course and teacher evaluation. *Research in Higher Education, 38*, 575-592.

Shulman, L.S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*, 4-14.

Steinert, Y., Mann, K., Centeno, A., Dolmans, D., Spencer, J., Gelula, M., & Prideaux, D. (2006). A systematic review of faculty development initiatives designed to improve teaching effectiveness in medical education: BEME Guide No. 8. *Medical Teacher, 28,* 497-526.

Stes, A., De Maeyer, S., Gijbels, D., & Van Petegem, P. (2011). *Effects of teachers' instructional development on students' study approaches in higher education. Studies in Higher Education,* doi:10.1080/03075079.2011.562976

Stes, A., Coertjens, L., & Van Petegem, P. (2010). *Instructional development for teachers in higher education: impact on teaching approach. Higher Education, 60*, 187-204.

Stes, A., Min-Leliveld, M., Gijbels, D., & Van Petegem, P. (2010). The impact of instructional development in higher education: The state-of-the-art of the research. *Educational Research Review, 5,* 25-49.

VSNU (2008). Overeenkomst inzake wederzijdse erkenning basiskwalifikatie onderwijs. http://www.vsnu.nl/docentkwaliteit.html. Accessed January 2013.

Weimer, M., & Lenze, L.F. (1997). Instructional interventions: A review of the literature on efforts to improve instruction. In K. R. Perry, & J. C. Smart (Eds.), *Effective teaching in higher education: Research and practice*, 205-240. New York: Agathon Press.

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

2

Measuring the quality of individual lectures:

Multilevel factor analysis on the Instructional Skills Questionnaire (ISQ)

This chapter is an adapted version of:
Knol, M.H., Dolan, C.V., Mellenbergh, G.J., & Van der Maas, H.L.J. (submitted).
Measuring the quality of individual lectures: Multilevel factor analysis on the
Instructional Skills Questionnaire (ISQ).

Abstract

This study concerns the psychometric quality of the Instructional Skills Questionnaire (ISQ), a student evaluation of teaching questionnaire with specific questions on lecturing skills. It was developed to be used following a single lecture, to serve as a formative feedback instrument for university professors. The ISQ contains seven dimensions of instructional skills. Dutch students in 75 courses evaluated three 90-minute lectures of their respective professors with the ISQ. Confirmatory two-level factor analysis confirmed a seven dimensional factor structure on professor level on each measurement occasion. The professor level reliabilities of the seven dimensions were found to be good. In addition, the factor structure at the student level was analyzed. Results showed that students differed in their perception of classroom interaction and of the clarity, interest and importance of the subject matter. Specific professor level factors and student level factors significantly predicted students' perception of their learning outcomes. These results supported the proposed theoretical framework on the relationship between the ISQ teaching dimensions and the student learning process, thus providing evidence for the construct validity of the instrument. In sum, this study offers a reliable and valid instrument to evaluate single lectures.

Introduction

The purpose of the present study is to investigate the Instructional Skills Questionnaire (ISQ), a student evaluation of teaching instrument, which can be used to evaluate a single lecture at the university level. Common students' evaluations of teaching (SET) instruments are designed to evaluate complete courses in universities. Unfortunately, the impact of SETs collected at the end of the term on teaching behavior is small (Kember, Leung & Kwan, 1999; Marsh, 2007a). To be effective, feedback should be well timed, specific, reliable, and should target malleable behavior (McLaughlin & Pfeifer, 1988). Therefore, the present study concerns the development of a SET instrument that enables the evaluation of single lectures at the university level, by means of detailed questions on lecturing behavior. With this instrument, professors can be provided with more specific and relevant feedback on their teaching behavior during their course to improve the quality of their teaching.

Below, we first present the theoretical background of the ISQ dimensions of lecturing behavior and the expected relationships between the ISQ dimensions and the student learning process. Second, we present our results concerning the reliability and internal structure of the ISQ. Third, we explore differences between students in how they perceive/rate a lecture. Fourth, we present results on the relationship between teaching behavior and the students' perceptions of their learning outcomes, and the relationship between student differences in ratings and the students' perceptions of their learning outcomes, to validate the theoretical framework of the ISQ. Instead of common exploratory and confirmatory factor analysis, we used two-level factor analyses. This enabled us to investigate both professor level factors (differences between professors) and student level factors (differences between students). We conclude with a discussion on the implications for measuring and teaching lectures.

Dimensions of teaching

There is an extensive body of research concerning SETs (for an overview, see Marsh, 2007b; Richardson, 2005; Wright & Jenkins-Guarnieri, 2012). Outcomes of SETs have proven to be reliable and stable, reasonably valid, as judged by a variety of indicators of effective teaching, and relatively unbiased (Marsh, 2007b).

An important aspect of SETs is multidimensionality. Over the past thirty-five years, Feldman differentiated twenty to twenty-eight teaching dimensions, based on students' views on effective teaching, on SET ratings, and on content analyses of single items and multiple-item scales found in the higher education research literature (Feldman, 1976b, 1983, 1984, 1989a, 1989b, 2007). Feldman (2007) related these dimensions to domains of student

achievement and overall evaluations. He found the dimensions most highly related to both domains to be 1) teacher clarity and comprehensibility; 2) teacher stimulation of interest in the subject matter; 3) perceived outcome or impact of instruction; and 4) teachers' preparation (organization of the course). At the same time, many SET instruments have been developed using factor analysis or a theory based approach. The development of and research into three instruments formed the theoretical basis of the ISQ.

First, a thoroughly investigated instrument is the Students' Evaluation of Education Quality (SEEQ), developed by Marsh and colleagues (Marsh, 1984, 1987; Marsh & Hocevar, 1991b). The SEEQ includes nine dimensions of teaching effectiveness (Organization/Clarity, Breath of Coverage, Instructor Enthusiasm, Individual Rapport, Group Interaction, Workload/ Difficulty, Learning/Value). The reliability and validity of the SEEQ has been established in different settings (Marsh & Dunkin, 1992,1997; Marsh & Roche, 1993).

Second, De Neve and Janssen (1982) developed the Evalec (EVAluation of LECturing). The authors adopted a theoretical point of view, and focused on specific lecturing behaviors, which facilitate the student learning process. The five Evalec dimensions (Validating, Stimulating, Conversation, Directing and Structuring) were designed to address different lecture components according to Van Gelder's model for didactic analysis (Van Gelder, 1975). These components include introducing clear objectives, tuning in on the students entry level and interests, applying effective teaching-learning strategies (e.g., clear exposition, well-selected content, useful learning aids, eliciting discussions), and evaluating the outcome.

Third, based on the SEEQ, Evalec and the work of Feldman, Vorst and Van Engelenburg (1992) developed a Dutch course evaluation instrument for the University of Amsterdam, the Uvalon. The Uvalon included six dimensions on course characteristics, seven dimensions of teaching behavior, and two dimensions on student behavior. The psychometric quality of the Uvalon was investigated and confirmed in several internal reports from the University of Amsterdam (Vorst & Van Engelenburg, 1992; Verbeek, De Jong, & Vermeulen, 2002, 2005).

Due to the Uvalon's theoretical and empirical foundation, the ISQ was based on Uvalon's seven dimensions of teaching behavior (*Structure*, *Explication*, *Stimulation*, *Validation*, *Instruction*, *Conversation* and *Interaction*). Since the ISQ is meant to evaluate single lectures and to serve as a formative feedback instrument for university professors, we retained only the dimensions pertaining to teaching behavior. We renamed the dimensions *Conversation* and *Interaction* to *Comprehension* and *Activation*, respectively, to more accurately convey their meaning. We hypothesized that both dimensions reflect interaction between the professor and the students, but with different purposes. The items of *Conversation* dimension focus on providing occasion for students to ask questions, and for the professor to check whether

students understand the subject matter (hence the new label *Comprehension*). Thus the purpose is to have students and professors regulate the students' comprehension of the subject matter during class. The items of the *Interaction* dimension concern getting students involved and activated (hence the label *Activation*). The seven ISQ dimensions are defined as follow:

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

To indicate the relationship between the instruments, the dimensions of the SEEQ, Evalec, Uvalon and the ISQ are listed in Table 2.1. In addition, the relationship with Feldman's categories is indicated. In terms of content validity, the relationship with Feldman's categories and the other instruments show that the Uvalon / ISQ teaching dimensions contain the most important teaching behaviors.

Like De Neve and Janssen in the development of the Evalec, we take a theoretical point of view on the relationship between the selected teaching dimensions, and how they facilitate the student learning process. In the next paragraph, we propose a theoretical framework for the ISQ, which we test in the present study.

Table 2.1 Dimensions of the evaluation instruments SEEQ, Evalec, Uvalon, ISQ and the relationship with Feldman's categories

SEEQ (Marsh, 1982b, 1987)	**Evalec** (De Neve & Janssen, 1982)	**Uvalon** (Vorst & Van Engelenburg, 1992)	**ISQ**	**Feldman's Categories** (1976b, 2007)
		Teaching behavior		
Organization / Clarity	Structuring	Structure	Structure	Teacher's Preparation; Organization of the Course (I)
Breath of Coverage		Explication	Explication	Clarity and Understandableness (I) Teacher's Knowledge of Subject Matter (I) Teacher's Intellectual Expansiveness (I)
Instructor Enthousiasm	Stimulating	Stimulation	Stimulation	Teacher's Stimulation of Interest in the Course and Its Subject Matter (I) Teacher's Enthusiasm (for Subject or for Teaching) (I)
Learning / Value	Validating	Validation	Validation	Nature and Value of the Course Material (Including Its Usefulness and Relevance) (III)
Organization / Clarity	Directing	Instruction	Instruction	Clarity of Course Objectives and Requirements (III)
Individual Rapport	Conversation	Conversation	Regulation	Teacher's Sensitivity to, and Concern with, Class Level and Progress (II) Teacher's Availability and Helpfulness (II) Teacher's Concern and Respect for Students (II)
Group Interaction		Interaction	Activation	Intellectual Challenge and Encouragement of Independent Thought (by the Teacher and the Course) (II) Teacher's Encouragement of Questions and Discussion, and Openness to Opinions of Others (II)

Table 2.1 *Continued*

SEEQ (Marsh, 1982b, 1987)	Evalec (De Neve & Janssen, 1982)	Uvalon (Vorst & Van Engelenburg, 1992)	ISQ	Feldman's Categories (1976b, 2007)
		Course characteristics		
Learning / Value		Learning/Value		Perceived Outcome or Impact of Instruction (III)
Examinations / Grading		Examination		Teacher's Fairness; Impartiality of Evaluation of Students; Quality of Examinations (III)
				Nature Quality, and Frequency of Feedback from the Teacher to Students (III)
Assignments / Readings		Literature		Nature and Usefulness of Supplementary Materials and Teaching Aids (III)
Difficulty / Workload		Workload /Difficulty		Difficulty of the Course (and Workload) (III)
				Difficulty of the Content (and Workload) (III)
		Student characteristics		
		Entry level		
		Time invested		

Notes: Feldman (1976b) clustered the categories in Presentation (I), Facilitation (II), and Regulation (III). This is indicated in parentheses following each category. Categories not included are: Teacher's Elocutionary Skills, Personality Characteristics ("Personality") of the Teacher, Teacher Motivates Students to Do Their Best, Teacher's Encouragement of Self-Initiated Learning, Teacher's Productivity in Research Related Activities, Classroom Management, Pleasantness of Classroom Atmosphere, Individualization of Teaching, Teacher Pursued and/or Met Course Objectives.

Relating teaching behavior to student learning

Based on the literature, Vermunt (1996) distinguished three domains of activities relevant to students' learning: cognitive, affective, and regulative learning activities. Cognitive activities serve to process learning content (e.g., looking for relations among parts of the subject matter, thinking of examples). These lead directly to learning. Affective learning activities are directed at coping with the feelings that arise during learning, and lead to an emotional state that may positively, neutrally, or negatively affect the learning process (e.g., motivating oneself). Regulative learning activities are directed at regulating the cognitive and affective learning activities, and therefore indirectly facilitate learning results (e.g., orienting on a learning task). Vermunt and Verschaffel (2000) noted great similarities between these learning activities and teaching activities as found in the literature. They adopted the terms learning functions and teaching functions. *Learning functions* are categorized into cognitive/processing, affective, and regulation functions (parallel to the distinction between learning activities). *Teaching functions* refer to the functions that promote student learning. Cognitive/processing functions of teaching concern presenting and clarifying the subject matter. Affective functions refer to creating and maintaining a positive motivational and emotional climate for students. Regulation functions concern guiding students' learning processes (Vermunt & Verschaffel, 2000).

We propose that the teaching functions are closely related to the ISQ teaching dimensions. By structuring the lecture and the content (dimension *Structure*) and by explaining the subject matter clearly (dimension *Explication*), we hypothesize that the cognitive / processing learning activities are addressed. By interesting students in the subject matter (*Stimulation*), we hypothesize that affective learning activities are addressed. By telling students what is important to learn (*Instruction*) and why (*Validation*), we hypothesize that regulative learning activities are addressed. A representation of the hypothesized relationships between teaching behavior and student learning outcomes is given in Figure 2.1.

In the present study, we analyzed these relationships between the ISQ dimensions and student perceptions of their cognitive, affective and regulative learning outcomes. We interpret the results in terms of our theoretical framework and the construct validity of the ISQ.

Multilevel factor analysis of the ISQ

Normally the internal structure of a questionnaire is established by means of exploratory or confirmatory factor analysis. Here we used *two-level* exploratory and confirmatory factor analyses, as this provides us with the means to study the structure of the ISQ at the student

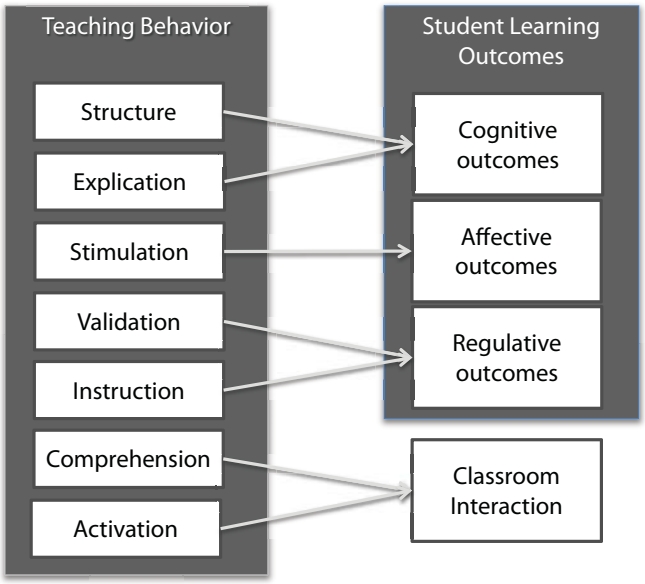

Figure 2.1 Theoretical framework on the relationship between the seven ISQ dimensions on teaching behavior and student learning outcomes.

level (level 1) and the professor level (level 2) using appropriate statistical modeling (Hox, 2002; Muthén & Muthén, 1998). Specifically, the level 2-factor model reflects the differences between the professors in the average responses of their students. The level 1 factor model reflects individual differences in students in their ratings of their professor. The professor level factor structure is of prime interest in any evaluation of professors. However, as the professor level data are based on student ratings, the ISQ necessarily provides student level data. The student level factor structure is of interest as it is at this level that student level variables, such as individual learning processes, are relevant. Specifically, we want to establish the structure of the student level ratings because this facilitates the evaluation of the relationship between students' perception of learning outcome variables and student level factors on the one hand, and the relationship between students' perception of learning outcome variables and professor level factors on the other. Finally, understanding the student level structure ultimately allows us to use the ISQ both as a professor level instrument (interpreting the mean ratings in terms of professor level latent variables that characterize individual differences of professors) and as a student level instrument (interpreting the students individual ratings in terms of latent variables that characterize individual differences

of students within a class). To our knowledge, MFA has not yet been applied to student evaluations of a single lecture (for an example on course evaluations, see Toland & De Ayala, 2005).

We had clear expectations concerning the factor structure on the professor level, and some conjectures concerning the student level factor structure. Since SET instruments are meant to differentiate between teachers on the given dimensions, the teacher-level is the appropriate unit of analysis when investigating the factor structure of SET instruments (Cranton & Smith, 1990; Gilmore, Kane, & Naccarato, 1978; Howard & Maxwell, 1980; Marsh, 1987, 2007b). Based on the theoretical background of the ISQ, we hypothesized that the seven-factor model provides an adequate account of the professor level covariance structure. We tested this hypothesis by means of confirmatory analysis. In addition, we investigated the student level with exploratory analysis, to obtain a better understanding of differences between students in how they perceive and rate a lecture. We chose to start with exploratory analyses followed by confirmatory analysis. Finally, we investigated the relationship between students' perceptions of their cognitive, affective, and regulative learning outcomes variables and the teacher and student level factor structures. In sum, the present study aims to address the following research questions:

1. What is the psychometric quality of the ISQ, in terms of internal structure and reliability of the subscales?
2. Are there structural differences between students in their rating of their professor? If so, what factor model provides an informative account of the student level covariance structure?
3. What are the relationships between students' perceptions of their cognitive, affective and regulative learning outcomes of the single lecture and the teacher and student level factors?

In this study, students rated three lectures per professor, which provided a dataset for each measurement occasion (denoted T_1, T_2, and T_3 below). The first measurement occasion was used to answer all three research questions. Measurement occasions two and three were used to (quasi) cross-validate the factor models.

Method

Participants

Professors

In total, 95 university professors from five departments of a Dutch university were scheduled to give a minimum of 3 lectures during a course in 2009-2010. From the 95 professors, 87 professors agreed to participate. Of the 87, 12 professors dropped out due to circumstance beyond their control (e.g., illness, rescheduling). This resulted in a final sample of 75 professors (63 male, 12 female, age $M = 46.8$, $SD = 9.6$) from the departments of Law ($N = 20$), Economics ($N = 24$), Science ($N = 13$), Social and Behavioral Sciences ($N = 13$), and Humanities ($N = 5$). Out of the 225 lectures (3 lectures per professor) that were scheduled to be rated by the students, 7 lectures were not rated by mistake. This resulted in 73 rated lectures on T_1, 74 rated lectures on T_2, and 71 rated lectures on T_3.

Students

The students in the selected courses rated their professors by completing the ISQ after three lectures during the course. In total, the ISQ was completed 14,298 times: 5,900 times on the first measurement occasion, 4,649 times on the second measurement occasion and 3,749 times on the third measurement occasion. Student-ID numbers were missing on 1,927 ISQ forms (13.5% of all completed forms). Forms with missing student ID numbers were given a unique substitute student ID number, which resulted in a total of 9,616 unique teacher-student combinations.

A mean response rate of 90.2% was observed in 76 randomly selected lectures. The mean class size, in terms of ISQ forms completed, was 80.8 students on measurement occasion one ($SD = 63.3$, $min = 13$, $max = 356$), 62.8 students on measurement occasion two ($SD = 48.6$, $min = 13$, $max = 215$), and 52.8 students on measurement occasion three ($SD = 42.3$, $min = 8$, $max = 201$).

Measures

The Instructional Skills Questionnaire (ISQ) comprises seven dimensions of lecturing skills (*Structure, Explication, Stimulation, Validation, Instruction, Comprehension* and *Activation*) measured by 28 items (listed in Appendix I). Each dimension is measured by four items, two indicative items and two contra-indicative items, on a 7-point likert scale (response

options ranging from strongly disagree to strongly agree). The contra-indicative items were recoded prior to analyses. The items were adapted from Vorst and Van Engelenburg's (1992) Uvalon items. From the pool of Uvalon items, 28 items concerning lecturing behavior were selected. Prior to this study, the ISQ items were tested at the department of Psychology of the University of Amsterdam. The questionnaire was administered at the end of twenty-five lectures of twenty-five different professors. In total, 609 forms were completed. This dataset was too small to perform multilevel confirmatory factor analyses. Cronbach's α of the seven dimensions on professor level were .66 on *Structure*, .76 on *Explication*, .93 on *Stimulation*, .84 on *Validation*, .72 on *Instruction*, .88 on *Comprehension* and .95 on *Activation*. In this first version, a 5-point likert scale was used. With SET ratings, the variation is often small. To increase resolution, we used a 7-point likert scale in the final version of the instrument. A few items were adapted to improve the reliability of the subscales and items.

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

Missing item responses (3.7%) were imputed with the student's mean of the other three items of that specific dimension. Out of 14,596 forms, 298 forms were excluded; 218 forms remained incomplete after imputation, and 80 forms were marked as extreme outliers. Extreme outliers were detected with the Inter Quartile Range (IQR; distance between the first and the third quartile). For each professor on each measurement occasion separately, the IQR was calculated on the total mean score. A form was considered an extreme outlier if the rating was at least two times the IQR lower than the first quartile, or two times the IQR higher than the third quartile. This equals a deviance of 3.6 times the standard deviation from the mean.

The final dataset contained 14298 forms with 527, 237 and 255 remaining missing ratings on the student level variable *Cognition*, *Affection* and *Regulation* respectively.

Procedure

Prior to the start of their courses, all professors received procedural instructions by email. Professors informed their students by a standardized email that they (i.e., the professors) would be participating in a research project on the quality of the lectures at the university. Students were invited to participate by evaluating three lectures during the course. A standard lecture at this university takes 90 minutes with a 15-minute break after 45 minutes. In the

final fifteen minutes of the lecture, professors reserved five minutes for an evaluation break. Research assistants distributed the questionnaires and collected them during this break. Students were instructed to focus on the current lecture, while completing the ISQ. They were asked to provide their student ID number for research purposes, and were assured of anonymity by means of an extra statement on the ISQ form.

We note that in this project professors were randomly assigned to one of three conditions; a feedback-only condition, in which professors received the student feedback each time shortly after the rated lecture so they could improve their upcoming lectures (N = 24); a feedback-plus-consultation condition, in which professors received student feedback and collaborative consultation with a consultant after each rated lecture to improve the subsequent lecture (N = 26); and a control condition, in which professors received the student feedback at the end of the course (N = 25). The interventions took place after the first measurement occasion.

Statistical analyses

Internal structure and reliability of the ISQ

First, we performed a multilevel confirmatory factor analysis (MCFA) on the data of T_1 with the expected seven-factor model on the professor level (level 2), and an unconstrained model on the student level (level 1). By an unconstrained model we mean that the model contained the exact covariance structure of the data on the student level to fully represent the existing variance on the student level (no constrains). With this model we verified the psychometric quality of the ISQ, in terms of its ability to differentiate between professors on seven specific dimensions.

According to Schermelleh-Engel, Moosbrugger, and Muller (2003) a good fit is represented by the following fit indices: RMSEA < .05, SRMR < .05, CFI > .97, TLI > .97. Acceptable fit indices are: RMSEA < .08, SRMR < .10, CFI > .95, TLI > .95. In addition, item loadings should be significant ($p < .01$). We note that the measures CFI and TLI were not developed for item level analyses. As incremental fit measures they are based on the comparison between a model in which the variables are uncorrelated (a baseline) and the model as specified (Kenny, 2012). As the correlations among items cannot be expected in general to correlate as highly as the correlation among subtests (often comprising several items), the standard criteria are too stringent. We therefore adopt the criteria of CFI > .90 and TLI > .90 for the student level. On the professor level the items are expected to correlate higher due to the aggregation of the data. We therefore contain the CFI and TLI criteria as formulated by Schermelleh-Engel and colleagues (2003) for the professor level.

Second, to cross-validate the hypothesized seven-factor model, we fitted the same model with MCFA on the second and third measurement occasion (a more accurate term is "quasi-cross-validation", given the repeated measures). To take into account the experimental conditions (see procedure), two dummy coded variables were added to the model as teacher-level covariates; *C1* (feedback-only condition: coded 1, control and feedback-plus-consultation condition: coded 0) and *C2* (feedback-plus-consultation condition: coded 1, control and feedback-only condition: coded 0).

Third, the reliability of the subscales on each measurement occasion is given in terms of Cronbach's alpha on the professor level. To obtain a measure of clustering, the intra-class correlations were also calculated for each measurement occasion. These represent the variance of the ratings between professors (level 2) divided by the total variance of the ratings for each item, thus indicating the dependency of the variation in ratings on the professor.

Exploratory analyses on student level

To explore the student level covariance structure, we carried out an exploratory factor analysis of the student level data at T_1. Specifically, we performed seven multilevel exploratory factor analyses (MEFA) with one to seven factors on the student level, while leaving the professor level unconstrained. Based on the Kaiser criterion (drop all components with eigenvalues under 1.0), the Cattell scree test (plot eigenvalues against eigenvalue numbers, find the 'elbow' in the curve and keep the number of factors above the elbow), fit indices, and interpretability of the factors, we selected a final model for the student level.

Next, the completely specified multilevel model, containing the seven-factor model on professor level and the selected factor model on student level, was tested with multilevel confirmatory factor analysis (MCFA) to the data of T_1. We studied the modification indices and allowed additional correlations between items which were amenable to substantive interpretation. SET items pertaining to different dimensions often correlate or have minor cross-loadings on other factors. Marsh (2007b) warned that CFAs on student evaluation instruments therefore tend to result in a poor fit, while EFAs clearly reveal a replicable structure in the data due to the fact that EFA allows each item to cross load on other factors as well. Although we allowed several interpretable correlations between items, we did not allow items to cross-load on other factors, to preserve and test the selected hypothesized models.

To quasi-cross-validate this model, we fitted the combined model with MCFA to the data of T_2 and T_3. Again, the dummy coded variables *C1* and *C2* were added to the model as professor level covariates to account for effects of the interventions.

Relationship between student and professor level factors and student learning outcomes

The student level variables *Cognition*, *Affection* and *Regulation*, pertaining to the students' perception of their learning outcomes, were added as dependent variables to the combined multilevel model (containing the seven-factor model on professor level and the selected model on student level). In this expanded model we regressed each dependent variable on all the dimensions on professor level and on the dimensions on student level. We applied Bonferroni's correction for multiple testing by dividing the alpha of .05 by the number of regression analysis on that level. To quasi-cross-validate the results, the same analyses were done on the data of T_2 and T_3.

Results

Internal structure and reliability of the ISQ

Professor level Cronbach's alphas of each dimension at each measurement occasion are listed in Table 2.2. The reliability of the subscales on the professor level is high. Cronbach's alphas range from .88 to .98, with a mean of .93 on the first measurement occasion, from .92 to .98, with a mean of .94 on the second measurement occasion, and from .91 to .98, with a mean of .94 on the third measurement occasion. One reason that these values are quite high is that the teacher scores are based on the average test scores of their students. The averages are necessarily subject to less error variance than the student level data.

Table 2.2 Cronbach's alphas on professors mean scores for each dimension on each measurement occasion

Measurement occasion	T_1	T_2	T_3
N professors	*N* = 73	*N* = 74	*N* = 71
Dimension			
Structure	0.91	0.93	0.93
Explication	0.94	0.95	0.94
Stimulation	0.98	0.98	0.98
Validation	0.92	0.93	0.94
Instruction	0.88	0.93	0.91
Comprehension	0.92	0.92	0.93
Activation	0.97	0.97	0.97

The intra-class correlations of the items varied between .04 and .33, with a mean of .16, indicating that an average of 16% of the variance of the ratings is between professors. Intra-class correlations on each measurement occasion are given in Table 2.3.

Multilevel confirmatory factor analysis (MCFA) on the first measurement occasion with the expected seven-factor model on the professor level and an unconstrained model on the student level yielded a good fit, as indicated by the goodness-of-fit-indices (T_1: RMSEA = .021, SRMR between = .087, CFI = .978, TLI = .952). All the factor loadings were statistically significant on a 1% significance level, with a mean standardized loading of .92 (*min* = .57, *max* = 1.00, *median* = .96). Factor loadings and correlations on measurement occasion one are given in Table 2.3 and Table 2.4. Correlations among the seven factors varied from .22 to .93 (*M* = .52). The relationships between the factors are quite consistent with the proposed theory on teaching behavior and student learning. *Structure* correlated .84 with *Explication*. Both professor level factors were hypothesized to address the cognitive domain of the student learning process. *Validation* and *Instruction* correlated .72 and were hypothesized to address the regulative domain of the student learning process. *Comprehension* and *Activation* correlate .93 and were hypothesized to reflect the amount of interaction a professor has with his students. In addition, both *Structure* and *Explication* also correlated between .66 and .80 with *Validation* and *Instruction*. *Explication* and *Validation* correlated .74 and .67 with *Stimulation*. The parameters of the dummy coded variables C1 and C2, which were added to correct for condition, were not significant for any of the ISQ dimensions as expected.

The good fit of this model was confirmed on the second measurement occasion (T_2: RMSEA = .022, SRMR between = .077, CFI = .978, TLI = .953) and the third measurement occasion (T_3: RMSEA = .025, SRMR between = .068, CFI = .975, TLI = .946). Again, all the factor loadings were statistically significant, with a mean loading of .95 on both the second and third measurement occasion (T_2: *min* = .78, *max* = 1.00, *median* = .97, T_3: *min* = .79, *max* = 1.00, *median* = .97). Correlations among the seven factors varied from .36 to .94 (*M* = .66) on the second measurement occasion and form .34 to .95 (*M* = .64) on the third measurement occasion and showed a similar pattern. In each model three residual variances were fixed to zero, as they assumed small negative values. Factor loadings and correlations on each measurement occasion are given in Table 2.3 and Table 2.4.

Exploratory analyses on student level

To explore the student level covariance structure, we carried out multilevel exploratory factor analyses (MEFA) with one to seven factors on the student level, while leaving the professor

level unconstrained. According to the Kaiser criterion, the final model should contain six factors or less. The Cattell scree test indicated a four-factor model or a six-factor model for the student level. Fit indices showed an acceptable fit with four factors or five factors and a good fit with six factors or more.

To arrive at the definite factor model, we evaluated the 4 and the 6 factor models in terms of ease of interpretation. As expected the inter-item correlations are appreciably smaller at the student level (compared to the professor levels). Consequently, the factor loadings are lower. To a large extent the patterns of factor loadings within the four-factor model turned out to correspond to the theoretical model on student learning. Items pertaining to the dimension *Structure* and *Explication* loaded on factor one. Items pertaining to the dimension *Stimulation* loaded on factor two. Items pertaining to the dimension *Validation* and *Instruction* loaded on factor three. Items pertaining to the dimension *Comprehension* and *Activation* loaded on factor four. Factors one, two and three, also contained some high loading items pertaining to other dimensions. Based on its content and theory on student learning, the factors were labelled as the extent to which students perceive the subject matter to be clear (Factor 1: *Clarity*, cognitive processing learning function 1), interesting (Factor 2: *Interest*, affective learning function 2), and important to them (Factor 3: *Importance*, regulative learning function 3), and as the extent to which interaction takes place between the student and the professor (Factor 4: *Interaction*). The six-factor model was decidedly more difficult to interpret. This model included the four factors mentioned above, and two additional factors, which defied clear interpretation. In this model, the four labeled factors did not contain high loading items pertaining to other dimensions like in the four-factor model. We therefore accepted the 4 common factor model with Factor 1 (*Clarity*), comprising the *Structure* items and *Explication* items, Factor 2 (*Interest*) comprising the *Stimulation* items, Factor 3 (*Importance*) comprising the *Validation* and *Instruction* items, and Factor 4 (*Interaction*) comprising the *Comprehension* and *Activation* items.

We tested this model using a confirmatory factor analysis again at measurement occasion one. The tested model contained the four-factor model on the student level (level 1) and the seven-factor model on the professor level (level 2), with professor level dummy variables C1 and C2 to take into account differences between the conditions)[1]. Four residual variances were set to zero on the professor level, as they turned negative close to zero. Fit indices indicated a good fit according to the RMSEA (.045) SRMR within (.050) and SRMR

1 On measurement occasion one, the dummy coded covariate Condition was not expected to influence the results, since the interventions took place after the first measurement occasion. Without this covariate, the analyses resulted in similar fit indices. The correction was still made to be able to cross-validate this model with consistent models on measurement occasion two and three.

Table 2.3 ISQ item standardized factor loadings and intra class correlations (ICC) on teacher level on measurement occasion one, two and three (student level unrestricted)

	T_1			T_2			T_3		
	Estimate	(*SE*)	ICC	Estimate	(*SE*)	ICC	Estimate	(*SE*)	ICC
Structure									
STR1	0.959	(0.018)	0.12	0.944	(0.025)	0.13	0.970	(0.013)	0.13
STR2	0.739	(0.072)	0.08	0.784	(0.061)	0.09	0.789	(0.062)	0.09
STR3	1.000	(0.000)	0.11	1.000	(0.000)	0.12	1.000	(0.000)	0.12
STR4	0.871	(0.055)	0.13	0.959	(0.016)	0.14	0.985	(0.010)	0.15
Explication									
EXPL1	1.000	(0.000)	0.15	0.985	(0.015)	0.16	0.994	(0.012)	0.13
EXPL2	0.969	(0.015)	0.12	0.962	(0.018)	0.12	0.939	(0.027)	0.13
EXPL3	0.975	(0.011)	0.19	0.971	(0.015)	0.18	0.989	(0.010)	0.15
EXPL4	0.747	(0.056)	0.15	0.883	(0.039)	0.17	0.860	(0.048)	0.15
Stimulation									
STIM1	1.000	(0.000)	0.24	1.000	(0.000)	0.26	0.989	(0.007)	0.26
STIM2	0.934	(0.019)	0.29	0.967	(0.011)	0.33	0.950	(0.016)	0.30
STIM3	0.977	(0.007)	0.24	0.980	(0.007)	0.26	0.974	(0.012)	0.23
STIM4	0.988	(0.005)	0.24	0.992	(0.004)	0.28	0.999	(0.011)	0.28
Validation									
VAL1	0.805	(0.063)	0.08	0.816	(0.051)	0.09	0.806	(0.061)	0.06
VAL2	0.907	(0.064)	0.08	0.934	(0.024)	0.10	0.956	(0.016)	0.09
VAL3	0.997	(0.044)	0.10	0.980	(0.021)	0.12	0.978	(0.015)	0.11
VAL4	0.901	(0.054)	0.11	0.967	(0.020)	0.11	1.000	(0.000)	0.10

Table 2.3 *Continued*

	T_1			T_2			T_3		
	Estimate	(*SE*)	ICC	Estimate	(*SE*)	ICC	Estimate	(*SE*)	ICC
Instruction									
INSTR1	0.972	(0.052)	0.05	1.000	(0.000)	0.05	0.883	(0.073)	0.04
INSTR2	0.993	(0.053)	0.05	0.952	(0.024)	0.06	0.996	(0.021)	0.05
INSTR3	0.566	(0.121)	0.06	0.931	(0.033)	0.08	0.936	(0.040)	0.06
INSTR4	0.824	(0.067)	0.09	0.968	(0.022)	0.13	0.932	(0.029)	0.10
Comprehension									
COMP1	0.940	(0.034)	0.13	0.954	(0.024)	0.10	0.934	(0.029)	0.09
COMP2	0.945	(0.026)	0.26	0.983	(0.015)	0.26	0.966	(0.016)	0.26
COMP3	0.806	(0.067)	0.17	0.920	(0.027)	0.17	0.911	(0.029)	0.18
COMP4	0.947	(0.032)	0.16	0.863	(0.046)	0.14	0.953	(0.023)	0.14
Activation									
ACT1	0.922	(0.029)	0.28	0.957	(0.037)	0.26	0.951	(0.016)	0.25
ACT2	1.000	(0.007)	0.19	0.990	(0.027)	0.17	1.000	(0.000)	0.16
ACT3	0.967	(0.010)	0.21	0.968	(0.024)	0.18	0.974	(0.009)	0.20
ACT4	0.958	(0.020)	0.31	0.972	(0.031)	0.29	0.975	(0.011)	0.26

Note: on each measurement occasion three residual variances were fixed to zero, as they assumed small negative values, resulting in loadings of 1 (T_1: str3, expl1 and stim1, T_2: str3, stim1 and instr1, T_3: str3, val4, and act2). In addition, in some cases the factor loadings were so high that they rounded to 1.

Table 2.4 Factor correlations on teacher level on measurement occasion one, two and three (student level unrestricted)

Measurement occasion	Structure	Explication	Stimulation	Validation	Instruction	Comprehension
T_1						
Explication	0.84					
Stimulation	0.48	0.74				
Validation	0.66	0.70	0.67			
Instruction	0.80	0.70	0.45	0.72		
Comprehension	0.31	0.36	0.41	0.28	0.33	
Activation	0.22	0.32	0.48	0.34	0.27	0.93
T_2						
Structure						
Explication	0.92					
Stimulation	0.76	0.89				
Validation	0.83	0.84	0.84			
Instruction	0.77	0.70	0.68	0.79		
Comprehension	0.37	0.36	0.47	0.45	0.43	
Activation	0.47	0.51	0.65	0.59	0.51	0.94
T_3						
Structure						
Explication	0.88					
Stimulation	0.68	0.88				
Validation	0.81	0.84	0.86			
Instruction	0.82	0.72	0.67	0.83		
Comprehension	0.34	0.35	0.48	0.45	0.38	
Activation	0.39	0.47	0.60	0.58	0.47	0.95

between (.088). Fit indices CFI (.813) and TLI (.788) were too low. Marsh (2007b) warned that SETs items tend to cross correlate between factors, resulting in an inadequate fit with confirmatory factor analysis. Therefore, the modification indices were investigated and on the student level nine inter-item correlations across dimensions were allowed. We also noted that items with negative worded items tended to correlate within each dimension. Therefore, a fifth method factor, was added to the model containing all negative worded items (for other examples of adding a method factor, see e.g., Marsh, 1996). The correlations between the method factor and the other four student level factors were set to zero, since we don't hold any theory on why these factors should correlate with a methodological factor. As in the previous model, four residual variances were set to zero, as they turned negative close to zero. This resulted in a good fit (RMSEA = .032, CFI = .991, TLI = .896, SRMR within = .038 and SRMR between = .087).

To quasi-cross-validate this final model, we fitted the model to the data of measurement occasions two and three. Fit indices indicated a good fit on both measurement occasions. Fit indices on all three measurement occasions are given in Table 2.5.

Relationship between student and professor level factors and student learning outcomes

We added the student level variables *Cognition*, *Affection* and *Regulation* as dependent variables to the combined multilevel model (containing the seven-factor model on professor level and the five-factor model on student level with additional cross correlations and Condition as a professor level dummy coded covariate). Table 2.6 shows the results of the regression analyses on measurement occasion one for each dependent variable on all the

Table 2.5 Fit indices of the final model fitted on measurement occasion one, two and three

Measurement occasion	Within level	Between level	Degrees of Freedom	RMSEA	CFI	TLI	SRMR within	SRMR between
T_1	5 factors	7 factors	696	0.032	0.991	0.896	0.038	0.087
T_2	5 factors	7 factors	695	0.033	0.911	0.896	0.040	0.077
T_3	5 factors	7 factors	695	0.036	0.906	0.890	0.039	0.068

Note: On each measurement occasion, the model contained the following five factors on student level: Cognition, Affection, Regulation, Interaction and Negative Worded Items, and the following seven factors on teacher level: Structure, Explication, Stimulation, Validation, Comprehension, Activation. Correction was made for Condition. On T_1, four residual variances were set to zero. On T_2 and T_3, three residual variances were set to zero. Nine inter-item correlations were allowed.

Table 2.6 Regression coefficients for student level and teacher level factors on students' perceptions of their learning outcomes on measurement occasion one

Dependent variable	Independent variable	Estimate	(*SE*)	Est./*SE*	*p* value
	Student level factors				
Cognitive Learning Outcome: "I learned a lot from this lecture"	Clarity	0.168	(0.042)	4.032	0.000
	Interest	0.381	(0.027)	13.884	0.000
	Importance	0.107	(0.030)	3.520	0.000
	Interaction	0.001	(0.019)	0.060	0.952
Affective Learning Outcome: "Because of this lecture, I want to learn more about the subject matter"	Clarity	-0.067	(0.034)	-1.955	0.051
	Interest	0.615	(0.025)	24.758	0.000
	Importance	0.091	(0.033)	2.749	0.006
	Interaction	-0.028	(0.017)	-1.593	0.111
Regulative Learning Outcome: "Because of this lecture, I now know what I have yet to study"	Clarity	0.128	(0.031)	4.114	0.000
	Interest	-0.033	(0.030)	-1.081	0.280
	Importance	0.412	(0.035)	11.884	0.000
	Interaction	0.080	(0.026)	3.044	0.002
	Teacher level factors				
Cognitive Learning Outcome: "I learned a lot from this lecture"	Structure	0.741	(0.166)	4.463	0.000
	Explication	-0.252	(0.183)	-1.382	0.167
	Stimulation	0.797	(0.140)	5.708	0.000
	Validation	-0.216	(0.153)	-1.414	0.157
	Instruction	-0.019	(0.145)	-0.130	0.896
	Comprehension	-0.482	(0.292)	-1.653	0.098
	Activation	0.392	(0.319)	1.228	0.219
Affective Learning Outcome: "Because of this lecture, I want to learn more about the subject matter"	Structure	0.279	(0.167)	1.674	0.094
	Explication	-0.384	(0.177)	-2.167	0.030
	Stimulation	1.148	(0.100)	11.474	0.000
	Validation	0.050	(0.128)	0.387	0.699
	Instruction	-0.187	(0.133)	-1.407	0.159
	Comprehension	0.009	(0.312)	0.029	0.977
	Activation	-0.095	(0.321)	-0.297	0.766
Regulative Learning Outcome: "Because of this lecture, I now know what I have yet to study"	Structure	0.554	(0.308)	1.798	0.072
	Explication	-0.623	(0.312)	-2.000	0.046
	Stimulation	0.342	(0.222)	1.538	0.124
	Validation	-0.488	(0.220)	-2.213	0.027
	Instruction	0.969	(0.180)	5.383	0.000
	Comprehension	-0.587	(0.536)	-1.096	0.273
	Activation	0.385	(0.552)	0.697	0.486

dimensions on professor level and on the four dimensions related to the student learning process on student level. We applied Bonferroni's correction for multiple testing. Thus, on the student level regression analyses, we applied an alpha of .004 (alpha of .05 divided by twelve tests) and on the professor level regression analysis, we applied an alpha of .002 (alpha of .05 divided by twenty-eight tests) when interpreting the results. Results show a significant effect of the student level factors *Clarity*, *Interest* and *Importance* on the cognitive learning outcome variable, of the factor *Interest* on the affective learning outcome variable, and of the factors *Clarity*, *Importance* and *Interaction* on the regulative learning outcome variable.

On the professor level we found significant effects of the factors *Structure* and *Stimulation* on the cognitive learning outcome variable, of the factor *Stimulation* on the affective learning outcome variable, and of the factor Instruction on the regulative learning outcome variable. A full representation of the final student and professor level model and the relationship with students' perceptions of their learning outcomes are given in Figure 2.2 and 2.3.

Again these results were replicated on measurement occasions two and three. All relationships found on measurement occasion one remained significant on both other measurement occasions, except for the effect of Interaction on the regulative learning outcome variable (on the student level) and the effect of *Structure* on the cognitive learning outcome variable (on the professor level). No additional significant relationships were found on measurement occasions two and three.

Discussion

Course evaluation instruments often do not serve well as a source of formative feedback for university professors on their teaching behavior. The aim of the present study was to investigate the psychometric qualities of a new theory-based student evaluation of teaching instrument, the Instructional Skills Questionnaire (ISQ). This instrument is suitable to assess detailed behavior, following each lecture, and therefore can be used to provide professors with immediate and specific feedback concerning their teaching behavior. Our conceptualization of teaching in terms of the seven ISQ dimensions 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 teaching behavior (2007).

We investigated the internal structure and reliability of the seven ISQ dimensions of teaching behavior at three measurement occasions. The mean intra-class correlation of .16 indicated that variation in ratings on ISQ items do depend on the professors. The professor

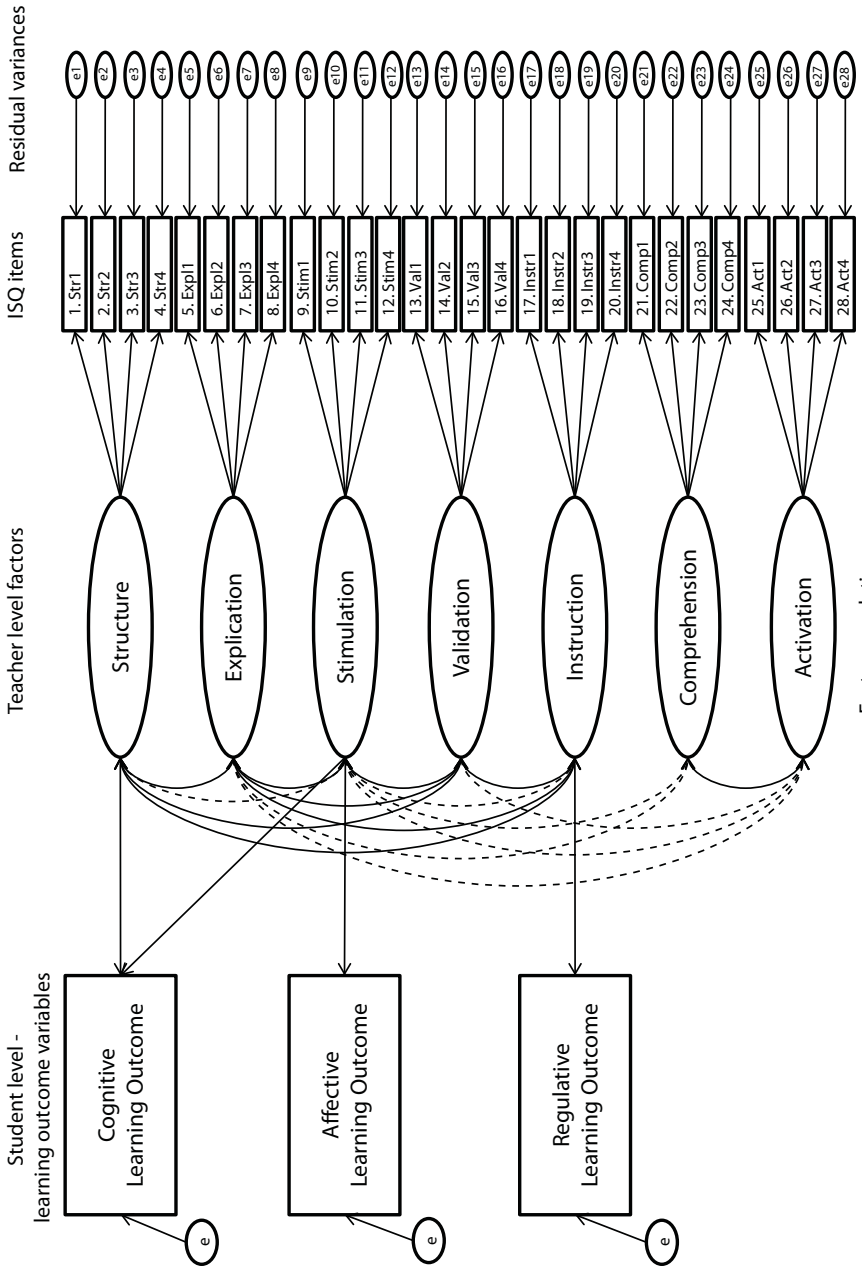

Figure 2.2 Professor level factor structure and the relationship with students' perception of their learning outcomes.

level reliabilities of the seven dimensions were found to be good. The fit-indices of the two-level confirmatory factor models indicated that the professor level seven-factor model fitted the data well on all three measurement occasions. Thus we conclude that the ISQ adequately measures seven dimensions of the professors' lecturing behavior. The instrument provided reliable and internally valid ratings on professors from a wide variety of departments at a Dutch university, on multiple occasions.

We note that at the professor level several ISQ factors correlated highly (the factor correlations ranged from .22 to .93 with a mean of .52). In almost all SET instruments certain factors tend to correlate highly, e.g., CFA on the SEEQ instrument resulted in factor correlations ranging from .02 to .87 with a median of .72 (Marsh et al., 2009). Whereas high correlations may be source of concern in other fields, we do not see the present high correlations as a problem. Highly correlating factors varied in the learning outcome variables they predicted, thus indicating the discriminant validity of the dimensions vis-à-vis outcome measures. For example, the professor level mean ratings on *Structure* significantly predicted the cognitive learning outcome variable ("I learned a lot from this lecture"), while *Explication* did not. Structure correlates highly with *Explication*, but we conclude that they do not reflect the same behavior. The correlation merely means that professors who tend to provide clear explanations also tent to structure the lecture more sufficiently. The same reasoning applies to the correlated factors *Instruction* (providing instructions on what is important to study) and *Validation* (indicating why the subject matter is important), which are both hypothesized to help regulate the students learning process. While these factors are correlated, only Instruction predicted the regulative learning outcome variable ("Because of this lecture, I now know what I have yet to study") significantly. Finally, *Comprehension* and *Activation* correlated highly, but showed different correlations with other factors. Another reason not to collapse factors is that they address different teaching goals (to check whether students understand the subject matter and provide room for questions versus involving students in the lecture). Retaining these as distinct factors also guarantees the specificity of the feedback. The provision of specific feedback is the ultimate purpose of the instrument.

Overall, the findings on the relationship between ISQ dimensions and the student level dependent variables *Cognition*, *Affection* and *Regulation* (representing the students' perceptions on their learning outcomes) provide support for the proposed theoretical framework on the relationship between teaching behavior and the student learning process. Aside from the direct relationship found between *Structure* and the cognitive learning outcome variable, and between Instruction and the regulative learning outcome variable, results showed that *Stimulation* significantly predicted the affective learning outcome variable

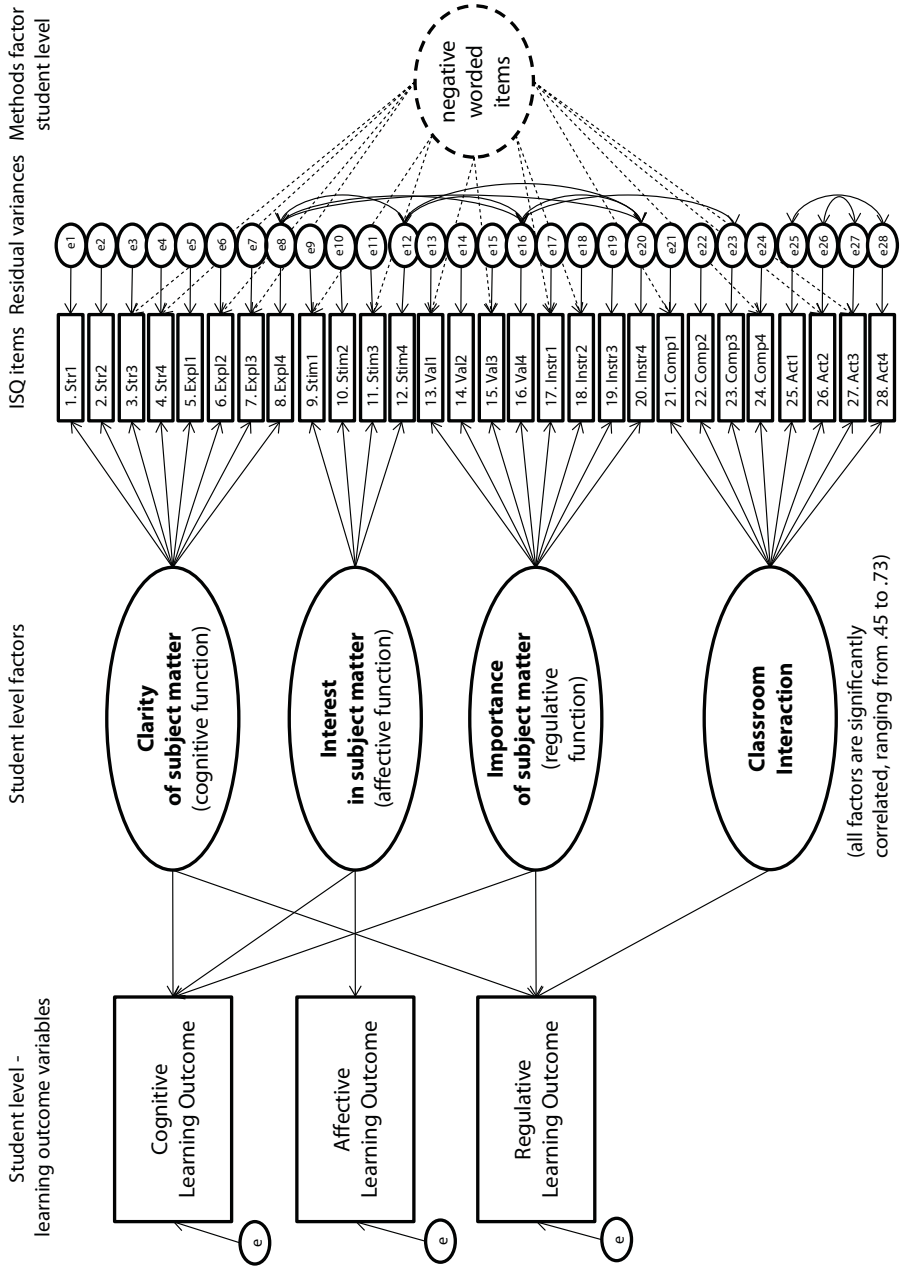

Figure 2.3 Student level factor structure and the relationship with student level learning outcome variables.

("Because of this lecture, I want to learn more about the subject matter"), as hypothesized. We note that the relationship between *Structure* and the cognitive learning outcome variable was only found on measurement occasion one. The other relationships were also found on measurement occasion two and three. In addition, *Stimulation* turned out to predict the cognitive learning outcome variable ("I learned a lot from this lecture") as well on all three measurement occasions. Although, we had not anticipated the direct effect of *Stimulation* on the cognitive learning outcome variable, it is consistent with Feldman (2007). Specifically, Feldman (2007) related teaching dimensions to domains of student achievement and overall evaluations. He found the dimensions most highly related to both to be teacher comprehensibility and clarity (identical to the dimensions *Structure* and *Explication*) and teacher stimulation of interest in the subject matter (identical to the dimension *Stimulation*). The direct relationships between ISQ teaching dimensions and student learning outcome variables support the construct validity of the instrument.

Next, the application of two level factor analysis allowed us to explore the structure of individual differences at the student level. On the basis of the student level exploratory factor analysis we identified four factors, which denoted *Clarity* (the extent to which the student perceives the subject matter to be clear), *Interest* (the extent to which the student perceive the subject matter to be interesting), *Importance* (the extent to which the student perceives the subject matter to be important) and *Interaction* (the extent to which the student interacts in classroom). We introduced an additional method factor, which served to accommodate the excess variance shared by negatively worded items. As such, this factor is of little substantive interest.

The structure, as established at the student level is again consistent with the proposed theoretical framework. The student level *Clarity* factor included mainly items from the professor level dimension *Structure* and *Explication*. Similarly, the *Interest* factor included the *Stimulation* items, *Importance* included the *Validation* and *Instruction* items, and, finally, the Interaction factor included the *Comprehension* and *Activation* items. Students differ on these factors, which relate to the cognitive domain (*Clarity* of the subject matter), affective domain (*Interest* in the subject matter), and regulative domain (*Importance* of the subject matter) of the student learning process. These results support the theoretical framework of the ISQ.

We studied the relationship between the student level factors and the students' perception of their learning outcomes. We found the factors *Clarity*, *Interest* and *Importance* significantly predicted the cognitive learning outcome variable ("I learnt a lot from this lecture"), the factor *Interest* significantly predicted the affective learning outcome variable ("Because of this lecture, I want to learn more about the subject matter"), and the factors

Clarity, *Importance* and *Interaction* significantly predicted the regulative learning outcome variable ("Because of this lecture, I now know what I have yet to study"). These exploratory and confirmatory results on the student level provide new insights on individual differences in how students perceive the same lecture. In addition, these results provide additional support of the proposed theoretical framework, and insight into the direct effects of student perceptions of the lecture on the perceived learning outcomes. Therefore, the ISQ serves both as an instrument to study differences between professors, and provides information on differences between students within the class.

In sum, this study offers a reliable and valid instrument to evaluate single lectures. The content validity, internal structure, construct validity and reliability of the ISQ teaching dimensions have been confirmed. With this instrument, professors can be provided with immediate, specific and reliable feedback on their teaching behavior and on differences between students during a course. In addition, it enables researchers to measure differences between professors in teaching behavior and differences between students in how they perceive a lecture. Finally, this study provided new insights into the classroom dynamics that characterize university lectures. Even though the lecture is not the most popular teaching format, to say the least, it is still used at universities worldwide on a daily basis. This study shows that professors have a direct influence on how useful a lecture actually is in terms of the student learning process.

References

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

De Neve, H.M.F., & Janssen, P.J. (1982). Validity of student evaluation of instruction. *Higher Education, 11,* 543-552.

Feldman, K.A. (1976b). The superior university teacher from the student's view. *Research in Higher Education, 5,* 243-288.

Feldman, K.A. (1983). The seniority and instructional experience of university teachers as related to the evaluations they receive from their students. *Research in Higher Education, 18,* 3-124.

Feldman, K.A. (1984). Class size and students' evaluations of university teacher and courses: A closer look. *Research in Higher Education, 21,* 45-116.

Feldman, K.A. (1989a). Instructional effectiveness of university teachers as judged by teachers themselves, current and former students, colleagues, administrators, and external neutral) observers. *Research in Higher Education, 30,* 137-194.

Feldman, K.A. (1989b). Association between student ratings of specific instructional dimensions and student achievement: Refining and extending the synthesis of data from multisection validity studies. *Research in Higher Education, 30,* 583-645.

Feldman, K.A. (2007). Identifying exemplary teachers and teaching: Evidence from student ratings. In R.P. Perry & J.C. Smart (Eds.), *The Scholarship of teaching and learning in higher education: An evidence-based perspective,* 93-129. Dordrecht: Springer.

Gilmore, G.M., Kane, M.T., & Naccarato, R.W. (1978). The generalizability of student ratings of instruction: Estimates of teacher and course components. *Journal of Educational Measurement, 15,* 1-13.

Howard, G.S., & Maxwell, S.E. (1980). The correlation between student satisfaction and grades: A case of mistaken causation? *Journal of Educational Psychology, 72,* 810-820.

Kember, D., Leung, D.Y.P., & Kwan, K.P. (2002). Does the use of student feedback questionnaires improve the overall quality of teaching? *Assessment and Evaluation in Higher Education, 27,* 411-425.

Kenny, D.A. (2012). *Measuring Model Fit.* http://davidakenny.net/cm/fit.htm. Last revision on July 5th 2012.

Marsh, H.W. (1984). Students evaluations of university teaching - dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology, 76,* 707-754.

Marsh, H.W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research, 11,* 253-388.

Marsh, H.W. (1996). Positive and negative global self-esteem: A substantively meaningful distinction or artifactors? *Journal of Personality and Social Psychology, 70,* 810-819.

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

Marsh. H.W. (2007b). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. In R.P. Perry & J.C. Smart (Eds.), *The Scholarship of teaching and learning in higher education: An evidence-based perspective,* 319-383. Dordrecht: Springer.

Marsh, H.W., & Dunkin, M. (1992). Students' evaluations of university teaching: A multidimensional perspective. In J.C. Smart (Ed.), *Higher Education: Handbook on Theory and Research,* 143-234. New York: Agathon.

Marsh, H.W., & Dunkin, M.J. (1997). Students' evaluations of university teaching: A multidimensional perspective. In R.P. Perry & J.C. Smart (Eds.), *Effective Teaching in Higher Education: Research and Practice,* 241-320. New York: Agathon.

Marsh, H.W., & Hocevar, D. (1991b). Students' evaluations of teaching effectiveness: The stability of mean ratings of the same teachers over a 13-year period. *Teaching and Teacher Education, 7,* 303-314.

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

Marsh, H.W., Muthén, B.O., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A.J.S., & Trautwein, U. (2009). Exploratory structural equation modeling, integrating CFA and EFA: Application to students' evaluations of university teaching. *Structural Equation Modeling, 16,* 439-476.

McLaughlin, M.W., & Pfeifer, R.S. (1988). *Teacher evaluation: Improvement, accountability, and effective learning.* New York: Teacher University Press.

Richardson, T.T.E. (2005). Instruments for obtaining student feedback: a review of the literature. *Assessment and Evaluation in Higher Education, 30,* 378-415.

Schermelleh-Engel, K., Moosbrugger, H., & Muller, H. (2003). Evaluating the fit of structural equation models: Tests of Significance and Descriptive Goodness-of-Fit measures. *Methods of Psychological Research online, 8,* 23-74.

Verbeek, F., De Jong, U., & Vermeulen, A. (2002). *Rapportage Uvalon.* Amsterdam, The Netherlands: SCO Kohnstamn Institute, University of Amsterdam.

Verbeek, F., De Jong, U., & Vermeulen, A. (2005). *Jaarverslag Uvalon 2003 en 2004.* Amsterdam, The Netherlands: SCO Kohnstamn Institute, University of Amsterdam.

Toland, M.D., & De Ayala, R.J. (2005). A multilevel factor analysis of students' evaluations of teaching. *Educational and Psychological Measurement, 65,* 272-296.

Wright, S.L., & Jenkins-Guarnieri, M.A. (2012). Student evaluations of teaching: combining the meta-analyses and demonstrating further evidence for effective use, *Assessment and Evaluation in Higher Education, 37,* 683-699.

Van Gelder, L. (1975). *Didactische analyse I.* Groningen: Pedagogisch Instituut.

Vermunt, J.D. (1996). Metacognitive, cognitive and affective aspects of learning styles and strategies: A phenomenographic analysis. *Higher Education, 31,* 25-50.

Vermunt, J.D., & Verschaffel, L. (2000). Process-oriented teaching. In R.J. Simons, J. van der Linden, & T. Duffy (Eds.), *New Learning,* 209-225. Dordrecht: Kluwer Academic.

Vorst, H.C.M., & Van Engelenburg, B. (1992). *UVALON UvA-pakket voor onderwijsevaluatie.* Amsterdam: Psychological Methods Department, University of Amsterdam.

Appendix I

English translation of the Dutch Instructional Skills Questionnaire

Structure:	the extent to which the subject matter is handled systematically and in an orderly way
1	The lecture has a clear structure
2	The instructor gives clear summaries
3	The subject matter is presented incoherently
4	The lecture is unorganized
Explication:	the extent to which the instructor explains the subject matter, especially the more complex topics
1	The instructor explains the subject matter clearly
2	The instructor is unclear
3	The instructor's explanations are hard to follow
4	The instructor gives clarifying examples
Stimulation:	the extent to which the instructor interests them for the subject matter
1	The lecture is boring
2	The instructor enlivens the subject matter
3	It is hard to stay focused on the lecture
4	The instructor interests you in the subject matter
Validation:	the extent to which the instructor stresses the benefits and the relevance of the subject matter for educational goals or future occupation
1	Little is said about the application of the subject matter
2	The instructor indicates the relevance of the subject matter
3	The utility of the subject matter is hardly discussed
4	The instructor shows why the subject matter is important
Instruction:	the extent to which the instructor provides instructions about how to study the subject matter
1	The instructor is unclear about which aspects of the subject matter are important
2	It is often unclear what the main and side issues are
3	It is clear what the instructor requires of me
4	The instructor indicates which parts of the subject matter are essential
Comprehension:	the extent to which the instructor creates opportunities for questions and remarks regarding the subject matter
1	The instructor provides insufficient occasion to ask questions
2	The instructor encourages students to ask questions about the subject matter
3	The instructor checks whether students understand the subject matter
4	The instructor hardly addresses the students' comments
Activation:	the extent to which the instructor encourages students to think about and work with the subject matter
1	Students are encouraged to think along during the lecture
2	The instructor provides little opportunity for discussions
3	During this lecture there is hardly any occasion to discuss the subject matter
4	The instructor involves students in the lecture

3

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

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

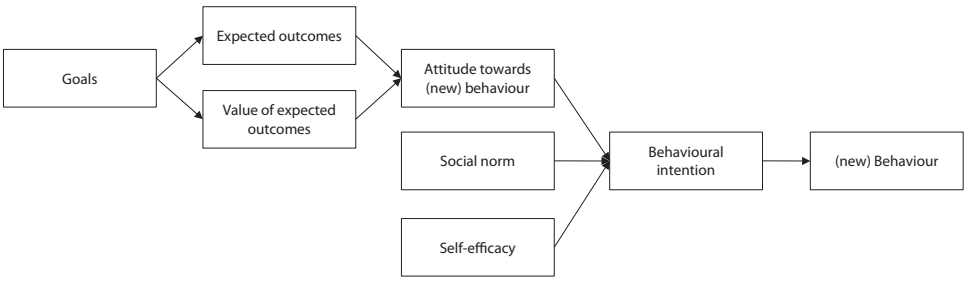

Figure 3.1 Combination of attitude-behavior models, conditions for behavioral change.

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 Enthousiasm, 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 Kohnstamn 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:

$$\text{Level 1:} \quad Y_{tij} = \beta_{0ij} + e_{tij} \tag{1.1}$$

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

$$\text{Level 3:} \quad \beta_{00j} = \gamma_{000} + v_{00j} \tag{1.3}$$

Here the student rating on dimension Y_{tij} on occasion t of student *i* in the class of professor *j* is modeled by the intercept β_{0ij} and a residual error term e_{tij}. 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 β_{0ij} is decomposed by a residual error term for students u_{0ij} (random intercept on student level), a residual error term for professors v_{00j} (random intercept on professor level) and a fixed effect parameter γ_{000} (the overall mean). The variances of the three residual error terms are denoted by

$$\text{var}\,(e_{tij}) = \sigma^2, \qquad \text{var}\,(u_{0ij}) = \tau_0^2, \qquad \text{var}\,(v_{00j}) = \varphi_0^2 \tag{1.4}$$

This model was used to calculate the intra-class correlation for the professor level by dividing the variance of the professor level (φ_0^2) by the total variance ($\sigma^2 + \tau_0^2 + \varphi_0^2$). For the proportion of the total variance that is due to differences between students we divided the variance of the student level (τ_0^2) by the total variance ($\sigma^2 + \tau_0^2 + \varphi_0^2$).

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:

$$\text{Level 1:} \quad Y_{tij} = \beta_{0ij} + \beta_1 \text{Time}_{ij} + \beta_2 \text{Condition}_j + e_{tij} \qquad (2.1)$$

$$\text{Level 2:} \quad \beta_{0ij} = \beta_{00j} + u_{0ij} \qquad (2.2)$$

$$\text{Level 3:} \quad \beta_{00j} = \gamma_{000} + v_{00j} \qquad (2.3)$$

Here β_{0ij} is the intercept, β_1 is the fixed effect parameter for *Time*, β_2 is the fixed effect parameter for *Condition* and e_{tij} is the residual error term. Again the intercept β_{0ij} is allowed to be random over professors and students by the decomposition into one fixed component (γ_{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:

$$\text{Level 1:} \quad Y_{tij} = \beta_{0ij} + \beta_{1ij} \text{Time}_{ij} + \beta_2 \text{Condition}_j + e_{tij} \qquad (3.1)$$

$$\text{Level 2:} \quad \beta_{0ij} = \beta_{00j} + u_{0ij} \qquad (3.2)$$

$$\text{Level 2:} \quad \beta_{1ij} = \beta_{10j} + u_{1ij} \qquad (3.3)$$

$$\text{Level 3:} \quad \beta_{00j} = \gamma_{000} + v_{00j} \qquad (3.4)$$

$$\text{Level 3:} \quad \beta_{10j} = \gamma_{100} + v_{10j} \qquad (3.5)$$

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 β_{0ij} is the intercept, β_1 is the random effect parameter for *Time*, β_2 is the fixed effect parameter for *Condition* and e_{tij} is the residual error term. Again the intercept β_{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 β_{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 γ_{100} represents the overall average regression coefficient for *Time* (mean slope).

The slope variances are denoted by

$$\text{var}\,(u_{1ij}) = \tau_1^{\,2}\,,\ \ \text{var}\,(v_{10j}) = \varphi_1^{\,2} \tag{3.6}$$

The intercept-slope covariances are denoted by

$$\text{cov}\,(u_{0ij}\,,\,u_{1ij}) = \tau_{01}\,,\qquad \text{cov}\,(v_{00j}\,,\,v_{10j}) = \varphi_{01} \tag{3.7}$$

Effects of SET consultation

With the fourth model we analyzed the effect of the intervention by adding the interaction effect *Time*Condition* (Model 4). Model 4 is defined through the equations:

$$\text{Level 1:}\quad Y_{tij} = \beta_{0ij} + \beta_{1ij}\,\text{Time}_{ij} + \beta_2\,\text{Condition}_j + \beta_3\,\text{Time*Condition}_{ij} + \text{e}_{tij} \tag{4.1}$$

$$\text{Level 2:}\quad \beta_{0ij} = \beta_{00j} + u_{0ij} \tag{4.2}$$

$$\text{Level 2:}\quad \beta_{1ij} = \beta_{10j} + u_{1ij} \tag{4.3}$$

$$\text{Level 3:}\quad \beta_{00j} = \gamma_{000} + v_{00j} \tag{4.4}$$

$$\text{Level 3:}\quad \beta_{10j} = \gamma_{100} + v_{10j} \tag{4.5}$$

The parameters are the same as in Model 3. The additional fixed effect parameter for *Time*Condition* (β_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 *Time*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 it's mean difference of T_4 and T_1 by its pooled standard deviation ($\sqrt{((SD\ (T_{1\_condition})^2 + SD\ (T_{4\_condition})^2) / 2}$). 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 ($\sqrt{((SD\ (T_{1\_control})^2 + SD\ (T_{4\_control})^2 + SD\ (T_{1\_exp})^2 + SD\ (T_{4\_exp})^2) / 4}$). 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 $SD\ (e_{tij})$ equals the standard deviation of Y_{tij} 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 $SD\ (e_{tij})$, 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 $SD\ (e_{tij})$. 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:

$$\begin{aligned} \text{Level 1:}\quad & Y_{tij} = \beta_{0ij} + \beta_1\ \text{Condition}_j + \beta_2\ T_1T_{2_{ij}} + \beta_3\ T_2T_{3_{ij}} + \\ & \beta_4\ T_3T_{4_{ij}} + \beta_5\ T_1T_2{}^*\text{Condition}_{ij} + \\ & \beta_6\ T_2T_3{}^*\text{Condition}_{ij} + \beta_7\ T_3T_4{}^*\text{Condition}_{ij} + e_{tij} \end{aligned} \qquad (5.1)$$

Level 2: $\beta_{0ij} = \beta_{00j} + u_{0ij}$ (5.2)

Level 3: $\beta_{00j} = \gamma_{000} + v_{00j}$ (5.3)

The parameter β_{0ij} represents the intercept. The intercept is random over professors and students. The fixed effect parameter β_1 represents the main effect of *Condition*. The fixed effect parameters β_2, β_3 and β_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*$_{ij}$ (β_5) represents the effect of the first SET consultation. The fixed interaction effect parameters for T_2T_3**Condition*$_{ij}$ (β_6) and T_3T_4**Condition*$_{ij}$ (β_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:

Level 1: $Y_{tij} = \beta_{0ij} + \beta_1 \text{Time}_{ij} + \beta_2 \text{Control-versus-NoTarget}_j$
$+ \beta_3 \text{Control-versus-Target}_j + \beta_4 \text{Time*Control-versus-NoTarget}_{ij}$
$+ \beta_5 \text{Time*Control-versus-Target}_{ij} + e_{tij}$ (5.1)

Level 2: $\beta_{0ij} = \beta_{00j} + u_{0ij}$ (5.2)

Level 3: $\beta_{00j} = \gamma_{000} + v_{00j}$ (5.3)

The parameter β_{0ij} represents the intercept. The intercept is random over professors and students. The fixed effect parameter β_1 represents the main effect of Time on the specific time interval. The fixed effect parameters β_2 and β_3 represent the contrasts *Control-versus-NoTarget* and *Control-versus-Target*. The fixed interaction effect parameter for *Time*Control-versus-NoTarget*$_{ij}$ (β_4) and *Time*Control-versus-Target*$_{ij}$ (β_5) represent the effects of SET consultation for non-targeted dimensions and targeted dimensions compared to the control condition.

Results

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 ($\chi^2[3] = 5.834$, $p = 0.12$).

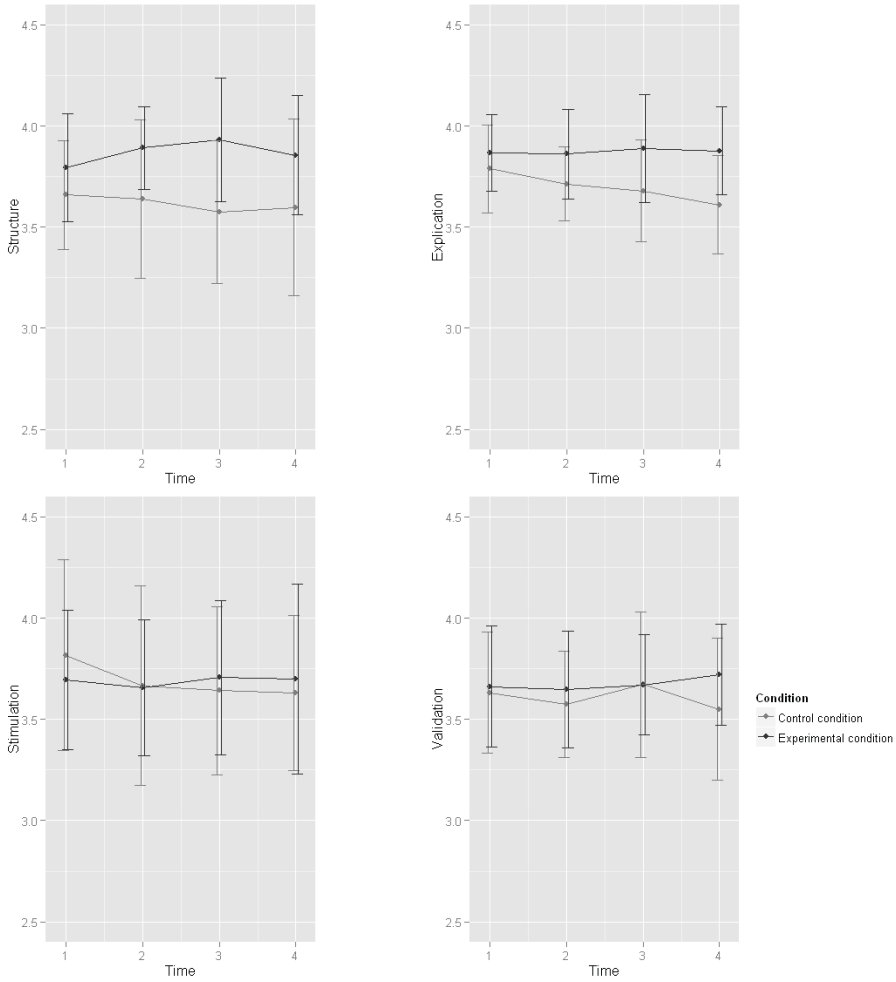

Figure 3.2a-2h Mean ratings on the eight dependent variables in the experimental condition ($N = 12$) and control condition ($N = 13$) on the four measurement occasion.

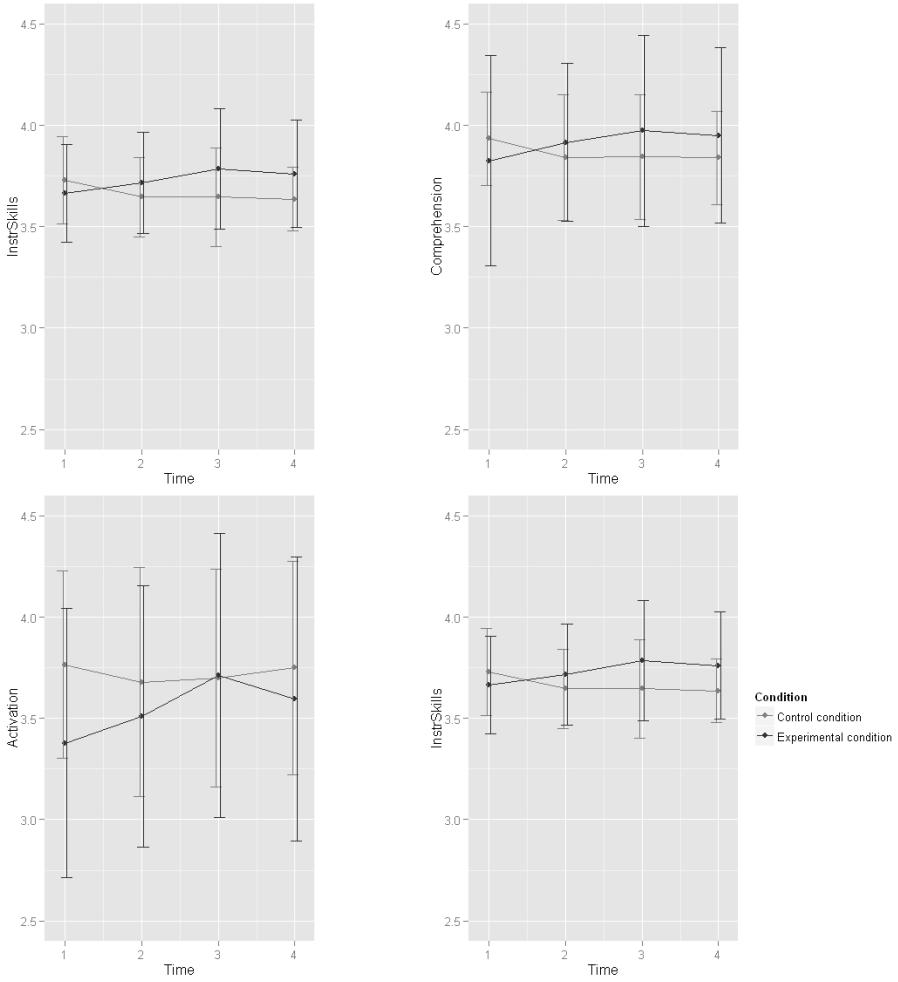

Figure 3.2a-2h *Continued.*

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*[3].

3 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

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

Table 3.1 *Continued*

	Measurement occasion 3							Measurement occasion 4						
	Professor level (between classes)			Student level (within classes)				Professor level (between classes)			Student level (within classes)			
	N	*M*	*SD*	*N*	*SD* Min	*SD* Median	*SD* Max	*N*	*M*	*SD*	*N*	*SD* Min	*SD* Median	*SD* Max
Control group	13			157				13			145			
Structure		3.56	0.27		0.22	0.58	0.69		3.58	0.41		0.25	0.57	0.74
Explication		3.65	0.20		0.28	0.56	0.97		3.58	0.16		0.26	0.56	0.95
Stimulation		3.63	0.43		0.50	0.58	1.02		3.59	0.37		0.36	0.59	1.06
Validation		3.64	0.33		0.26	0.46	0.93		3.51	0.30		0.21	0.62	0.86
Instruction		3.39	0.26		0.27	0.52	0.70		3.48	0.22		0.23	0.54	0.84
Comprehension		3.82	0.28		0.24	0.51	0.77		3.84	0.19		0.31	0.53	0.78
Activation		3.66	0.49		0.32	0.59	0.79		3.73	0.49		0.28	0.56	0.80
Total Instructional Skills		**3.61**	**0.20**		**0.20**	**0.42**	**0.59**		**3.60**	**0.14**		**0.20**	**0.38**	**0.74**
Experimental group	11			225				12			280			
Structure		3.89	0.25		0.35	0.45	0.59		3.83	0.28		0.38	0.50	0.64
Explication		3.80	0.23		0.37	0.51	0.90		3.85	0.18		0.37	0.49	0.65
Stimulation		3.59	0.34		0.34	0.52	0.79		3.64	0.43		0.39	0.59	0.74
Validation		3.62	0.16		0.41	0.51	0.83		3.69	0.20		0.38	0.50	0.62
Instruction		3.56	0.28		0.27	0.52	0.75		3.63	0.28		0.38	0.52	0.89
Comprehension		3.96	0.39		0.36	0.53	0.79		3.95	0.41		0.36	0.48	0.72
Activation		3.64	0.59		0.28	0.51	0.86		3.60	0.64		0.41	0.58	0.79
Total Instructional Skills		**3.71**	**0.22**		**0.23**	**0.36**	**0.54**		**3.74**	**0.25**		**0.25**	**0.38**	**0.47**

Note: student level *N* refers to the number of evaluation forms completed at measurement occasion one, two, three and four.

Table 3.2 Deviance tests between the four models for *Total Instructional Skills*

	-2 log likelihood	deviance	*df*	*p*-value
Model 1 (random intercept only)	1655.1			
Model 2 (random intercept + Time + Condition)	1653.4	1.7	2	0.421
Model 3 (random intercept + random slope + Time + Condition)	1599.8	53.6	4	<0.001
Model 4 (random intercept + random slope + Time + Condition +Time*Condition)	1591.7	8.1	1	0.004

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 3.3 Estimates and standard errors of the explanatory variables under Model 3 for the dependent variables *Structure*, *Stimulation*, *Validation* and *Instruction*

		Structure Model 3		Stimulation Model 3		Validation Model 3		Instruction Model 3	
		Estimate	(*SE*)	Estimate	(*SE*)	Estimate	(*SE*)	Estimate	(*SE*)
Fixed Part									
β_0	Constant	3.360	(0.069)	3.762	(0.104) *	3.588	(0.071)	3.427	(0.057)
β_1	Time	-0.001	(0.021)	-0.044	(0.017) *	-0.005	(0.016)	0.025	(0.021)
β_2	Condition	0.189	(0.096)	-0.089	(0.148)	0.066	(0.095)	0.052	(0.076)
Random Part									
Intercept variance									
φ_0^2	Level 3: Professor	0.051	(0.017) *	0.122	(0.039) *	0.060	(0.020) *	0.033	(0.012) *
τ_0^2	Level 2: Student	0.136	(0.017) *	0.263	(0.025) *	0.131	(0.017) *	0.140	(0.019) *
σ^2	Level 1: Time	0.141	(0.010) *	0.164	(0.012) *	0.138	(0.010) *	0.153	(0.011) *
Slope variance									
φ_1^2	Level 3: Professor	0.008	(0.003) *	0.003	(0.002)	0.003	(0.002)	0.008	(0.003) *
τ_1^2	Level 2: Student	0.002	(0.004)	0.011	(0.005) *	0.007	(0.004) *	0.012	(0.005) *
Intercept-slope covariance									
φ_{01}	Level 3: Professor	-0.004	(0.005)	0.002	(0.006)	-0.006	(0.005)	-0.006	(0.005)
τ_{01}	Level 2: Student	0.000	(0.007)	-0.014	(0.010)	0.000	(0.007)	-0.009	(0.008)

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*

		Instructional Skills Model 4			Explication Model 4			Comprehension Model 4			Activation Model 4		
		Estimate	(*SE*)		Estimate	(*SE*)		Estimate	(*SE*)		Estimate	(*SE*)	
Fixed Part													
β_0	Constant	3.710	(0.060)		3.763	(0.055)		3.922	(0.101)		3.748	(0.146)	
β_1	Time	-0.040	(0.015)	*	-0.057	(0.019)	*	-0.040	(0.024)		-0.026	(0.035)	
β_2	Condition	-0.045	(0.086)		0.074	(0.076)		-0.086	(0.144)		-0.340	(0.209)	
β_3	Time*Condition	0.065	(0.020)	*	0.053	(0.025)	*	0.086	(0.033)	*	0.103	(0.049)	*
Random Part													
Intercept variance													
φ_0^2	Level 3: Professor	0.041	(0.013)	*	0.025	(0.010)	*	0.118	(0.036)	*	0.258	(0.077)	*
τ_0^2	Level 2: Student	0.091	(0.008)	*	0.153	(0.018)	*	0.114	(0.011)	*	0.175	(0.022)	*
σ^2	Level 1: Time	0.047	(0.003)	*	0.141	(0.010)	*	0.158	(0.009)	*	0.175	(0.012)	*
Slope variance													
φ_1^2	Level 3: Professor	0.001	(0.001)		0.001	(0.001)		0.004	(0.002)	*	0.011	(0.004)	*
τ_1^2	Level 2: Student	0.005	(0.002)	*	0.008	(0.004)	*	0.000	(0.000)		0.005	(0.005)	
Intercept-slope covariance													
φ_{01}	Level 3: Professor	-0.002	(0.002)		-0.001	(0.002)		-0.009	(0.006)		-0.007	(0.013)	
τ_{01}	Level 2: Student	0.002	(0.003)		-0.003	(0.007)		0.000	(0.000)		-0.001	(0.009)	

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*

Condition	**Control condition**		**Experimental condition**	
Comparison	**(T_4 vs T_1)**		**(T_4 vs T_1) vs Control (T_4 vs T_1)**	
Effect size	Cohen's *d*	Multilevel model	Cohen's *d*	Multilevel model
Calculation	$M(T_{4\_C})$ - $M(T_{1\_C})$	3**Time*	($M(T_{4\_E})$ - $M(T_{1\_E})$) - ($M(T_{4\_C})$ - $M(T_{1\_C})$)	3**Time***Condition*
	pooled *SD*	*SD* (e_{tij})	pooled *SD*	*SD* (e_{tij})
Explication	-0.77	-0.46	0.86	0.42
Comprehension	-0.41	-0.31	0.59	0.65
Activation	-0.03	-0.18	0.39	0.73
Total Instructional Skills	**-0.49**	**-0.55**	**0.84**	**0.88**

Notes: the control condition is denoted C, the experimental condition is denoted E. Pooled *SD* for the comparison T_4 vs T_1 of the control condition was calculated with $\sqrt{((SD(T_{1\_C})^2 + SD(T_{4\_C})^2)/2)}$. Pooled *SD* for the comparison T_4 vs T_1 of the experimental condition versus the control condition was calculated with $\sqrt{((SD(T_{1\_C})^2 + SD(T_{4\_C})^2 + SD(T_{1\_E})^2 + SD(T_{4\_E})^2)/4)}$.

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 β_5) and no additional effects of the second and third consultation (parameters β_6 and β_7). The parameter of T_1T_2 (β_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 β_5). The

Table 3.6 Estimates and standard errors of the explanatory variables under Model 5 for the dependent variable *Total Instructional Skills*

		Instructional Skills Model 5		
		Estimate	(*SE*)	
Fixed Part				
β_0	Constant	3.726	(0.059)	
β_1	Condition	-0.080	(0.083)	
β_2	T1T2	-0.080	(0.030)	*
β_3	T2T3	-0.040	(0.032)	
β_4	T3T4	0.016	(0.035)	
β_5	T1T2*Condition	0.123	(0.037)	*
β_6	T2T3*Condition	0.063	(0.041)	
β_7	T3T4*Condition	0.019	(0.044)	
Random Part				
Intercept variance				
φ_0^2	Level 3: Professor	0.037	(0.011)	*
τ_0^2	Level 2: Student	0.102	(0.006)	*
σ^2	Level 1: Time	0.056	(0.003)	*

Notes: * $p < .05$. Units level 3: 25 professors, units level 2: 1,333 students, units level 1: 1,954 questionnaires.

ratings of the control condition decreased further between the second and third measurement occasion, but not significantly (parameter β_3), and remains stable between the third and fourth measurement occasion (parameter β_4). Compared to the control condition, ratings of the experimental condition increased after the second consultation, but not significantly (parameter β_6), and remained stable after the third consultation (parameter β_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

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 (T_1 vs. T_2) 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

Ajzen, I., & Fishbein, M. (1980). *Understanding attitudes and predicting social behavior.* Englewood Cliffs, NJ: Prentice-Hall.

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes, 50*, 179-211.

Bandura, A. (1977). *Social learning theory.* New York, NY: General Learning Press.

Bandura, A. (1997). *Self-efficacy: The exercise of control.* New York, NY: W.H. Freeman.

Benton, S.L., & Cashin, W.E. (2012). *Student ratings of teaching: A summary of research and literature (IDEA Paper no. 50).* Manhattan, KS: The IDEA Center. Http://www.theideacenter.org/sites/default/files/idea-paper_50.pdf

Brinko, K.T. (1990). Instructional consultation with feedback in higher education. *Journal of Higher Education, 61,* 65-83.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.)*. New Jersey: Lawrence Erlbaum.

Dresel, M., & Rindermann, H. (2011). Counseling university instructors based on student evaluations of their teaching effectiveness: a multilevel test of its effectiveness under consideration of bias and unfairness variables. *Research in Higher Education, 52,* 717-737.

De Neve, H.M.F., & Janssen, P.J. (1982). Validity of student evaluation of instruction. *Higher Education, 11,* 543-552.

Eagly, A.H., & Chaiken, S. (1998). Attitude structure and function. In D. T. Gilbert, S. T. Fiske & G. Lindzey (Eds), *The Handbook of Social Psychology (4th ed),* 269-322. New York, NY: McGraw-Hill.

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

Hattie, J.A.C., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*, 81-112.

Hox, J.J. (2002). *Multilevel Analysis. Techniques and Applications.* Mahwah, NJ: Lawrence Erlbaum Associates.

Kember, D., Leung, D.Y.P., & Kwan, K.P. (2002). Does the use of student feedback questionnaires improve the overall quality of teaching? *Assessment & Evaluation in Higher Education, 27,* 411-425.

Knapper, C., & Piccinin, S. (1999). Consultation about teaching: An overview. In Knapper, C. and Piccinin, S. (Eds), *Using Consultants to Improve Teaching. New Directions for Teaching and Learning, Number 79.* San Francisco, CA: Jossey-Bass.

Lang, J.W.B., & Kersting, M. (2007). Regular feedback from student ratings of instruction: Do college teachers improve their ratings in the long run? *Instructional Science, 35,* 187-205.

Lenze, L.F. (1996). Instructional development: What works? *National Education Association, Office of Higher Education Update, 2,* 1-4.

Levinson-Rose, J., & Menges, R.J. (1981). Improving college teaching: A critical review of research. *Review of Educational Research, 51,* 403-434.

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

Madden, T., Ellen, P., & Ajzen, I. (1992). A comparison of the theory of planned behavior and the theory of reasoned action. *Personality and Social Psychology Bulletin, 18,* 3-9.

Marsh, H.W. (1984). Students evaluations of university teaching - dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology, 76,* 707-754.

Marsh, H.W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research, 11,* 253-388.

Marsh, H.W., & Hocevar, D. (1991b). Students' evaluations of teaching effectiveness: The stability of mean ratings of the same teachers over a 13-year period. *Teaching and Teacher Education, 7,* 303-314.

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

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

Marsh, H.W. (2007b). Students' Evaluations of University Teaching: Dimensionality, Reliability, Validity, Potential Biases and Usefulness. In Perry, R.P. & Smart, J.C. (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective*, 319-384. New York, NY: Springer.

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

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

McLaughlin, M.W., & Pfeifer, R.S. (1988). *Teacher Evaluation: Improvement, Accountability, and Effective Learning.* New York, NY: Teachers College Press.

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

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

Prebble, T., Hargraves, H., Leach, L., Naidoo, K., Suddaby, G., & Zepke, N. (2004). *Impact of student support services and academic development programmes on student outcomes in undergraduate tertiary study: A synthesis of the research.* Report to the Ministry of Education, Massey University College of Education.

Richardson, T.T.E. (2005). Instruments for obtaining student feedback: a review of the literature. *Assessment and Evaluation in Higher Education, 30,* 378-415.

Rindermann, H., Kohler, J., & Meisenberg, G. (2007). Quality of instruction improved by evaluation and consultation of instructors. *International Journal for Academic Development, 12,* 73-85.

Sheppard, B.H., Hartwick, J., & Warshaw, P.R. (1988). The theory of reasoned action: A meta-analysis of past research with recommendations for modifications and future research. *Journal of Consumer Research, 15*, 325-343.

Snijders, T.A.B., & Bosker, R.J. (1999). *Multilevel Analysis: An Introduction to Basicand Advanced Multilevel Modeling.* Thousand Oaks, CA: Sage.

Stes, A., Min-Leliveld, M., Gijbels, D., & Van Petegem, P. (2010). The impact of instructional development in higher education: The state-of-the-art of the research. *Educational Research Review, 5*, 25-49.

Theall, M., & Franklin, J. (2001). Looking for Bias in All the Wrong Places: A Search for Truth or a Witch Hunt in Student Ratings of Instruction?. In Theall, M. P. Abrami, and Mets, L. (Eds.) *The Student Ratings Debate: Are they Valid? How Can We Best use Them? New Directions for Institutional Research, 109,* 45-56. Jossey-Bass: San Francisco.

Van den Putte, B. (1993). *On the theory of reasoned action*. Dissertation, University of Amsterdam.

SCO Kohnstamn Institute (2002). *Rapportage Uvalon.* Amsterdam: University of Amsterdam.

SCO Kohnstamn Institute (2005). *Jaarverslag Uvalon 2003 en 2004.* Amsterdam: University of Amsterdam.

Vorst, H.C.M., & Van Engelenburg, B. (1994). *UVALON, UvA-pakket voor onderwijsevaluatie.* The Netherlands: Psychological Methods Department, University of Amsterdam.

Weimer, M., & Lenze, L.F. (1997). Instructional interventions: A review of the literature on efforts to improve instruction. In K. R. Perry, & J. C. Smart (Eds.), *Effective teaching in higher education: Research and practice*, 205-240. New York, NY: Agathon Press.

4

Experiment II:

Effects of intermediate student feedback and collaborative consultation on university professors' self-assessed learning on lecturing

Abstract

The present study concerns experimental effects of intermediate students' evaluations of teaching (SETs) with or without consultation (SET consultation) on professors' self-assessed knowledge, attitudes, focus of attention and skills on lecturing. In total, 75 professors from five different university departments participated in the study. In 75 university courses, students gave specific feedback on the professors' lecturing skills at the end of three single lectures during the course. The study contains a randomized controlled design with three conditions; a control condition in which professors received the student feedback at the end of the course ($N = 25$), a feedback-only condition in which professors received the student feedback each time shortly after the rated lecture ($N = 24$), and a feedback-plus-consultation condition in which professors received student feedback and collaborative consultation with a consultant after each rated lecture ($N = 26$). The results of the post-tests, administered to professors (response $N = 70$), showed significant changes on their self-reported knowledge, attitudes, focus of attention and skills on lecturing for the feedback-plus-consultation condition compared to the control condition. Feedback-only had no significant impact. The implications of these findings for the practice of intermediate student feedback and consultation are discussed.

Introduction

A main purpose of collecting students' evaluations of teaching (SETs) at universities is to provide professors with feedback so they can improve the quality of their teaching. Although SETs are considered to be useful sources of information by most professors, SETs collected at the end of the course tend to have little effect on teaching behavior (Kember, Leung & Kwan, 2002, see also Marsh, 2007a; Marsh & Hocevar, 1991b). Providing university professors with student feedback *during* the course has proven to be more effective (Cohen, 1980; Menges & Brinko, 1986). More effective still is to augment SETs with individual consultation (see reviews by Cohen, 1980; Menges & Brinko, 1986; Penny & Coe, 2004).

What remains largely unknown at present is the impact of intermediate SETs with and without consultation on the views of the professors concerning the benefits of these interventions on the various aspects of their teaching. This study addressed the effects of these interventions on professors' self-reported learning on lecturing, using an experimental design including a control condition.

The outline of this paper is as follows. First, we outline four levels of evaluating interventions, which aim to facilitate the professional development of professors. We address the importance of experimental research on the effects at the level of the professors' perception of what they have learned. Next, we present detailed experimental findings of these interventions on this level of evaluation. We conclude this paper with a discussion.

Levels of evaluation

Kirkpatrick (1959, 1976, 1994) distinguished four levels of evaluation of training programs in business and industry: reaction, learning, behavior, and results. The reaction level of evaluation concerns participants' satisfaction with the program. The learning level concerns the knowledge, attitudes, and skills that participants acquire as a result of the program. The behavioral level of evaluation concerns participants' behavioral changes on the job, due to the program, and the result level of evaluation concerns the effects of the program on the organization. Guskey (2000) adapted Kirkpatrick's evaluation model to the educational field, specifically to evaluate the professional development of teachers. Guskey's five-level evaluation model comprises participants' reaction (level 1), participant's learning (level 2), organization support and change (level 3), participant's application of new knowledge and skills (level 4), and student learning outcomes (level 5). This model implies a hierarchic arrangement of levels, from simple to more complex, with the higher level building on the lower levels (Guskey, 2000).

In three reviews of the literature, authors have found small effects of providing professors with intermediate student feedback, and medium effects of feedback with additional individual consultation on students' evaluation of professors' teaching skills (Cohen, 1980; Menges & Brinko, 1986; Penny & Coe, 2004). In Guskey's model, effects in terms of an improvement in student ratings relate to the fourth level of effect: professors utilize the feedback, adapt their teaching behavior, and consequently obtain better student ratings on their teaching skills. The fact that small to medium effects were found is encouraging, given that it is not simple to obtain effects at this level for the following reasons. First, professors have to be willing to act immediately on the intermediate feedback and consultation. Second, professors have to interpret the ratings carefully (Theall & Franklin, 2001), 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, i.e., result in an increase in subsequent student ratings.

Beside these findings, the question remains whether there are appreciable differences in the amount and the content of the knowledge, attitudes, and skills that professors perceive to have acquired (level 2). As the process of feedback and consultation, optimization of teaching behavior, and student re-evaluation is multifaceted and complex, it is important to evaluate its impact in sufficient detail. De Neve (1991) found that the professor's interpretation of student ratings, and his or her 'thinking about lecturing' are important conditions for the effective use of students feedback. More insight in changes that do or do not occur on level 2 increases our understanding of findings on level 4. For example, it's possible that, in the light of student feedback, activating students becomes more important to certain professors and they increase their focus of attention to this domain of teaching. At the same time, they may not increase their knowledge on teaching strategies on how to activate students and/or may not be successful in improving this teaching domain. As a result, they do not feel they improve their teaching skills on this matter. In addition, professors' characteristics, such as age or teaching quality, may moderate the effects on their knowledge, focus of attention, attitudes and skills. With greater knowledge on the specific impact of intermediate feedback with and without consultation, we can optimize the professors' development as lecturers.

At universities, the effectiveness of faculty development practices is rarely investigated thoroughly. The authors of reviews on the effects of faculty development programs stressed the importance of more experimental research in this field (Levinson-Rose & Menges, 1981; Prebble et al., 2004; Steinert et al., 2006; Stes, Min-Leliveld, Gijbels, & Van Petegem, 2010; Weimer & Lenze, 1997). Many studies in this field employ weak designs; they often lack a

control condition, appropriate randomization, or are limited to small selected samples (e.g., professors who approach a professional development center to improve their teaching and are highly motivated to change). In the present study, we aim to overcome these limitations.

The present investigation

In the present study we focus on improving the quality of lectures at a Dutch University. We investigated the impact of intermediate student feedback with and without consultation in professors from a wide variety of departments and other teaching characteristics. In the Netherlands, bachelor programs generally take three years, and focus on a specific field of study from inception. At this university, courses take eight weeks (varying in workload), and students attend several courses per semester. The courses that feature in the present study included at least seven lectures (one lecture a week, final exams take place in the eight week). A standard lecture takes 90 minutes with a 15-minute break. Most courses include additional meetings during each week in small groups to discuss course assignments with a teaching assistant.

Students in this study evaluated three lectures during their course. Prior to the course, professors were randomly assigned to three conditions; a feedback-only condition, in which professors received the student feedback shortly after each rated lecture, a feedback-plus consultation condition, in which professors received the student feedback with consultation shortly after each rated lecture, and a control condition, in which professors received the student feedback at the end of the course.

With a post-test, administered to all participating professors, we investigated the professors' satisfaction with the program to which they were assigned (Guskey's level 1), and professors' self-reported learning on different dimension of lecturing skills (Guskey's level 2). In terms of teaching dimensions, we defined seven specific dimensions of lecturing skills in Chapter 2; structuring the subject matter (Structure), explaining the subject matter (Explication), interest students for the subject matter (Stimulation), stressing the relevance of the subject matter (Validation), providing instructions on how to study the subject matter (Instruction), creating opportunities for questions and remarks (Comprehension), and encouraging students to think about the subject matter (Activation). Professors' self-reported learning, in terms of changes in their knowledge, attitudes and skills, were investigated on these seven dimensions, and on designing, teaching, and evaluating their lectures in general. In addition, we investigated the increase in the professors' focus of attention to these dimensions, and to their plans for improvement. In so doing we hoped to gain more insight

on what professors reflect on, and to what extent they make use of, the student feedback. We addressed the following specific research questions:

1. How satisfied are professors with intermediate feedback with or without consultation compared to professors in a control condition? Specifically, to what extent do they find the complete program, the lecture evaluations, and the consultation to be useful to improve their teaching, and recommendable to colleagues?
2. What are the self-assessed effects of intermediate feedback with or without consultation on professors' knowledge, attitudes, focus of attention, and skills relating to designing, teaching and evaluating lectures in general, and with reference to the seven teaching dimensions, compared to the control condition?
3. Do professors in the various conditions differ in the extent to which they make plans to improve their teaching after the lecture evaluations?

In addition, we explored the moderating effects of the professors' teaching quality, age and class size on all dependent variables.

Method

Participants

Professors

In total, 95 university professors met the following inclusion criteria: 1) professors were scheduled to give a minimum of 3 lectures during a course in 2009-2010; 2) the number of enrolled students in the course was at least 25; and 3) professors did not follow any other professional development program while participating in this study. Of the 95 professors invited to participate, 87 agreed to participate. During the experiment, 12 professors dropped out for reasons unrelated to one of the conditions (e.g., illness, rescheduling). This resulted in a final sample of 75 professors (63 male, 12 female, 12 teaching assistants, 45 assistant or associate professors, 18 full professors, M_{age} = 46.8, SD_{age} = 9.6) from the departments of Economics (N = 24), Law (N = 20), Humanities (N = 5), Social and Behavioral Sciences (N = 13), and Science (N = 13). Post-test A and B (see below) were administered to these professors. Not all professors completed the post-tests. Some professors did not respond to

the request (by email) and a reminder to complete the post-tests for unknown reasons. In the control condition (N = 25), 24 professors completed post-test part A and 22 professors completed post-test part B. In the feedback-only condition (N = 24), 20 professors completed post-test part A and B. In the feedback-and-consultation condition (N = 26), 26 professor completed post-test part A and B.

Consultants

For this study, five consultants (two male, three female) were trained in SET consultation by the first author. The consultants were experienced university professors and/or faculty development staff. The consultants were trained in using a collaborative approach to consultation. In the collaborative approach, the consultant serves as a facilitator, who encourages the professor to reflect on the current situation, teaching effectiveness, and possible alternative teaching strategies to achieve his or her goals (Brinko, 1990). To standardize the consultation process, a consultation protocol (see procedure) was instated, and there were regular meetings between the consultants and first author.

Procedure

Professors were assigned to the control condition (N = 25), feedback-only condition (N = 24), or the feedback–plus-consultation condition (N = 26) according to a randomized block design. In this design, professors were grouped according to their department (departments of Law, Economics, Science, Social and Behavioral Sciences, and Humanities) and the quality of their teaching (high vs. medium quality) based on previous course evaluations. The participating professors made available course evaluation ratings of the same or a similar course that they had given in the previous academic year. The course evaluation instruments and questions differed in formulation and scale. The professor's quality was therefore recoded into two categories, high quality and medium quality professors (there were no notably low quality professors), based on the questions related to the quality of the professor. Professors with a mean rating of 8 or higher on relevant ten-point scale questions or ratings of 4 or higher on five-point scale questions were considered to be high quality professors (*High Quality*: coded as 1). Professors with lower ratings fell in the category medium quality professors (*Medium Quality*: coded as 0). The students' total mean ratings on the first measurement occasion in this study, confirmed significant higher ratings for high quality professors compared to medium quality professors, according to a multilevel t-test (high quality: β = .373, SE = .0798, $t(1)$ = 4.68). The randomized block design resulted in ten groups of professors. Professors

of the same department and quality of teaching were randomly assigned to one of three conditions to assure equal distributions of these two variables across the three conditions.

The average course took eight weeks with one 90-minute lecture a week. In all three conditions, students evaluated three 90-minute lectures during the course by completing the Instructional Skills Questionnaire (ISQ, see Chapter 2). The ISQ measures seven dimensions of lecturing as stated in the introduction of this chapter: *Structure*, *Explication*, *Stimulation*, *Validation*, *Instruction*, *Comprehension*, and *Activation*. Students were instructed to focus on the current 90-minute lecture, while completing the ISQ.

The median class sizes, in terms of ISQ forms completed, were 48.5 (*min* = 10, *max* = 365, *M* = 72.3, *SD* = 64.8), 46 (*min* = 11, *max* = 215, *M* = 66.7, *SD* = 53.1), and 43 students (*min* = 8, *max* = 190, *M* = 57.9, *SD* = 38.2) in the control, the feedback-only, and feedback–plus-consultation condition, respectively.

Control condition (*N* = 25)

At the end of their course, professors in the control condition were requested to complete the post-test part A. After completion, the professors received their ISQ results pertaining to their three lectures. Professors where then requested to complete post-test part B after studying the ISQ results.

Feedback-only condition (*N* = 24)

Professors in the feedback-only condition received their ISQ results three times, within a week after each evaluated lecture by email. They were free to use the results as they saw fit. At the end of the course, professors were requested to complete the post-test part A and B.

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

In the feedback–plus-consultation condition, each professor met with a consultant between each evaluated lecture to discuss the ISQ-results. In total there were four meetings: an introductory meeting meant for the consultant and professor to get acquainted, two consultation meetings (after the first and after the second evaluated lecture) and a final meeting to evaluate the program (after the third evaluated lecture).

The consultation protocol was based on the collaborative approach and involved a five-step procedure. The steps were 1) the evaluation of the previous lecture, 2) the evaluation of the student ratings, 3) the selection of items pertaining to specific dimensions of the ISQ that were considered to be open to improvement, 4) the analysis of the current situation

and problems that explain the selected ratings, and 5) the formulation of strategies for improvement.

The consultant's role was to facilitate behavioral change. The professor decided which items of the ISQ, as completed by their students, were to be addressed, the identification of areas of improvement, and the formulation of an action plan. The consultants were free to adopt a more directive role, if they considered that expedient; e.g., by providing alternative interpretations of student ratings, alternative views when exploring problems in teaching effectiveness, and alternative strategies for improvement. Nevertheless, as stipulated in the protocol, every step of the consultation started and ended with the professor's opinion and conclusions. In the collaborative approach it is important that the professor accepts and identifies with the conclusions of the consultation process.

The second consultation followed the same protocol as the first. At the beginning of the second consultation, the professor reported on his or her experiences in carrying out the plans made during the previous consultation. The consultant encouraged the professor to reflect on causes of success or failure. In the final meeting, 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. At the end of the course (approximately two weeks later), professors were requested to complete the post-tests part A and B, which are discussed below.

Dependent variables

Post-test A

Post-test A was developed for this experiment similar to a facet design (Guttman, 1954). The purpose of this approach was to arrive at a systematic set of statements on different concepts regarding professors' learning (knowledge, attitude, focus of attention, skills) by teaching phase (designing, teaching and evaluating ones lecture) and by teaching dimension (structure, explication, stimulation, validation, instruction, comprehension, and activation). As a result, 40 items were formulated; 12 items on domains of learning (4) * teaching phase (3) and 28 items on domains of learning (4) * teaching dimension (7).

First, the 12 items on the domains of learning*teaching phase were systematically formulated. For example, the three questions concerning *improved skills* on each teaching phase were phrased as follows: "During this course, I've became better at designing a lecture" (phase: designing), "During this course, I've became better at teaching a lecture" (phase: teaching), and "During this course, I've became better at evaluating my lectures" (phase:

evaluating). The questions related to the other domains of learning (knowledge*teaching phase, attitude*teaching phase and focus of attention*teaching phase) were formulated in similar ways.

Next, the 28 items on the domains of learning * teaching dimension were systematically formulated. For example, the questions concerning the *knowledge gained* on each teaching dimension were phrased as follows. The questions started with the statement: "During the period of this course *I gained new ideas on the following*:". This statement was followed by the specified questions: "...structuring the subject matter" (*Structure*), "...providing clear explanations of the difficult aspects of the subject matter" (*Explication*), "...enlivening the subject matter" (*Stimulation*), "...clarifying the relevance of the subject matter" (*Validation*), "...indicating what is expected of students" (*Instruction*), "...checking whether students understand the subject matter and providing opportunity for questions" (*Comprehension*), "...encouraging students to think along during the lecture" (*Activation*).

The same seven specific sentences were used to investigate professors' perceptions of *attitude change*, *increased focus of attention* and *gained skills*. Each time the seven questions started with a different statement related to the domains of learning. The complete Post-test A form is given in Appendix I. All items were measured with 5-point Likert scales (strongly disagree-strongly agree).

As mentioned in the procedure section, post-test A was administered to professors in the control condition *before* they received the student feedback (ISQ results). This procedure was followed to be able to compare the specific learning outcomes of professors who did not receive any feedback during the course, with the specific learning outcomes of professors in the two experimental conditions.

Post-test B

Post-test B contained items to measure the professors' satisfaction with the complete program, the lecture evaluations, and the consultation (feedback-plus-consultation condition only). Furthermore, post-test B measured professors' perceived satisfaction with their learning outcomes in general.

Satisfaction with the *program* was assessed by asking professors whether they thought their time was well spent, and to what extent they did *not* find the program useful. Satisfaction with *lecture evaluations* was assessed by asking professors whether they found lecture evaluations useful in improving their teaching, and a useful supplement to regular end-of-the-course evaluations. Professors were also asked whether they would recommend lecture evaluations to junior and/or senior colleagues. Satisfaction with *consultation* was

assessed by asking professors in the feedback-plus-consultation condition whether they were satisfied with the quality of the consultant, whether they found consultation useful for their teaching, and a useful supplement to the lecture evaluations. Professors were also asked whether they would recommend consultation to junior and/or senior colleagues.

In addition to the outcome variables pertaining to post-test A, professors in the two experimental conditions were asked whether they had made plans for improvement on the basis of the first and second lecture evaluation. Professors in all three conditions were asked whether they made plans to improve their teaching in their next course. Finally, the professors were asked how much they felt they had learned from the program to which they were assigned. The complete Post-test B form is given in Appendix II.

As mentioned in the procedure section, post-test B was administered to professors in the control condition *after* they received the student feedback (ISQ results). This procedure was followed to be able to question professors in the control condition about their satisfaction with the lecture evaluations and future plans for improvement based on the ISQ results.

Statistical analyses

Group differences on each post-test item were analyzed with a one-way ANOVA, with *Condition* (control condition, feedback-only condition, and feedback-plus-consultation condition) as between subject factor. We chose to investigate each item of post-test A and B separately, because we were interested in detailed findings on professors' self-reported changes. Considering the number of tests on multiple dependent variables, we adopted a test-wise alpha of .01. When the *F*-test result on a dependent variable was significant, we tested the specific contrasts *Feedback-only_versus_Control*, and Feedback-plus-consultation_versus_Control, both with an alpha of .01.

Cohen's *d* was calculated on each post-test item of each experimental condition compared to the control condition, to indicate the size of the differences between these conditions in terms of their standard deviation. We calculated Cohen's *d* by dividing the difference in mean ratings (M_E - M_C) by its pooled standard deviation ($\sqrt{}$ (($(N_C - 1)\ SD_C^2 + (N_E-1)\ SD_E^2$) /($N_C + N_E$ - 2)). Cohen's *d* of .2, .5 and .8 are generally accepted to represent small, medium, and large effects, respectively (Cohen, 1988).

Finally, the moderating effects of professors' *Age*, baseline *Quality of Teaching* (see procedure) and *Class Size* on each post-test item were analyzed with three one-way ANCOVAs. The covariate *Class Size* represented the number of students who completed the student ratings form on the first measurement occasion (baseline ratings) ($M = 80.5$, $SD = 65.3$,

min = 13, *max* = 365). The analyses included *Condition* as between subject factor, either *Age*, *Quality of Teaching* or *Class Size* as covariate and the interaction term *Condition*Covariate*. The main effects and the interaction effects were tested with an alpha of .01.

Results

Perceived satisfaction with the programs

Perceived satisfaction (Guskey's first level of evaluation) was measured with items of post-test B. Table 4.1 contains descriptives concerning the satisfaction of the professors with the complete program, with the lecture evaluations and with the consultation (in the condition including consultation). Table 4.1 also contains the results of tests of differences between the three conditions concerning the satisfaction with the complete program and the evaluation.

Considering mean ratings, professors in the feedback-plus-consultation condition were most satisfied with the complete program and lecture evaluations.

Additionally, they reported to be satisfied with the consultation itself (mean ratings 4.00, or higher, on a 5-point Likert scale). Compared to the professors in the control condition, they considered the lecture evaluations to be more useful to improve their teaching (Cohen's d = .80, p = .013). The feedback-only condition did not differ from the control condition with respect to perceived usefulness of the lecture evaluations, notwithstanding the difference between the feedback-only condition and the control condition in timing of the feedback. The professors in the three conditions did not differ on the other items related to satisfaction. In general, professors in all three conditions were positive about the lecture evaluation and would recommend them to their colleagues (particularly to their junior colleagues).

Perceived learning outcomes of the programs

Professors' self-reported learning outcomes (Guskey's second level of evaluation) was measured with forty items of post-test A, relating to the domains of learning (knowledge, attitude, focus of attention and skills), and 4 items of post-test B, relating to plans made for improvement of teaching and relating to learned from the program (evaluations and consultation) in general. Table 4.2 contains descriptives of the detailed outcome measures, as evaluated by the professors, and one-way ANOVA tests of the differences between the conditions.

Table 4.1 Descriptives in the three conditions concerning satisfaction with the complete program, lecture evaluations and consultation and results of tests of the differences between the conditions

	Condition														
	Control			Feedback-only				Feedback-plus-consultation							
	N	*M*	(*SD*)	*N*	*M*	(*SD*)	Cohen's *d* compared to control	*N*	*M*	(*SD*)	Cohen's *d* compared to control	*F*		η^2
Complete program														
Time well spent	21	3.57	(0.93)	19	3.53	(0.84)	-0.05	26	3.77	(0.99)	0.21	0.45		0.01
Program was *not* useful	21	3.43	(1.03)	19	3.47	(1.43)	0.04	26	2.54	(1.42)	-0.71	3.81		0.11
Lecture evaluation														
useful to improve teaching	21	3.14	(0.96)	19	2.68	(1.25)	-0.41	26	3.96	(1.08)	0.80	7.93	***	0.20
useful as a supplement to course evaluations	21	3.10	(1.14)	19	3.26	(1.28)	0.14	26	3.81	(1.10)	0.64	2.43		0.07
recommend to a junior colleague	21	3.86	(1.20)	19	3.84	(1.46)	-0.01	26	4.04	(1.22)	0.15	0.17		0.01
recommend to a senior colleague	21	3.38	(1.20)	19	3.32	(1.42)	-0.05	26	3.85	(1.08)	0.41	1.31		0.04
Consultation														
satisfied with quality of the consultant								26	4.35	(0.75)				
useful to improve teaching								26	4.12	(0.91)				
useful as a supplement to lecture evaluations								26	4.23	(0.99)				
recommend to a junior colleague								26	4.19	(1.10)				
recommend to a senior colleague								26	4.00	(1.02)				

Note: ** $p < .01$, *** $p < .001$. Cohen's *d* was calculated by dividing the difference in mean ratings of the experimental and control condition ($M_E - M_C$) by its pooled standard deviation ($\sqrt{(((N_C - 1)\, SD_C^2 + (N_E - 1)\, SD_E^2) / (N_C + N_E - 2)}$). Each *F*-test has $df_1 = 2$ and $df_2 = 63$ degrees of freedom. Eta-squared expresses the effect size. Contrast analyzes were conducted when the *F*-test was significant with $p < 0.01$. No contrast analyzes are indicated with '-'.

Table 4.2 Descriptives and comparison between three conditions on outcome variables of the programs

	Condition																
	Control			Feedback-only					Feedback-plus-consultation								
	N	*M*	(*SD*)	*N*	*M*	(*SD*)		Cohen's *d* compared to control	*N*	*M*	(*SD*)		Cohen's *d* compared to control	df_1	df_2	*F*		η^2
Gained knowledge on																		
Designing	24	2.33	(1.05)	20	2.85	(1.27)		0.45	25	3.68	(1.03)	***	1.30	2	66	9.178	***	0.22
Teaching	24	2.29	(1.12)	20	2.55	(1.00)		0.24	25	3.56	(0.82)	***	1.29	2	66	11.252	***	0.25
Evaluating	24	2.75	(1.19)	20	3.65	(1.35)	**	0.71	25	4.20	(0.82)	***	1.43	2	66	10.375	***	0.24
Teaching dimensions:																		
Structure	24	2.88	(1.36)	20	2.40	(1.31)	-	-0.35	25	3.28	(1.24)	-	0.31	2	66	2.526		0.07
Explication	24	2.75	(1.33)	20	2.20	(1.06)	-	-0.45	25	2.72	(1.10)	-	-0.02	2	66	1.481		0.04
Stimulation	24	2.58	(0.93)	20	2.20	(1.11)		-0.38	25	3.72	(1.10)	***	1.11	2	66	13.278	***	0.29
Validation	24	2.58	(1.10)	20	2.55	(1.19)	-	-0.03	25	3.40	(1.08)	-	0.75	2	66	4.402		0.12
Instruction	24	2.63	(1.13)	20	2.70	(1.08)		0.07	25	3.64	(1.11)	**	0.90	2	66	6.229	**	0.16
Comprehension	24	2.46	(1.10)	20	2.20	(0.95)		-0.25	24	3.25	(1.03)		0.74	2	65	6.328	**	0.16
Activation	24	2.46	(1.14)	20	2.30	(1.08)	-	-0.14	25	3.16	(1.07)	-	0.64	2	66	4.073		0.11
Attitude change on																		
Designing	24	2.46	(1.02)	20	2.25	(1.16)	-	-0.19	25	2.88	(1.13)	-	0.39	2	66	1.940		0.06
Teaching	24	2.50	(1.06)	20	2.50	(1.10)	-	0.00	25	2.92	(1.00)	-	0.41	2	66	1.274		0.04
Evaluating	24	2.54	(1.18)	20	2.55	(1.19)	-	0.01	25	3.24	(1.13)	-	0.61	2	66	2.837		0.08
Teaching dimensions:																		
Structure	24	2.46	(1.22)	20	2.25	(1.16)	-	-0.17	25	2.88	(1.09)	-	0.37	2	66	1.765		0.05
Explication	24	2.54	(1.25)	20	2.40	(1.27)	-	-0.11	25	2.92	(1.00)	-	0.34	2	66	1.218		0.04
Stimulation	24	2.33	(1.09)	20	2.50	(1.15)		0.15	25	3.44	(1.08)	***	1.02	2	66	7.070	**	0.18
Validation	24	2.42	(1.14)	20	2.35	(1.14)		-0.06	25	3.32	(0.95)	**	0.86	2	66	6.069	**	0.16
Instruction	24	2.46	(1.02)	20	2.45	(1.19)		-0.01	25	3.48	(1.05)	**	0.99	2	66	7.171	**	0.18
Comprehension	24	2.25	(1.03)	20	2.40	(1.14)	-	0.14	25	3.08	(1.00)	-	0.82	2	66	4.285		0.12
Activation	24	2.58	(1.10)	20	2.45	(1.28)	-	-0.11	25	3.24	(1.09)	-	0.60	2	66	3.172		0.09

Table 4.2 *Continued*

	Condition																
	Control			Feedback-only					Feedback-plus-consultation								
	N	*M*	(*SD*)	*N*	*M*	(*SD*)		Cohen's *d* compared to control	*N*	*M*	(*SD*)		Cohen's *d* compared to control	df_1	df_2	*F*		η^2
Increased focus of attention to																		
Designing	23	2.39	(1.20)	20	2.50	(1.40)		0.08	25	3.16	(1.25)	-	0.63	2	65	2.542		0.07
Teaching	23	2.17	(1.07)	20	2.60	(1.35)	-	0.35	25	3.64	(1.15)	***	1.32	2	65	9.678	***	0.23
Evaluating	23	2.48	(1.12)	20	2.95	(1.32)	-	0.39	25	3.96	(0.73)	***	1.58	2	65	12.167	***	0.27
Teaching dimensions:																		
Structure	23	2.61	(1.20)	20	2.35	(1.23)		-0.21	25	2.96	(1.21)	-	0.29	2	65	1.448		0.04
Explication	23	2.57	(1.27)	20	2.30	(1.08)		-0.22	25	2.84	(1.14)	-	0.23	2	65	1.187		0.04
Stimulation	23	2.39	(1.12)	20	2.45	(1.10)	-	0.05	25	3.56	(1.23)	***	0.99	2	65	7.744	**	0.19
Validation	23	2.52	(1.20)	20	2.30	(0.92)	-	-0.21	25	3.36	(1.22)		0.69	2	65	5.644	**	0.15
Instruction	23	2.22	(1.00)	20	2.40	(1.23)	-	0.16	25	3.56	(0.92)	***	1.40	2	65	11.630	***	0.26
Comprehension	23	2.43	(1.12)	20	2.25	(0.97)	-	-0.18	25	3.44	(1.19)	**	0.87	2	65	7.838	***	0.19
Activation	22	2.55	(1.22)	20	2.35	(1.14)	-	-0.17	25	3.52	(1.16)	**	0.82	2	64	6.623	**	0.17

Table 4.2 continues on next page

Table 4.2 *Continued*

	Condition																
	Control			Feedback-only				Feedback-plus-consultation									
	N	*M*	(*SD*)	*N*	*M*	(*SD*)	Cohen's *d* compared to control	*N*	*M*	(*SD*)		Cohen's *d* compared to control	df_1	df_2	*F*		η^2
Improved skills on																	
Designing	23	2.52	(1.08)	20	2.25	(1.12)	-0.25	24	3.04	(1.20)	-	0.46	2	64	2.804		0.08
Teaching	23	2.78	(1.17)	20	2.30	(1.13)	-0.42	24	3.25	(0.99)	-	0.43	2	64	4.120		0.11
Evaluating	23	2.35	(1.11)	20	2.40	(1.19)	0.05	24	3.50	(1.10)	***	1.04	2	64	7.657	**	0.19
Teaching dimensions:																	
Structure	23	2.61	(1.23)	20	2.30	(1.22)	-0.25	24	2.83	(1.13)	-	0.19	2	64	1.093		0.03
Explication	23	2.61	(1.16)	20	2.10	(1.02)	-0.46	24	2.58	(1.02)	-	-0.02	2	64	1.512		0.05
Stimulation	23	2.39	(0.94)	20	2.25	(1.02)	-0.14	24	3.25	(1.19)	**	0.80	2	64	5.968	**	0.16
Validation	23	2.57	(1.08)	20	2.15	(0.99)	-0.40	24	3.17	(1.34)	-	0.49	2	64	4.328		0.12
Instruction	23	2.35	(0.93)	20	2.25	(1.21)	-0.09	24	3.38	(1.13)	**	0.99	2	64	7.447	**	0.19
Comprehension	23	2.35	(1.07)	20	2.00	(0.92)	-0.35	24	3.13	(1.08)		0.72	2	64	7.016	**	0.18
Activation	23	2.70	(1.15)	20	2.10	(1.02)	-0.55	23	3.26	(1.01)		0.52	2	63	6.391	**	0.17
Plans for improvement																	
plans after evaluation 1				18	2.67	(1.24)		25	3.84	(1.07)	**	1.03	1	41	11.070	**	0.21
plans after evaluation 2				18	2.28	(1.32)		22	3.41	(1.26)	**	0.88	1	38	7.651	**	0.17
plans for next year	20	3.20	(1.11)	19	2.53	(1.31)	-0.55	24	3.92	(1.10)		0.65	2	60	7.573	**	0.20
Learned																	
Learned from evaluations (plus consultation)	21	3.29	(0.96)	19	2.79	(1.27)	-0.45	26	4.15	(0.78)	**	0.99	2	63	10.858	***	0.26

Note: ** $p < .01$, *** $p < .001$. Cohen's *d* was calculated by dividing the difference in mean ratings of the experimental and control condition (M_E - M_C) by its pooled standard deviation ($\sqrt{(((N_C - 1)\ SD_C^2 + (N_E - 1)\ SD_E^2) / (N_C + N_E - 2))}$). Cohen's *d* on plans made after evaluation 1 and 2 were calculated on differences between the two experimental conditions. Eta-squared expresses the effect size. Contrast analyzes were conducted when the *F*-test was significant with $p < 0.01$. No contrast analyzes are indicated with '-'.

We found a significant main effect of condition ($p < .01$) on twenty-five out of forty-four outcome variables. Most effects concerned variables related to professors self-reported *gained knowledge* on the teaching phases and teaching dimensions, *increased focus of attention* to the teaching phases and teaching dimensions, *plans made for improvement of teaching* and the amount that professors perceived to *have learned from the program* in general. Planned contrasts of the two experimental conditions versus the control condition showed that nearly all significant *F* values were due to differences between the feedback-plus-consultation condition and the control condition. With respect to twenty of these twenty-five variables, including the variable how much was *learned from the program* in general, the differences between the feedback-plus-consultation condition and the control condition were significant ($p < .01$). The effect sizes were large, with Cohen's *d* ranging from .80 to 1.58 (mean value of 1.08).

More specifically, professors in the feedback-plus-consultation condition reported significant changes, compared to the control condition, in knowledge, focus of attention, as well as skills at *evaluating* their lectures. In addition, we found significant differences compared to the control condition in self-reported gained knowledge on how to *teach* and *design* lectures and increased focus of attention to *teaching*. On the teaching dimensions *Stimulation* and *Instruction*, professors reported to have gained more knowledge, changed their attitudes, increased their focus of attention as well as improved their skills. Additionally, compared to the control condition, the dimension *Validation* became more important to them (attitude change) and they stated to be more attentive to the dimensions *Comprehension* and *Activation*. Compared to the feedback-only condition, professors in the feedback-plus-consultation condition made significantly ($p < .01$) *more plans for improvement* during their course. These differences were large as well (after evaluation 1: Cohen's $d = 1.03$, after evaluation 2: Cohen's $d = .88$).

The feedback-only condition showed a significant difference with the control condition on only one variable; *gained knowledge* on *evaluating* their lectures (item: "During this course, I came to know more about how students experience my lectures", Cohen's $d = .71$). In short, according to professors themselves, intermediate feedback-plus-consultation had a significant impact, while intermediate feedback-only had little effect.

Analyses with the professors' *Age*, baseline *Quality of Teaching*, and *Class Size* as covariates showed one significant main effect of age on the dependent variable *improved skills on teaching* ($F(5,61) = 8{,}740$, $p = .004$, $eta^2 = .125$), indicating that in all three conditions younger professors improved more than older professors. There were no significant interaction effects (*Condition*Covariate*) with all three covariates on any of the other dependent variables, meaning that professors in the three conditions did not differ on the outcome variables depending on professors' age, baseline quality of teaching and class size.

Discussion

The aim of the present study was to investigate the detailed impact of intermediate student feedback on specific lectures with and without individual consultation on professors' self-perceived learning in an experimental design, including a control condition. First, we investigated professors' satisfaction with the interventions (Guskey's first level of effect; Guskey, 2000). Second, we studied the effects of these interventions on professors' self-assessed knowledge, attitudes, focus of attention and skills on lecturing (Guskey's second level of effect). Third, the effects on their planning of improvements to their teaching, and the effects on the perceived general benefit of the interventions were investigated. Finally, the moderating effects of professors' age, baseline quality of teaching and class size on all dependent variables were analyzed.

With respect to Guskey's first level of effects (satisfaction), we found that 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). Additionally, we found that professors in the feedback-plus-consultation condition considered lecture evaluations to be most useful to improve their teaching, compared to the professors in the other conditions. Professors in the feedback-plus-consultation condition stated that they would recommend consultation to both junior and senior colleagues.

With respect to Guskey's second level of effects (learning) we found that, compared to the control condition, the professors in the feedback-plus-consultation condition benefitted considerably from the intervention. In contrast, the professors in the feedback-only condition did not differ appreciably from the professors in the control condition. In terms of Cohen's *d*, significant effects of intermediate feedback-plus-consultation were large ($d > .8$) on twenty outcome variables, including the amount that professors perceived to have *learned from the program* in general. The feedback-only condition showed only one medium effect on the outcome variable; *gained knowledge* on evaluating their lectures (item: "During this course, I came to know more about how students experience my lectures").

Considering specific teaching phases, professors in the feedback-plus-consultation condition reported to have *gained more knowledge* on how to design, teach and evaluate their lecture, *increased their focus of attention* to their teaching and to evaluating their lectures and *became more skilled* in evaluating their lectures. Considering specific teaching dimensions, we observed significant changes in *knowledge*, *attitude*, *focus of attention* and *skills* on the ISQ dimensions Stimulation and Instruction. Validation *became more important* as well and professors *increased their focus of attention* to *Comprehension* and *Activation*. Furthermore, the feedback-plus-consultation condition made significantly ($p < .01$) *more*

plans for improvement during their course, compared to the feedback-only condition. These differences were large as well.

Given the selected sample and the randomized block design, we expect the results to generalize to most professors teaching comparable courses (with respect to length, teaching format, and general organization) at this university. Additionally, analyzes showed that professors' baseline quality of teaching, their age and their class size did not influence these results.

We note that we did not observe improvements in the feedback-plus-consultation on all teaching dimensions. Specifically we found no effects on the teaching dimensions Structure and Explication. Possibly, most professors had little to gain with respect to structuring a lecture and explaining the subject matter, as these teaching dimensions are basic to designing and teaching a lecture. Structure and Explication can be characterized as teacher-focused teaching dimensions aimed at knowledge transmission, while the other five dimensions can be characterized as student-focused teaching dimensions aimed at facilitating student learning. According to the literature, a student focused approach to teaching is more demanding than a teacher focused approach (Saroyan & Snell, 1997). It requires greater advance planning and preparation, a greater involvement of the student in the instruction, *and the ability to incorporate pedagogical principles relating to the delivery of instruction to facilitate the student learning process* (Saroyan & Snell, 1997). Effective professors tend to adopt a more student-focused teaching approach and this approach is associated with a deeper approach to learning by students (Gow & Kember, 1993; Prosser & Trigwell, 1998; Saroyan & Snell, 1997; Trigwell, Prosser, & Waterhouse, 1999; Young & Shaw, 1999). As professors in the present study were all members of the existing faculty staff, they are likely to be competent with respect to structuring and explaining the subject matter, and thus, as these findings clearly indicate, they made more progress on student-focused teaching dimensions such as Stimulation and Instruction.

Secondly, we note that fewer effects were found on self-assessed improved skills on teaching dimensions (items starting with "I became better at..."). Professors might need more time to successfully implement new strategies, and improve their teaching skills. Guskey (2000) noted that the most worthwhile changes in education require time for adaption, adjustment, and refinement.

The present study is limited in that the interventions occurred in a relatively short period of time; courses at this university only last eight weeks and are mostly thought once a year. Also, plans such as changing the amount of subject matter are more difficult to implement during the course, due to the fixed set up of most of the courses. Longer lasting interventions

and follow up investigation were therefore not feasible in the present study. We recommend longer lasting investigations on actual skills development at other universities for future research. In addition, the present study is limited to professors' self-reports on learning. No objective assessments were used to measure changes in professors' teaching competencies. Self-reports risk to be subject to socially desirable answers. Nonetheless, professors in the three conditions differ significantly in their self-reported ratings and they show variance within each condition in ratings on different outcome variables. To provide additional validity to these findings, further research needs to be conducted to complement these results with other objective measures (see Chapter 5).

The present results are consistent with previous findings in reviews on the effects of intermediate feedback-only and feedback-plus-consultation found on student ratings (Cohen, 1980; Menges & Brinko, 1990; Penny & Coe, 2004). Specifically, these reviews indicated that the effects of intermediate feedback-only on student 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 that intermediate feedback and consultation had on professors' self-perceived learning. Apparently, feedback only does not influence professors' perception of their knowledge, focus of attention, attitudes nor skills regardless of professors' age, baseline quality of teaching or class size. Even though the results may only be generalized to the university at which this study was conducted, these findings hint that the efforts undertaken at many universities to provide professors with feedback to improve their teaching probably require supplemental support.

In summary, the present results indicate a considerable impact of combining student feedback with individual consultation on professors' self-assessed learning. In addition, professors found the collaborative approach to be useful and recommendable to colleagues. Considering the impact of consultation and the substantial differences with the limited impact of intermediate feedback-only, we conclude that collaborative consultation based on student feedback is recommendable in faculty development.

References

Brinko, K.T. (1990). Instructional consultation with feedback in higher education. *Journal of Higher Education, 61,* 65-83.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.)*. New Jersey: Lawrence Erlbaum.

De Neve, H.M.F. (1991). University teachers' thinking about lecturing: student evaluation of lecturing as an improvement perspective for the lecturer, *Higher Education, 22*, 63-89.

Gow, L., & Kember, D. (1993). Conceptions of teaching and their relationship to student learning. *British Journal of Educational Psychology, 63,* 20-23.

Guskey, T.R. (2000). *Evaluating professional development.* Thousand Oaks, CA: Corwin Press.

Guttman, L. (1954). An outline of some new methodology for social research. *Public Opinion Quarterly, 18,* 395-404.

Kember, D., Leung, D.Y.P., & Kwan, K.P. (2002). Does the use of student feedback questionnaires improve the overall quality of teaching? *Assessment & Evaluation in Higher Education, 27,* 411-425.

Kirkpatrick, D.L. (1994). *Evaluating training programs: The four levels.* San Francisco, CA: Berrett-Koehler Publishers.

Levinson-Rose, J., & Menges, R.J. (1981). Improving college teaching: A critical review of research. *Review of Educational Research, 51,* 403-434.

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

Marsh, H.W., & Hocevar, D. (1991b). Students' evaluations of teaching effectiveness: The stability of mean ratings of the same teachers over a 13-year period. *Teaching and Teacher Education, 7,* 303-314.

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

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

Prebble, T., Hargraves, H., Leach, L., Naidoo, K., Suddaby, G., & Zepke, N. (2004). *Impact of student support services and academic development programmes on student outcomes in undergraduate tertiary study: A synthesis of the research.* Report to the Ministry of Education, Massey University College of Education.

Prosser, M., & Trigwell, K. (1998). *Understanding Learning and Teaching: The Experience in Higher Education*. Milton Keynes: Open University Press.

Saroyan, A., & Snell (1997). Variation in lecturing styles. *Higher Education, 33,* 85-104.

Steinert, Y., Mann, K., Centeno, A., Dolmans, D., Spencer, J., Gelula, M., & Prideaux, D. (2006). A systematic review of faculty development initiatives designed to improve teaching effectiveness in medical education: BEME Guide No. 8. *Medical Teacher, 28,* 497-526.

Stes, A., Min-Leliveld, M., Gijbels, D., & Van Petegem, P. (2010). The impact of instructional development in higher education: The state-of-the-art of the research. *Educational Research Review, 5*, 25-49.

Theall, M., & Franklin, J. (2001). Looking for Bias in All the Wrong Places: A Search for Truth or a Witch Hunt in Student Ratings of Instruction?. In Theall, M. P. Abrami, and Mets, L. (Eds.), *The Student Ratings Debate: Are they Valid? How Can We Best use Them? New Directions for Institutional Research, 109,* 45-56. San Francisco, CA: Jossey-Bass.

Trigwell, K., Prosser, M., & Waterhouse, F. (1999). Relations between teachers' approaches to teaching and students' approaches to learning. *Higher Education, 37,* 57-70.

Weimer, M., & Lenze, L.F. (1997). Instructional interventions: A review of the literature on efforts to improve instruction. In K. R. Perry, & J. C. Smart (Eds.), *Effective teaching in higher education: Research and practice,* 205-240. New York, NY: Agathon Press.

Young, S., & Shaw, D.G. (1999). Profiles of effective college and university teachers. *The Journal of Higher Education, 70,* 670-686.

Appendix I

Post-test A

Measured with 5-point Likert scales (strongly disagree-strongly agree)

Variables		Items
Gained knowledge on		
Teaching phases	**Designing**	During this course, I came to know more about which aspects are important in designing a lecture
	Teaching	During this course, I came to know more about teaching strategies I can use during a lecture
	Evaluating	During this course, I came to know more about how students experience my lectures
Teaching dimensions		During the period of this course *I gained new ideas on the following:*
	Structure	… structuring the subject matter
	Explication	… providing clear explanations of the difficult aspects of the subject matter
	Stimulation	… enlivening the subject matter
	Validation	… clarifying the relevance of the subject matter
	Instruction	… indicating what is expected of students
	Comprehension	… checking whether students understand the subject matter and providing opportunity for questions
	Activation	… encouraging students to think along during the lecture
Attitude change on		
Teaching phases	**Designing**	During this course, instructional design became more important to me
	Teaching	During this course, the way I teach became more important to me
	Evaluating	During this course, find out how students have experienced my lectures became more important to me
Teaching dimensions		During the period of this course *the following became more important to me:*
	Structure	… structuring the subject matter
	Explication	… providing clear explanations of the difficult aspects of the subject matter
	Stimulation	… enlivening the subject matter
	Validation	… clarifying the relevance of the subject matter
	Instruction	… indicating what is expected of students
	Comprehension	… checking whether students understand the subject matter and providing opportunity for questions
	Activation	… encouraging students to think along during the lecture

Appendix I continues on next page

Appendix I *Continued*

Variables		Items
Increased focus of attention to		
Teaching phases	**Designing**	During this course, I've paid more attention to the instructional design of my lectures
	Teaching	During this course, I've paid more attention to how I teach
	Evaluating	During this course, I've thought more about how my lectures went
Teaching dimensions		During the period of this course *I paid more attention to the following:*
	Structure	… structuring the subject matter
	Explication	… providing clear explanations of the difficult aspects of the subject matter
	Stimulation	… enlivening the subject matter
	Validation	… clarifying the relevance of the subject matter
	Instruction	… indicating what is expected of students
	Comprehension	… checking whether students understand the subject matter and providing opportunity for questions
	Activation	… encouraging students to think along during the lecture
Improved skills on		
Teaching phases	**Designing**	During this course, I've became better at designing a lecture
	Teaching	During this course, I've became better at teaching a lecture
	Evaluating	During this course, I've became better at evaluating my lectures
Teaching dimensions		During the period of this course *I became better at the following:*
	Structure	… structuring the subject matter
	Explication	… providing clear explanations of the difficult aspects of the subject matter
	Stimulation	… enlivening the subject matter
	Validation	… clarifying the relevance of the subject matter
	Instruction	… indicating what is expected of students
	Comprehension	… checking whether students understand the subject matter and providing opportunity for questions
	Activation	… encouraging students to think along during the lecture

Appendix II

Post-test B

Measured with 5-point Likert scales (strongly disagree-strongly agree)

	Abbreviation	Item
Satisfaction		
Complete program	Time well spent	The time that this program cost me was well spent
	Program was *not* useful	I did *not* find this program useful for my teaching
Lecture evaluation	useful for improvement of teaching	Lecture evaluations were useful in improving of my teaching
	useful in addition to course evaluations	Lecture evaluations were a useful supplement to regular end-of-the-course evaluations
	recommend to a junior colleague	I would recommend lecture evaluations to a junior colleague
	recommend to a senior colleague	I would recommend lecture evaluations to a senior colleague
Consultation	satisfied with quality of the consultant	I am satisfied with the quality of the consultant
	useful for improvement of teaching	I found the consultation useful for my teaching
	useful in addition to lecture evaluations	The consultation was a useful supplement to the lecture evaluations
	recommend to a junior colleague	I would recommend consultation to a junior colleague
	recommend to a senior colleague	I would recommend consultation to a senior colleague
Outcome variables		
Plans for improvement	plans after evaluation 1	Following the first lecture evaluation, I made plans to improve my following lectures
	plans after evaluation 2	Following the second lecture evaluation, I made plans to improve my subsequent lectures
	plans for next course	Following the complete program, I made plans to improve my teaching in my next course
Learned	Learned from evaluations (plus consultation)	I learned from the lecture evaluations (plus consultation)

5

Experiment II:

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

Abstract

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

Introduction

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

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

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

Research on SETs and SET consultation

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

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

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

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

Penny and Coe (2004) studied the predictors of successful SET consultation. They investigated differences between studies in terms of features, participant characteristics, form of consultation and consultation components. They identified several consultation strategies that appeared to be important (e.g., active involvement of the professor in the learning process, sufficient time for dialogue, and use of high-quality feedback information). Ultimately, they

noted that none of the differences proved to be significant, possibly due to low statistical power given the small number of studies available. They concluded: "Thus the most robust finding may be that more [experimental] research is needed" (p. 236).

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

Approaches to consultation

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

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

These recent findings are important, because they address biasing variables, the importance of using appropriate multilevel procedures and effects in a non-English speaking country for the first time. Whether non-experimental results are due to the intervention or due to alternative explanations remains an open question. For example, in some studies the participating professors approached a teacher training center and were highly motivated to change. Additionally, the effects of a specific approach may be due to a Hawthorne effect (the

attention/social treatment one receives). Dresel and Rindermann (2011) noted the difficulty of conducting research, which is both internally and externally valid. With the present study we aim to augment recent non-experimental findings with up-to-date experimental results for a collaborative approach to consultation.

The present study

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

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

We used multilevel modeling to analyze the data. In the past, this statistical approach was poorly disseminated in terms of user friendly software. At present, many programs are available, in the form of dedicated software (MLwiN; Rasbash, Charlton, Browne, Healy & Cameron, 2009) and procedures (SPSS linear mixed; 2007) and libraries (R nlme; Pinheiro, Bates, DebRoy, Sarkar & the R Development Core Team, 2012). Multilevel regression analysis

allowed us to take the nested structure of the data into account (students were nested within classes and measurement occasions were nested within students). By doing so, we were able to account for the clustering due to random differences between the lectures, the students and the professors.

In summary the present study addresses the following research questions:

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

Context information

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

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

Method

Participants

Professors

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

Students

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

A mean response rate of 90.2% was observed in 76 randomly selected lectures. The median class sizes, in terms of ISQ forms completed, over the three measurement occasions were 48.5 students (*min* = 10, *max* = 365, M = 72.3, SD = 64.8), 46 students (*min* = 11, *max* = 215, M = 66.7, SD = 53.1), and 43 students (*min* = 8, *max* = 190, M = 57.9, SD = 38.2) in the control, the feedback-only, and feedback–plus-consultation condition, respectively.

There was an expected decrease in attendance over time in all three conditions, as a number of students invariably drop out of courses (N_{ISQ_forms} on T_1: 5,900, T_2: 4,649, T_3: 3,749). In the control condition more students dropped out compared to the two experimental conditions (control versus feedback-only: $\chi^2(2) = 15.99$, $p < .05$, control versus feedback-plus-consultation: ($\chi^2(2) = 13.56$, $p < .05$). The two experimental conditions did not differ ($\alpha = .05$) in terms of dropout ($\chi^2(2) = 4.83$, $p = .09$). This difference particularly concerned students who only attended and completed the ISQ once (instead of two or three times). Multilevel-regression analyses on the first measurement occasion (T_1) showed significantly ($\alpha = .05$) higher mean ratings for students who completed the ISQ twice (on T_1 and T_2, or T_1 and T_3) ($\beta = .127$, $SE = .023$, $t = 5.52$) or three times (on T_1, T_2 and T_3) ($\beta = .116$, $SE = .023$, $t = 5.04$) compared to once (on T_1). Ratings in the control condition might therefore be slightly biased in terms of more negative on T_1, or more positive on T_2 and T_3, compared to the experimental conditions.

Consultants

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

Procedure

The participating professors were assigned to the control condition, feedback-only condition, or the feedback–plus-consultation condition according to a randomized block design. In this design, professors were grouped according to the quality of their teaching (high vs. medium quality) based on previous course evaluations (see section on moderators for grouping information), and their department (departments of Law, Economics, Science, Social and Behavioral Sciences, and Humanities). This resulted in ten groups of professors. Professors of the same department and quality were randomly assigned to one of three conditions to assure equal distributions of these two variables across the three conditions.

Prior to the start of their courses, all professors received procedural instructions by email. Professors sent a standardized email to their students, informing them that they (i.e., the professors) would be participating in a research project on the quality of the lectures at the university. Students were invited to take part by evaluating three lectures during the course. In the final fifteen minutes of the lecture, professors reserved five minutes for an evaluation break. Research assistants distributed the questionnaires and collected them during this break. Students were explicitly instructed to evaluate the current lecture. They were asked to provide their student ID number for research purposes and were assured of anonymity in their evaluations with an extra statement on the ISQ form. The students did not know that their professors were participating in a randomized experiment.

Independent variables

Control condition (*N* = 25)

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

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

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

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

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

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

Consultation meetings. The consultation protocol, based on the collaborative approach, involved a five-step procedure: 1) the evaluation of the previous lecture, 2) the evaluation

of the student ratings, 3) the selection of items of the ISQ to improve, 4) the analysis of the current situation and problems that explain the selected ratings, and 5) the formulation of strategies for improvement.

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

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

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

Dependent variables

The dependent variables are the scores on the Instructional Skills Questionnaire (ISQ). The ISQ was based on the course-evaluation instrument of the University of Amsterdam, the Uvalon, developed by Vorst and Van Engelenburg (1992). The Uvalon was based on theories on effective instruction, and on research on the effects of teaching quality on student learning (Frey, Leonard, & Beatty, 1975; De Neve & Janssen, 1982; Janssen & De Neve, 1988; Marsh, 1984; Abrami, Apollonia & Cohen, 1990). The Uvalon contains seven dimensions on instructional behavior. Its psychometric quality was investigated and confirmed in several internal reports from the University of Amsterdam (Vorst & Van Engelenburg, 1992; Verbeek, De Jong, &Vermeulen, 2002, 2005).

The Uvalon was adapted to a one-lecture instrument, the ISQ, with a selection of specific questions on the seven dimensions on instructional behavior. With the ISQ the seven dimensions are measured with two positive (indicative) and two negative (contra-indicative) worded items on a 7-point likert scale. This resulted a total of 28 items and two open questions ("What was good about this lecture?" and "How can this lecture be improved?").

The psychometric quality of the ISQ was investigated and confirmed with confirmatory factor analyses (see Chapter 2). In more detail, the seven dimensions are:

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

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

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

The reliability of the subscales on the professor level were high. Cronbach's alphas ranged from .88 to .98, with a mean of .93 on T_1, from .92 to .98, with a mean of .94 on T_2, and from .91 to .98, with a mean of .94 on T_3. One reason that these values were quite high is that the professor scores were based on the average test scores of their students. The averages were necessarily subject to less error variance than the student level data.

Missing item responses (3.7%) were imputed with the student's mean of the other three items of that specific dimension. Out of 14,596 forms, 298 forms were excluded; 218 forms remained incomplete after imputation, and 80 forms were marked as extreme outliers. Extreme outliers were detected with the Inter Quartile Range (IQR; distance between the first and the third quartile). For each professor on each measurement occasion separately the IQR was calculated on *Total Instructional Skills*. A form was considered an extreme outlier when the rating was at least two times the IQR lower than the first quartile or two times the IQR higher than the third quartile. This equals a deviance of 3.6 times the standard deviation from the mean.

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

Moderators

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

Statistical analyses

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

Randomization check

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

Intra-class correlation

To obtain a measure of clustering, we calculated the intra-class correlations in the intercept only model, which we denote Model 1. Model 1 was fitted on data from all three measurement occasions, with time as level 1 variable, students as level 2 variable and professors as level 3 variable. Model 1 contained an intercept random over professors and students. The variances in ratings were decomposed into the variance of the ratings over time of a given student (σ^2), the variance of the ratings over students of a given professor (τ_0^2), and the variance of the ratings over professors (φ_0^2). The professor level intra-class correlation (IIC_T) and the student level intra-class correlation (IIC_S) were calculated as $ICC_T = \varphi_0^2/(\sigma^2 + \tau_0^2 + \varphi_0^2)$ and $ICC_S = \tau_0^2/(\sigma^2 + \tau_0^2 + \varphi_0^2)$, respectively. The latter quantifies the degree to which students of a given professor are alike (relative to students of different professors). The former quantifies

the degree to which the professors are alike (in terms of repeated measures by their students relative to the repeated measures of different students).

Effects of the interventions and correction for random effects

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

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

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

Cohen's *d* was calculated in two ways. The first Cohen's *d* was calculated for each condition by dividing it's mean difference of T_3 and T_1 by its pooled standard deviation ($\sqrt{((SD\ (T_{1\_condition})^2 + SD\ (T_{3\_condition})^2) / 2}$). The second Cohen's *d* was calculated for each experimental condition versus the control condition by dividing the mean difference of T_3

and T_1 of the experimental condition minus the mean difference of T_3 and T_1 of the control condition by its pooled standard deviation ($\sqrt{((SD\,(T_{1\_control})^2 + SD\,(T_{3\_control})^2 + SD\,(T_{1\_condition})^2 + SD\,(T_{3\_condition})^2)} / 4$). Multilevel effect sizes were calculated based on the multilevel modeling output of the final model. Two times the beta of *Time* represents the change in mean of the control condition between T_1 and T_3, two times the beta of *Time*ΔF* represents the change in mean of the feedback-only condition compared to the control condition between T_1 and T_3, and two times the beta of *Time*ΔFC* represents the change in mean of the feedback-plus-consultation condition compared to the control condition between T_1 and T_3. The residual standard deviation $SD(e_{tij})$ equals the standard deviation of Y_{tij} over time of a given student *i* for a given professor *j* (assuming homoskedasticity). To standardize the effect, we expressed the effect size as a function of this within student, within professor residual standard deviation. The effect size of the control condition was therefore calculated by dividing two times the beta of *Time* by the residual standard deviation $SD\,(e_{tij})$, the effect size of the feedback-only condition compared to the control condition was calculated by dividing two times the beta of *Time*ΔF* by the residual standard deviation $SD\,(e_{tij})$, and the effect size of the feedback-plus-consultation condition compared to the control condition was calculated by dividing two times the beta of *Time*ΔFC* by the residual standard deviation $SD\,(e_{tij})$. Effect sizes of .2, .5 and .8 were considered as small, medium and large effects respectively (Cohen, 1988).

Effects of targeted versus non-targeted dimensions

With Model 5, we analyzed the effect of dimensions that were targeted for improvement during the consultation meetings versus the effect of non-targeted dimensions on each of the seven specific teaching dimensions on each time interval. These additional exploratory analyses were done to link the effects of the feedback-plus-consultation intervention to the specific content of the consultation. In Model 5, professors in the feedback-plus-consultation condition were separated into two groups for each dimension on each time interval based on the consultation reports; a group which targeted the dimension for improvement (*Target*), meaning that they made concrete plans for improvement, and a group that did not target the dimension (*No Target*). *Condition* was therefore recoded into the dummy variables *Control-versus-Feedback-only* (denoted as Δ*F*), *Control_versus_Feedback-plus-Consultation_No Target* (denoted as Δ*FC_NoTarget*) and *Control_versus_Feedback-plus-Consultation_Target* (denoted as Δ*FC_Target*). *Time* was recoded for the specific time interval (in case of time interval T_1T_2; $T_1 = 0$ and $T_2 = 1$, and in case of time interval T_2T_3; $T_2 = 0$ and $T_3 = 1$). We did not have enough data to fit a model with these additional parameters plus the parameters for all possible random effects. We therefore limited the random effects to the professor level

and student level intercept in this model. To compensate, the parameters were tested with a more restricted alpha of .01.

Moderators

Finally, we modeled the moderating effects of two professor-level moderators, *Age*, *Quality of Teaching* and *Class Size*. Let M denote a moderator of interest. First we added the main effect of M (γ_{003}) to Model 4 (Model 6). Next, we added the interaction-effects of $M^*\Delta F$ (γ_{004}), $M^*\Delta FC$ (γ_{005}), M*Time (γ_{103}), $M^*Time^*\Delta F$ (γ_{104}), and $M^*Time^*\Delta FC$ (γ_{105}) to Model 6 (Model 7). The interaction effects $M^*Time^*\Delta F$ and $M^*Time^*\Delta FC$, represent the separate effects of the two interventions for professors with high and low ratings on the specific moderator, compared to the control condition. The significance of the main effects were tested with a deviance test on Model 4 compared to Model 6 on a 5% significance level. The significance of the interaction effects were tested with a deviance test on Model 6 compared to Model 7 on a 5% significance level. Appendix I includes the equations for Model 6 and 7.

Results

Descriptives

Table 5.1 shows mean ratings, the professor-level standard deviation and the student level standard deviation at each measurement occasion in each condition of all dependent variables. On each dependent variable, mean ratings did not differ with respect to *Condition* at baseline (e.g., *Total Instructional Skills* T_1: $\beta = .029$, $SE = .055$, $p = .596$), indicating successful randomization.

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

Effects of the interventions, in the presence of random effects

Table 5.2 shows the estimates and standard errors of the four models on *Total Instructional Skills*. Deviance tests show that Model 3 (with the two Time*Condition interaction effects) fitted the data significantly better than Model 2 (without the interaction effects) ($\chi^2(2) = 76.9$, $p < .001$). However, Model 4 (with an additional random slope at the professor and student level) fitted the data significantly better than Model 3 (without a random slope) ($\chi^2(2) =$

Table 5.1 Mean student ratings and standard deviations of the dependent variables on professor and student level on each measurement occasion in each condition

		T_1					T_2					T_3				
		$M_{Professors}$	$SD_{Professors}$	$SD_{Students}$ (within groups)			$M_{Professors}$	$SD_{Professors}$	$SD_{Students}$ (within groups)			$M_{Professors}$	$SD_{Professors}$	$SD_{Students}$ (within groups)		
				Min	*Median*	*Max*			*Min*	*Median*	*Max*			*Min*	*Median*	*Max*
Control condition																
Professor level variables	Structure	5.17	0.34	0.64	0.87	1.20	5.16	0.41	0.39	0.84	1.27	5.13	0.39	0.47	0.87	1.22
	Explication	5.36	0.39	0.69	0.86	1.12	5.29	0.46	0.50	0.84	1.15	5.21	0.41	0.77	0.93	1.17
	Stimulation	4.52	0.70	0.83	1.01	1.33	4.56	0.79	0.76	0.99	1.44	4.57	0.74	0.77	1.04	1.54
	Validation	4.80	0.38	0.47	0.88	1.16	4.91	0.38	0.65	0.89	1.07	4.87	0.34	0.61	0.88	1.11
	Instruction	4.72	0.29	0.72	0.87	1.23	4.72	0.31	0.49	0.85	1.14	4.65	0.30	0.69	0.89	1.53
	Comprehension	4.90	0.51	0.65	0.80	1.08	4.89	0.54	0.47	0.81	1.07	4.86	0.56	0.55	0.89	1.16
	Activation	4.62	0.53	0.67	0.90	1.08	4.75	0.60	0.52	0.91	1.13	4.70	0.58	0.67	0.89	1.24
	Total ISQ	4.87	0.31	0.54	0.64	0.81	4.90	0.39	0.45	0.67	0.86	4.86	0.34	0.49	0.67	1.01
Student level variables	Cognition	4.95	0.40	0.90	1.16	1.47	5.04	0.49	0.85	1.10	1.36	4.99	0.43	0.82	1.21	1.48
	Affection	4.26	0.45	1.12	1.39	1.61	4.34	0.52	1.07	1.32	1.68	4.34	0.54	0.71	1.38	1.64
	Regulation	4.79	0.41	0.88	1.33	1.69	4.96	0.34	0.90	1.16	1.74	4.94	0.38	0.80	1.27	1.59
Feedback-only condition																
Professor level variables	Structure	5.21	0.43	0.59	0.88	1.23	5.20	0.42	0.55	0.87	1.14	5.12	0.50	0.72	0.89	1.18
	Explication	5.43	0.52	0.51	0.84	1.23	5.40	0.48	0.65	0.85	1.18	5.32	0.52	0.49	0.85	1.47
	Stimulation	4.83	0.84	0.54	1.01	1.37	4.87	0.88	0.55	0.96	1.23	4.89	0.87	0.50	0.98	1.19
	Validation	4.97	0.35	0.53	0.88	1.27	4.97	0.45	0.61	0.81	1.09	4.95	0.43	0.58	0.89	1.22
	Instruction	4.73	0.33	0.61	0.95	1.24	4.72	0.47	0.68	0.88	1.23	4.73	0.43	0.72	0.89	1.21
	Comprehension	5.03	0.53	0.46	0.83	1.05	4.95	0.51	0.62	0.81	1.06	4.98	0.51	0.54	0.85	1.11
	Activation	4.93	0.67	0.46	0.94	1.20	4.89	0.72	0.56	0.91	1.33	4.98	0.66	0.59	0.85	1.11
	Total ISQ	5.02	0.42	0.30	0.64	0.92	5.00	0.48	0.51	0.62	0.88	4.99	0.48	0.46	0.68	1.10

Table 5.1 *Continued*

		T_1					T_2					T_3				
		$M_{Professors}$	$SD_{Professors}$	$SD_{Students}$ (within groups)			$M_{Professors}$	$SD_{Professors}$	$SD_{Students}$ (within groups)			$M_{Professors}$	$SD_{Professors}$	$SD_{Students}$ (within groups)		
				Min	*Median*	*Max*			*Min*	*Median*	*Max*			*Min*	*Median*	*Max*
Feedback-only condition																
Student level variables	Cognition	5.08	0.53	0.73	1.10	1.35	5.08	0.70	0.67	1.11	1.47	5.07	0.71	0.67	1.12	1.46
	Affection	4.44	0.69	0.85	1.30	1.58	4.40	0.80	0.76	1.31	1.51	4.43	0.77	0.86	1.33	1.56
	Regulation	4.82	0.32	0.90	1.25	1.61	4.81	0.49	0.91	1.28	1.61	4.87	0.40	0.99	1.33	1.61
Feedback-plus-consultation condition																
Professor level variables	Structure	5.19	0.44	0.61	0.83	1.30	5.26	0.44	0.54	0.84	1.26	5.31	0.42	0.58	0.91	1.07
	Explication	5.33	0.54	0.58	0.82	1.17	5.30	0.58	0.60	0.83	1.22	5.34	0.54	0.51	0.83	1.19
	Stimulation	4.62	0.75	0.68	1.03	1.29	4.72	0.79	0.64	0.98	1.22	4.77	0.73	0.72	0.98	1.28
	Validation	4.73	0.40	0.65	0.89	1.14	4.89	0.44	0.69	0.86	1.18	4.99	0.41	0.64	0.89	1.35
	Instruction	4.68	0.35	0.70	0.87	1.14	4.76	0.37	0.55	0.87	1.18	4.79	0.33	0.71	0.91	1.38
	Comprehension	5.11	0.61	0.58	0.76	1.12	5.20	0.51	0.60	0.76	0.96	5.23	0.53	0.57	0.75	1.02
	Activation	4.87	0.86	0.47	0.83	1.28	5.00	0.71	0.53	0.93	1.07	5.12	0.69	0.60	0.84	1.12
	Total ISQ	4.93	0.42	0.47	0.57	0.85	5.02	0.45	0.44	0.60	0.82	5.08	0.44	0.50	0.66	0.86
Student level variables	Cognition	4.96	0.48	0.82	1.10	1.52	5.09	0.48	0.80	1.03	1.41	5.22	0.45	0.88	1.08	1.84
	Affection	4.38	0.49	1.06	1.35	1.55	4.46	0.50	0.96	1.30	1.73	4.50	0.52	0.90	1.33	1.77
	Regulation	4.79	0.38	0.91	1.25	1.53	4.92	0.30	0.98	1.19	1.63	5.04	0.32	0.71	1.20	1.80

Table 5.2 Estimates and standard errors of the four models for *Total Instructional Skills*

Total Instructional Skills	Model 1			Model 2			Model 3			Model 4		
	Estimate		*SE*	Estimate		*SE*	Estimate		*SE*	Estimate		*SE*
Fixed Part												
Intercept	4.942	***	0.046	4.845		0.079	4.887	***	0.079	4.856	***	0.077
Time				0.020	**	0.006	-0.033	**	0.010	-0.008		0.025
ΔFeedback-only				0.119		0.112	0.085		0.113	0.116		0.111
ΔFeedback-plus-Consultation				0.124		0.110	0.024		0.111	0.060		0.109
Time*ΔFeedback-only							0.043	**	0.014	0.011		0.037
Time*ΔFeedback-plus-Consultation							0.126	***	0.014	0.084	*	0.036
Random Part												
Intercept												
Level 3: Professor	0.153	***	0.026	0.149	***	0.025	0.149	***	0.025	0.143	***	0.025
Level 2: Student	0.195	***	0.006	0.195	***	0.006	0.196	***	0.006	0.203	***	0.006
Level 1: Time	0.237	***	0.005	0.237	***	0.005	0.234	***	0.005	0.218	***	0.004
Slope												
Level 3: Professor										-0.004		0.006
Level 2: Student										0.000		0.000
Intercept-slope covariance												
Level 3: Professor										0.013	***	0.003
Level 2: Student										0.000		0.000
-2 log likelihood	27686.9			27674.1			27597.2			27265.0		

Notes: * $p < .05$, ** $p < .01$, *** $p < .001$, units level 3: 75 professors, units level 2: 9,616 students, units level 1: 14,298 observations, variances of .000 are smaller than .0001.

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

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

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

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

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

4 Tables of detailed results of all models on all dependent variables are available on request.

Table 5.3 Estimates and standard errors of Model 4 for each specific dependent variable

	Students' perceptions of learning outcomes									ISQ teaching dimensions 1-2					
	Cognition			Affection			Regulation			Structure			Explication		
	Estimate		*SE*	Estimate		*SE*	Estimate		*SE*	Estimate		*SE*	Estimate		*SE*
Fixed Part															
Intercept	4.974	***	0.092	4.266		0.106	4.809	***	0.072	5.174	***	0.077	5.367	***	0.095
Time	0.000		0.037	0.030		0.038	0.084	*	0.039	-0.024		0.030	-0.083	*	0.035
ΔFeedback-only	0.087		0.132	0.147		0.152	-0.047		0.103	0.008		0.111	0.056		0.136
ΔFeedback-plus-Consultation	-0.022		0.129	0.087		0.149	-0.028		0.101	0.004		0.109	-0.047		0.134
Time*ΔFeedback-only	-0.013		0.054	-0.030		0.056	-0.048		0.056	0.004		0.043	0.038		0.050
Time*ΔFeedback-plus-Consultation	0.121	*	0.053	0.050		0.055	0.049		0.056	0.092	*	0.043	0.078		0.049
Random Part															
Intercept															
Level 3: Professor	0.191	***	0.035	0.253	***	0.046	0.104	***	0.021	0.138	***	0.025	0.215	***	0.037
Level 2: Student	0.649	***	0.031	0.509	***	0.026	0.450	***	0.024	0.279	***	0.011	0.287	***	0.011
Level 1: Time	0.676	***	0.020	1.304	***	0.026	1.236	***	0.035	0.503	***	0.010	0.446	***	0.009
Slope															
Level 3: Professor	0.024	***	0.006	0.021	***	0.006	0.023	***	0.006	0.016	***	0.004	0.024	***	0.005
Level 2: Student	0.041	*	0.016	0.000		0.000	0.021		0.026	0.000		0.000	0.000		0.000
Intercept-slope covariance															
Level 3: Professor	0.003		0.010	0.006		0.012	-0.017		0.009	-0.008		0.007	-0.020	*	0.010
Level 2: Student	-0.036		0.018	0.000		0.000	-0.044		0.028	0.000		0.000	0.000		0.000
Units level 3:	75			75			75			75			75		
Units level 2:	9235			9490			9469			9616			9616		
Units level 1:	13771			14061			14043			14298			14298		

Table 5.3 *Continued*

	ISQ teaching dimensions 3-7														
	Stimulation			Validation			Instruction			Comprehension			Activation		
	Estimate		*SE*	Estimate		*SE*	Estimate		*SE*	Estimate		*SE*	Estimate		*SE*
Fixed Part															
Intercept	4.531	***	0.150	4.819	***	0.074	4.725	***	0.065	4.848	***	0.110	4.600	***	0.140
Time	0.005		0.042	0.026		0.032	-0.040		0.032	0.006		0.033	0.069		0.044
ΔFeedback-only	0.290		0.214	0.128		0.106	-0.050		0.093	0.122		0.157	0.256		0.200
ΔFeedback-plus-Consultation	0.091		0.210	-0.091		0.104	-0.058		0.092	0.259		0.154	0.261		0.197
Time*ΔFeedback-only	0.024		0.060	-0.028		0.046	0.054		0.047	-0.008		0.048	-0.019		0.063
Time*ΔFeedback-plus-Consultation	0.060		0.060	0.105	*	0.046	0.114	*	0.046	0.064		0.047	0.071		0.062
Random Part															
Intercept															
Level 3: Professor	0.542	***	0.091	0.125	***	0.023	0.094	***	0.017	0.291	***	0.049	0.477	***	0.080
Level 2: Student	0.425	***	0.015	0.315	***	0.019	0.335	***	0.020	0.261	***	0.010	0.227	***	0.012
Level 1: Time	0.613	***	0.012	0.468	***	0.014	0.477	***	0.014	0.418	***	0.008	0.641	***	0.012
Slope															
Level 3: Professor	0.035	***	0.007	0.019	***	0.004	0.019	***	0.004	0.022	***	0.005	0.040	***	0.008
Level 2: Student	0.000		0.000	0.024	*	0.011	0.040	***	0.011	0.000		0.000	0.000		0.000
Intercept-slope covariance															
Level 3: Professor	-0.019		0.019	-0.011		0.007	-0.011		0.006	-0.033	**	0.012	-0.066	***	0.020
Level 2: Student	0.000		0.000	-0.012		0.012	-0.019		0.012	0.000		0.000	0.000		0.000
Units level 3:	75			75			75			75			75		
Units level 2:	9616			9616			9616			9616			9616		
Units level 1:	14298			14298			14298			14298			14298		

Notes: * $p < .05$, ** $p < .01$, *** $p < .001$, variances of .000 are smaller than .0001.

Table 5.4 Effect sizes calculated with Cohen's *d* and calculated with multilevel regression Model 4 output

	Control		Feedback-only			Feedback-plus-Consultation		
Comparison:	**(T_3 vs T_1)**	**(T_3 vs T_1)**	**(T_3 vs T_1)**	**(T_3 vs T_1) vs Control (T_3 vs T_1)**		**(T_3 vs T_1)**	**(T_3 vs T_1) vs Control (T_3 vs T_1)**	
Effectsize:	Cohen's *d*	multilevel model	Cohen's *d*	Cohen's *d*	multilevel model	Cohen's *d*	Cohen's *d*	multilevel model
Calculation:	$M(T_{3\_C})$- $M(T_{1\_C})$	2*Time	$M(T_{3\_FB})$- $M(T_{1\_FB})$	$(M(T_{3\_FB})$-$M(T_{1\_FB}))$- $(M(T_{3\_C})$- $M(T_{1\_C}))$	2*Time*ΔFB	$M(T_{3\_FC})$- $M(T_{1\_FC})$	$(M(T_{3\_FC})$- $M(T_{1\_FC}))$- $(M(T_{3\_C})$- $M(T_{1\_C}))$	2*Time*ΔFC
	pooled *SD*	*SD* (e)	pooled *SD*	pooled *SD*	*SD* (e)	pooled *SD*	pooled *SD*	*SD* (e)
Professor level								
Structure	-0.12	-0.07	-0.19	-0.11	0.01	0.29	0.43	0.26
Explication	-0.38	-0.25	-0.22	0.08	0.11	0.02	0.34	0.23
Stimulation	0.07	0.01	0.06	0.00	0.06	0.20	0.13	0.15
Validation	0.19	0.08	-0.06	-0.25	-0.08	0.63	0.49	0.31
Instruction	-0.26	-0.12	0.00	0.22	0.16	0.34	0.60	0.33
Comprehension	-0.07	0.02	-0.09	-0.02	-0.03	0.22	0.30	0.20
Activation	0.15	0.17	0.07	-0.06	-0.05	0.33	0.25	0.18
Total ISQ	-0.05	-0.03	-0.06	-0.03	0.05	0.35	0.43	0.36
Student level								
Cognition	0.09	0.00	-0.02	-0.09	-0.03	0.56	0.50	0.29
Affection	0.15	0.05	-0.02	-0.14	-0.05	0.24	0.09	0.09
Regulation	0.38	0.15	0.15	-0.25	-0.09	0.71	0.26	0.09

Notes: the control condition is denoted C, the feedback-only condition is denoted FB, the feedback-plus-consultation condition is denoted FC. Pooled *SD* for the comparison T_3 vs T_1 of the control condition was calculated with $\sqrt{((SD(T_{1\_control})^2 + SD(T_{4\_control})^2) / 2)}$. Pooled *SD* for the comparison T_3 vs T_1 of the experimental conditions versus the control condition was calculated with $\sqrt{((SD(T_{1\_control})^2 + SD(T_{4\_control})^2 + SD(T_{1\_experimental})^2 + SD(T_{4\_experimental})^2) / 4)}$.

significant on all seven professor level dependent variables and on two out of three student level variables (*Cognition* and *Affection*). The inclusion of a random slope on the professor and student level in Model 4 rendered the effects of both interventions on multiple variables statistically insignificant. This indicates the importance of taking random variations in ratings over time into account when analyzing student ratings data.

Effects of targeted versus non-targeted dimensions

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

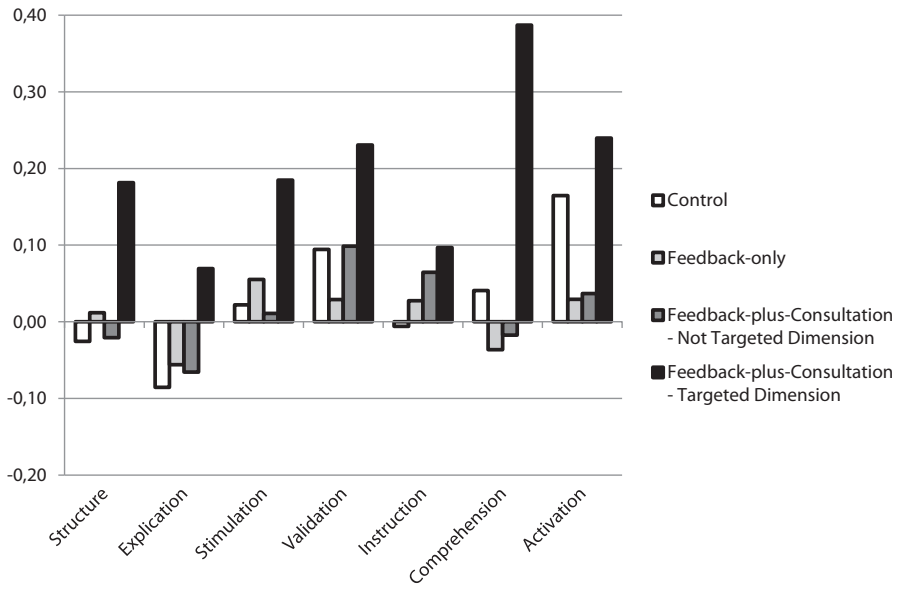

Figure 5.1 Improvement in mean ratings on the first time interval (T_1T_2) for each condition on each ISQ teaching dimension.

We note that professors who did not target the dimension *Comprehension* had ratings on T_1 that were significantly higher than the control condition (Δ*Feedback-plus-Consultation_No-Target*; $\beta = .401$, $SE = .160$, $p = .012$). On the other six dimensions, there were no significant differences between the conditions on their baseline ratings.

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

On the second time interval (T_2T_3) there were fewer effects of the experimental conditions. Professors who targeted the dimensions *Instruction* and *Activation* increased their ratings significantly compared to the control condition (*Instruction*: $\beta = .173$, $SE = .050$, $p < .001$, *Activation*: $\beta = .255$, $SE = .061$, $p < .001$). At the same time, professors who did not target the dimension *Validation* increased their ratings significantly over time compared to the control condition (*Validation*: $\beta = .154$, $SE = .046$, $p < .001$). Figure 5.2 shows the mean improvement for each condition on each dimension on the second time interval. We note that

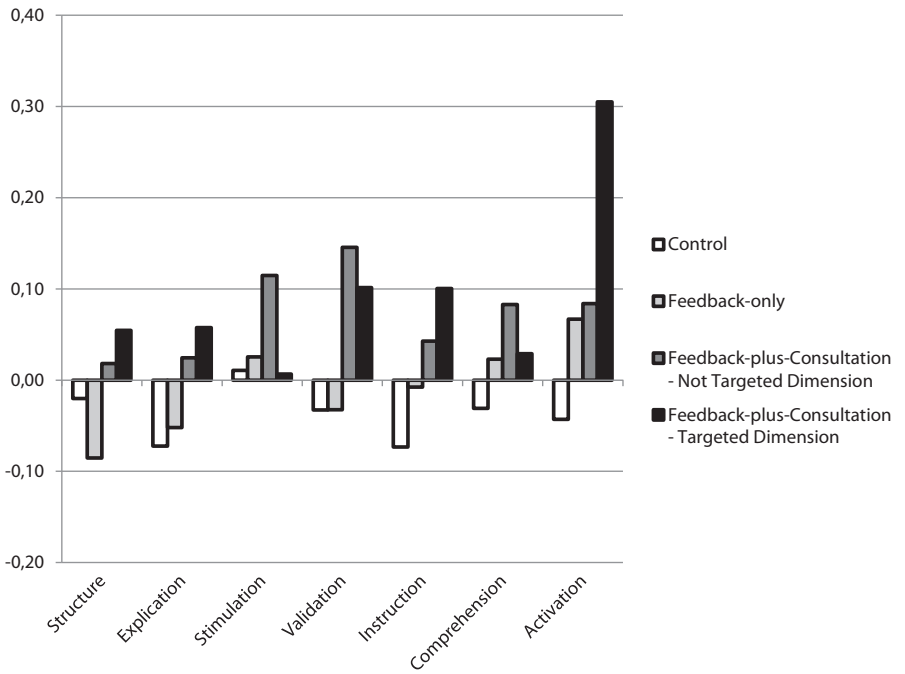

Figure 5.2 Improvement in mean ratings on the second time interval (T_2T_3) for each condition on each ISQ teaching dimension.

on this time interval, ratings of the control condition on *Explication* decreased significantly again (*Time*; β = -.075, *SE* =. 028, p = .007).

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

Moderators Age, Quality of Teaching and Class Size

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

Deviance tests between Model 4 and Model 6 indicated that there was a main effect of *Age* on five out of seven teaching dimensions (i.e., *Total Instructional Skills*, *Structure*, *Explication*, *Stimulation*, *Validation* and *Instruction*), and on the student learning variable *Cognition* (e.g., *Total Instructional Skills*: ($\chi^2(1) = 8.495$, p = .004). On all of these variables younger professors received higher ratings compared to older colleagues on a 5% level (e.g., *Total Instructional Skills*: β = -.014, *SE* = .005, p = .005).

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

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

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

Table 5.5 Estimates and standard errors of Model 7 with moderators *Age*, *Quality of Teaching* and *Class Size* for Total Instructional Skills

Total Instructional Skills	Model Moderator Age			Model Moderator Quality of Teaching			Model Moderator Class Size		
	Estimate		*SE*	Estimate		*SE*	Estimate		*SE*
Fixed Part									
Intercept	4.849		0.073	4.825		0.088	4.890		0.079
Time	-0.010		0.025	0.059		0.032	-0.015		0.026
ΔFeedback-only	0.200		0.109	-0.048		0.127	0.120		0.112
ΔFeedback-plus-Consultation	0.054		0.104	-0.149		0.124	0.031		0.111
Time*ΔFeedback-only	-0.008		0.037	-0.041		0.046	0.025		0.037
Time*ΔFeedback-plus-Consultation	0.087	*	0.036	0.019		0.045	0.098	**	0.036
Moderator	-0.014	*	0.007	0.088		0.132	-0.001		0.001
Moderator*Time	-0.002		0.002	-0.147	**	0.048	0.000		0.000
Moderator*ΔFeedback-only	-0.005		0.011	0.355		0.189	-0.001		0.002
Moderator*ΔFeedback-plus-Consultation	0.005		0.011	0.449	*	0.185	0.000		0.002
Moderator*ΔFeedback-only*Time	0.007		0.004	0.115		0.069	0.000		0.001
Moderator*ΔFeedback-plus-onsultation*Time	0.002		0.004	0.145	*	0.068	0.001		0.001

Table 5.5 *Continued*

Total Instructional Skills	Model Moderator Age			Model Moderator Quality of Teaching			Model Moderator Class Size		
	Estimate		*SE*	Estimate		*SE*	Estimate		*SE*
Random Part									
Intercept									
Level 3: Professor	0.126	***	0.022	0.101	***	0.018	0.137	***	0.024
Level 2: Student	0.203	***	0.006	0.203	***	0.006	0.207	***	0.006
Level 1: Time	0.218	***	0.004	0.218	***	0.004	0.217	***	0.004
Slope									
Level 3: Professor	0.012	***	0.003	0.011	***	0.002	0.012	***	0.003
Level 2: Student	0.000		0.000	0.000		0.000	0.000		0.000
Intercept-slope covariance									
Level 3: Professor	-0.002		0.005	-0.001		0.005	-0.004		0.006
Level 2: Student	0.000		0.000	0.000		0.000	0.000		0.000
-2 log likelihood	27252.080			27232.284			26170.058		
Deviance with Model 5 (df = 5)	4.452			16.130	**		2.682		

Notes: * $p < .05$, ** $p < .01$, *** $p < .001$. Age and Quality of Teaching; units level 3: 75 professors, units level 2: 9,616 students, units level 1: 14,298 observations. Class Size; Age and Quality of Teaching; units level 3: 73 professors, units level 2: 9,189 students, units level 1: 13,719 observations. Variances of .000 are smaller than .0001.

When the moderator *Quality of Teaching* was added, deviance tests showed significant effects of the interactions on *Total Instructional Skills* ($\chi^2(5) = 16.13$, $p = .007$), and the teaching dimensions *Explication* ($\chi^2(5) = 12.73$, $p = .026$), *Stimulation* ($\chi^2(5) = 16.09$, $p = .007$) and *Validation* ($\chi^2(5) = 19.81$, $p = .001$). Thus, *Quality of Teaching* moderated the treatment effects on these variables. On these four variables, medium quality professors in the experimental conditions did not improve their ratings compared to the control condition at all. The effects occurred mainly on high quality professors.

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

Discussion

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

5 Tables of detailed results of Model 6 and 7 on all dependent variables are available on request.

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

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

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

Exploratory analyses with the moderators Age and Class Size had no influence on the effects of the interventions on any of the dimensions. The effects are the same for younger as for older professors and for professors with smaller and larger classes. Exploratory analyses with the moderator *Quality of Teaching* showed that effects of the interventions partly depend on the professor's baseline quality of teaching. Results were different between medium quality professors and high quality professors on four dependent variables; *Explication*, *Stimulation*, *Validation*, and *Total Instructional Skills*. On these variables ratings of medium quality professors did not change over time in all three conditions. Ratings of high quality professors improved significantly in the feedback-plus-consultation condition on *Validation* and *Total Instructional Skills*, compared to the control condition. Ratings of high quality

professors in the feedback-only condition significantly increased in ratings on *Stimulation*. At the same time, high quality professors in the control condition decrease in ratings on *Explication*, *Stimulation*, *Validation* and on *Total Instructional Skills*. Thus, the effects of the interventions on high quality professors may be viewed as helping them to increase ratings as well as to prevent a decrease in ratings over time. It should be noted that baseline ratings of high quality professors were higher in the feedback-plus-consultation condition compared to the control condition. Effects on high quality professors might therefore be somewhat biased.

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

Professors in this study were members of the current staff of different departments and had not sought out our intervention, or any other, aimed at improving teaching skills. The present effects are therefore not dependent on the professors' intrinsic motivation. The randomized block design furthermore ensured an equal distribution of professors from different departments and high versus medium teaching quality to each condition. The results therefore generalize to professors at this university in general.

The present study has its limitations. We note that the effects are investigated by means of students' perceptions of teaching. Although SETs are proven to be valid and reliable in many different settings (see Marsh, 2007), other researchers on evaluation of teaching effectiveness have recommended the use of multiple sources of data to assess teaching quality (Benton & Cashin, 2012). Also, with respect to student learning outcomes, the results were based on students self-reports only. We therefore suggest future research to complement these findings from student ratings with additional measures of teaching effectiveness, such as classroom observations, and of student learning outcomes, such as course grades.

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

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

References

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

Brinko, K.T. (1990). Instructional consultation with feedback in higher education. *Journal of Higher Education, 61,* 65-83.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.)*. New Jersey: Lawrence Erlbaum.

Dresel, M., & Rindermann, H. (2011). Counseling university instructors based on student evaluations of their teaching effectiveness: a multilevel test of its effectiveness under consideration of bias and unfairness variables. *Research in Higher Education, 52,* 717-737.

De Neve, H.M.F., & Janssen, P.J. (1982). Validity of student evaluation of instruction. *Higher Education, 11,* 543-552.

Frey, P.W., Leonard, D.W., & Beatty, W.W. (1975). Student ratings of instruction - validation research. *American Educational Research Journal, 12,* 435-444.

Hattie, J.A.C., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*, 81-112.

Hox, J.J. (2002). *Multilevel Analysis. Techniques and Applications.* Mahwah, NJ: Lawrence Erlbaum Associates.

Janssen, P. J., & De Neve, H.M.F. (1988). *Studeren en doceren aan het hoger onderwijs: Vakmanschap als leeropdracht (Studying and Lecturing in Higher Education)*. Leuven: Acco.

Kember, D., Leung, D.Y.P., & Kwan, K.P. (2002). Does the use of student feedback questionnaires improve the overall quality of teaching? *Assessment & Evaluation in Higher Education, 27,* 411-425.

Levinson-Rose, J., & Menges, R.J. (1981). Improving college teaching: A critical review of research. *Review of Educational Research, 51,* 403-434.

Marsh, H.W. (1984). Students evaluations of university teaching - dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology, 76,* 707-754.

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

Marsh, H.W. (2007b). Students' Evaluations of University Teaching: Dimensionality, Reliability, Validity, Potential Biases and Usefulness. In Perry, R.P. & Smart, J.C. (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective*, 319-384. New York, NY: Springer.

Marsh, H.W., & Hocevar, D. (1991b). Students' evaluations of teaching effectiveness: The stability of mean ratings of the same teachers over a 13-year period. *Teaching and Teacher Education, 7,* 303-314.

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

McAlpine, L., & Weston, C. (2000). Reflection: Issues related to improving instructors' teaching and students' learning. *Instructional Science, 28,* 363-385.

McLaughlin, M.W., & Pfeifer, R.S. (1988). *Teacher Evaluation: Improvement, Accountability, and Effective Learning.* New York, NY: Teachers College Press.

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

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

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

Pinheiro, J., Bates, D., DebRoy, S. Sarkar, & the R Development Core Team (2012). *Nlme: Linear and Nonlinear Mixed Effects Models.* R package version 3.1-106.

Prebble, T., Hargraves, H., Leach, L., Naidoo, K., Suddaby, G., & Zepke, N. (2004). *Impact of student support services and academic development programmes on student outcomes in undergraduate tertiary study: A synthesis of the research.* Report to the Ministry of Education, Massey University College of Education.

R Development Core Team (2008). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. Http://www.R-project.org.

Rasbash, J., Charlton, C., Browne, W.J., Healy, M., & Cameron, B. (2009) *MLwiN Version 2.1.* Centre for Multilevel Modelling, University of Bristol.

Richardson, T.T.E. (2005). Instruments for obtaining student feedback: a review of the literature. *Assessment and Evaluation in Higher Education, 30,* 378-415.

Rindermann, H., Kohler, J., & Meisenberg, G. (2007). Quality of instruction improved by evaluation and consultation of instructors. *International Journal for Academic Development, 12,* 73-85.

SPSS Inc. (Released 2007). *SPSS for Windows, Version 16.0.* Chicago: SPSS Inc.

Stes, A., Min-Leliveld, M., Gijbels, D., & Van Petegem, P. (2010). The impact of instructional development in higher education: The state-of-the-art of the research. *Educational Research Review, 5*, 25-49.

Theall, M., & Franklin, J. (2001). Looking for Bias in All the Wrong Places: A Search for Truth or a Witch Hunt in Student Ratings of Instruction?. In Theall, M. P. Abrami, & Mets, L. (Eds.), *The Student Ratings Debate: Are they Valid? How Can We Best use Them? New Directions for Institutional Research, 109,* 45-56. San Francisco, CA: Jossey-Bass.

Verbeek, F., De Jong, U., & Vermeulen, A. (2002). *Rapportage Uvalon.* Amsterdam: SCO Kohnstamn Institute, University of Amsterdam.

Verbeek, F., De Jong, U., & Vermeulen, A. 2005. *Jaarverslag Uvalon 2003 en 2004.*Amsterdam: SCO Kohnstamn Institute, University of Amsterdam.

Vermunt, J.D., & Verschaffel, L. (2000). Process-oriented teaching. In R.J. Simons, J. van der Linden, & T. Duffy (Eds.), *New Learning*, 209-225. Dordrecht: Kluwer Academic.

Vorst, H.C.M., & Van Engelenburg, B. (1992). *UVALON UvA-pakket voor onderwijsevaluatie.* Amsterdam: Psychological Methods Department, University of Amsterdam.

Weimer, M., & Lenze, L.F. (1997). Instructional interventions: A review of the literature on efforts to improve instruction. In K. R. Perry, & J. C. Smart (Eds.), *Effective teaching in higher education: Research and practice,* 205-240. New York, NY: Agathon Press.

Appendix I

Multilevel regression models 1 to 7

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

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

$$\text{Level 1:} \quad Y_{tij} = \beta_{0ij} + e_{tij}\,, \tag{1.1}$$

$$\text{Level 2:} \quad \beta_{0ij} = \beta_{00j} + u_{0ij}\,, \tag{1.2}$$

$$\text{Level 3:} \quad \beta_{00j} = \gamma_{000} + v_{00j}\,. \tag{1.3}$$

By substitution, we obtain the single-equation:

$$Y_{tij} = \gamma_{000} + v_{00j} + u_{0ij} + e_{tij}\,. \tag{1.4}$$

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

The variances of the three residual error terms are denoted by

$$\text{var}\,(e_{tij}) = \sigma^2\,, \quad \text{var}\,(u_{0ij}) = \tau_0{}^2\,, \quad \text{var}(v_{00j}) = \varphi_0{}^2\,. \tag{1.5}$$

These represent the variance of the ratings over time of a given student *i* (σ^2), the variance of the ratings over students of a given professor *j* ($\tau_0{}^2$), and the variance of the ratings over professors ($\varphi_0{}^2$).

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

$$\text{Level 1:} \quad Y_{tij} = \beta_{0ij} + \beta_{1ij}\,\text{Time}_{ij} + e_{tij}\,, \tag{2.1}$$

$$\text{Level 2:} \quad \beta_{0ij} = \beta_{00j} + u_{0ij}\,, \tag{2.2}$$

$$\text{Level 2:} \quad \beta_{1ij} = \beta_{10j}, \tag{2.3}$$

$$\text{Level 3:} \quad \beta_{00j} = \gamma_{000} + \gamma_{001}{}^{*}\Delta \text{F j} + \gamma_{002}{}^{*}\Delta \text{FC}_{j} + \nu_{00j}, \tag{2.4}$$

$$\text{Level 3:} \quad \beta_{10j} = \gamma_{100}. \tag{2.5}$$

By substitution, we obtain the single-equation:

$$Y_{tij} = \gamma_{000} + \gamma_{100}\,\text{Time}_{ij} + \gamma_{001}\Delta \text{F}_{j} + \gamma_{002}\Delta \text{FC}_{j} + \nu_{00j} + u_{0ij} + e_{tij}. \tag{2.6}$$

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

Model 3. In Model 3 we added the Time*Condition interaction effects *Time**Δ*F* and *Time**Δ*FC*. Model 3 is defined in equations:

$$\text{Level 1:} \quad Y_{tij} = \beta_{0ij} + \beta_{1ij}\,\text{Time}_{ij} + e_{tij}, \tag{3.1}$$

$$\text{Level 2:} \quad \beta_{0ij} = \beta_{00j} + u_{0ij}, \tag{3.2}$$

$$\text{Level 2:} \quad \beta_{1ij} = \beta_{10j}, \tag{3.3}$$

$$\text{Level 3:} \quad \beta_{00j} = \gamma_{000} + \gamma_{001}\Delta \text{F}_{j} + \gamma_{002}\Delta \text{FC}_{j} + \nu_{00j}, \tag{3.4}$$

$$\text{Level 3:} \quad \beta_{10j} = \gamma_{100} + \gamma_{101}\Delta \text{F}_{j} + \gamma_{102}\Delta \text{FC}_{j}. \tag{3.5}$$

By substitution, we obtain the single-equation:

$$Y_{tij} = \gamma_{000} + \gamma_{100}\,\text{Time}_{ij} + \gamma_{001}\Delta \text{F}_{j} + \gamma_{002}\Delta \text{FC}_{j} + \gamma_{101}\text{Time}_{ij}{}^{*}\Delta \text{F}_{j} + \gamma_{102}\text{Time}_{ij}{}^{*}\Delta \text{FC}_{j} + \nu_{00j} + u_{0ij} + e_{tij}. \tag{3.6}$$

The parameters are the same as in Model 2. The parameters of *Time**Δ*F* (γ_{101}) and *Time**Δ*FC* (γ_{102}) represent the effects of the interventions.

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

$$\text{Level 1:} \quad Y_{tij} = \beta_{0ij} + \beta_{1ij}\,\text{Time}_{ij} + e_{tij}, \tag{4.1}$$

$$\text{Level 2:} \quad \beta_{0ij} = \beta_{00j} + u_{0ij}, \tag{4.2}$$

$$\text{Level 2:} \quad \beta_{1ij} = \beta_{10j} + u_{1ij}\,, \tag{4.3}$$

$$\text{Level 3:} \quad \beta_{00j} = \gamma_{000} + \gamma_{001}\Delta\text{F}_j + \gamma_{002}\Delta\text{FC}_j + v_{00j}\,, \tag{4.4}$$

$$\text{Level 3:} \quad \beta_{10j} = \gamma_{100} + \gamma_{101}\Delta\text{F}_j + \gamma_{102}\Delta\text{FC}_j + v_{10j}\,. \tag{4.5}$$

By substitution, we obtain the single-equation:

$$\begin{aligned} Y_{tij} = {} & \gamma_{000} + \gamma_{100}\,\text{Time}_{ij} + \gamma_{001}\Delta\text{F}_j + \gamma_{002}\Delta\text{FC}_j + \\ & \gamma_{101}\text{Time}_{ij}{}^{*}\Delta\text{F}_j + \gamma_{102}\text{Time}_{ij}{}^{*}\Delta\text{FC}_j + v_{10j} + u_{1ij} + v_{00j} + u_{0ij} + e_{tij}\,. \end{aligned} \tag{4.6}$$

Again the intercept β_{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 β_{1ij} for *Time* is allowed to be random over students and professors by including the random effects u_{1ij} and v_{10j}.

The slope variances are denoted by

$$\text{var}\,(u_{1ij}) = \tau_1{}^2\,, \quad \text{var}\,(v_{10j}) = \varphi_1{}^2\,. \tag{4.7}$$

The intercept-slope covariances are denoted by

$$\text{cov}\,(u_{0ij}\,, u_{1ij}) = \tau_{01}\,, \quad \text{cov}\,(v_{00j}\,, v_{10j}) = \varphi_{01}\,. \tag{4.8}$$

Model 5. In Model 5, for each dimension on each specific time interval, we split up the feedback-plus consultation condition in a group that targeted the dimension for improvement (*Target*) and a group that did not (*No Target*). *Condition* was therefore recoded into the dummy variables Control_versus_Feedback-only (denoted as Δ*F*), Control_versus_Feedback-plus-Consultation_No Target (denoted as Δ*FC_NoTarget*) and Control_versus_Feedback-plus-Consultation_Target (denoted as Δ*FC_Target*). Time was recoded for the specific time interval (in case of time interval T_1T_2; $\text{T}_1 = 0$ and $\text{T}_2 = 1$ and in case of time interval T_2T_3; $\text{T}_2 = 0$ and $\text{T}_3 = 1$). The random effects were limited to the intercept in this model. Model 5 is defined through the equations:

$$\text{Level 1:} \quad Y_{tij} = \beta_{0ij} + \beta_{1ij}\,\text{Time}_{ij} + e_{tij}\,, \tag{5.1}$$

$$\text{Level 2:} \quad \beta_{0ij} = \beta_{00j} + u_{0ij}\,, \tag{5.2}$$

$$\text{Level 2:} \quad \beta_{1ij} = \beta_{10j}\,, \tag{5.3}$$

$$\begin{aligned} \text{Level 3:} \quad \beta_{00j} = {} & \gamma_{000} + \gamma_{001}\Delta\text{F}_j + \gamma_{002}\Delta\text{FC\_NoTarget}_j + \\ & \gamma_{003}\Delta\text{FC\_Target}_j + v_{00j} \end{aligned} \tag{5.4}$$

$$\begin{aligned} \text{Level 3:} \quad \beta_{10j} = {} & \gamma_{100} + \gamma_{101}\Delta\text{F}_j + \gamma_{102}\Delta\text{FC\_NoTarget}_j + \\ & \gamma_{103}\Delta\text{FC\_Target}_j\,. \end{aligned} \tag{5.5}$$

By substitution, we obtain the single-equation:

$$Y_{tij} = \gamma_{000} + \gamma_{100}\,\text{Time}_{ij} + \gamma_{001}\Delta\text{F}_j + \gamma_{002}\Delta\text{FC\_NoTarget}_j + \gamma_{003}\Delta\text{FC\_Target}_j + \gamma_{101}\text{Time}_{ij}{}^{*}\Delta\text{F}_j + \gamma_{102}\text{Time}_{ij}{}^{*}\Delta\text{FC\_NoTarget}_j + \gamma_{103}\text{Time}_{ij}{}^{*}\Delta\text{FC\_Target}_j + \nu_{00j} + u_{0ij} + e_{tij}\,. \quad (5.6)$$

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

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

$$\beta_{00j} = \gamma_{000} + \gamma_{001}\Delta\text{F}_j + \gamma_{002}\Delta\text{FC}_j + \gamma_{003}\text{M}_j + \nu_{00j}\,, \quad (6.1)$$

$$\beta_{10j} = \gamma_{100} + \gamma_{101}\Delta\text{F}_j + \gamma_{102}\Delta\text{FC}_j + \gamma_{103}\text{M}_j + \nu_{10j}\,. \quad (6.2)$$

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

$$\beta_{00j} = \gamma_{000} + \gamma_{001}\Delta\text{F}_j + \gamma_{002}\Delta\text{FC}_j + \gamma_{003}\text{M}_j + \gamma_{004}\Delta\text{F}_j{}^{*}\text{M}_j + \gamma_{005}\Delta\text{FC}_j{}^{*}\text{M}_j + \nu_{00j}\,, \quad (7.1)$$

$$\beta_{10j} = \gamma_{100} + \gamma_{101}\Delta\text{F}_j + \gamma_{102}\Delta\text{FC}_j + \gamma_{103}\text{M}_j + \gamma_{104}\Delta\text{F}_j{}^{*}\text{M}_j + \gamma_{105}\Delta\text{FC}_j{}^{*}\text{M}_j + \nu_{10j}\,. \quad (7.2)$$

6

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

Arthur, L. (2009). From performativity to professionalism: lecturers' responses to student feedback. *Teaching in Higher Education, 14,* 441-454.

Benton, S.L., & Cashin, W.E. (2012). *Student ratings of teaching: A summary of research and literature (IDEA Paper no. 50).* Manhattan, KS: The IDEA Center.Http://www.theideacenter.org/sites/default/files/idea-paper_50.pdf

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

De Neve, H.M.F., & Janssen, P.J. (1982). Validity of student evaluation of instruction. *Higher Education, 11,* 543-552.

Feldman, K.A. (2007). Identifying exemplary teachers and teaching: Evidence from student ratings. In R. P. Perry & J. C. Smart (Eds.), *The Scholarship of teaching and learning in higher education: An evidence-based perspective,* 93-129. Dordrecht: Springer.

Guskey, T.R. (2000). *Evaluating professional development.* Thousand Oaks, CA: Corwin Press.

Handal, G. (1999). Consultation using critical friends. In Knapper, C. & Piccinin, S. (Eds.), *Using Consultants to Improve Teaching. New Directions for Teaching and Learning, 79,* 59-70. San Francisco, CA: Jossey-Bass.

Levinson-Rose, J., & Menges, R. J. (1981). Improving college teaching: A critical review of research. *Review of Educational Research, 51,* 403-434.

Lenze, L.F. (1996). Instructional development: What works? *National Education Association, Office of Higher Education Update, 2,* 1-4.

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

Marsh, H.W., & Hocevar, D. (1991b). Students' evaluations of teaching effectiveness: The stability of mean ratings of the same teachers over a 13-year period. *Teaching and Teacher Education, 7,* 303-314.

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

McAlpine, L., & Weston, C. (2000). Reflection: Issues related to improving instructors' teaching and students' learning. *Instructional Science, 28,* 363-385.

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

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

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

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

Prebble, T., Hargraves, H., Leach, L., Naidoo, K., Suddaby, G., & Zepke, N. (2004). *Impact of student support services and academic development programmes on student outcomes in undergraduate tertiary study: A synthesis of the research.* Report to the Ministry of Education, Massey University College of Education.

Rotem, A., & Glasman, N.S. (1979). On the effectiveness of students' evaluative feedback to university instructors. *Review of Educational Research, 49,* 497-511.

Schön., D. (1987). Educating the reflective practitioner: Toward a new design for teaching and learning in the professions. San Francisco: Jossey Bass.

Steinert, Y., Mann, K., Centeno, A., Dolmans, D., Spencer, J., Gelula, M., & Prideaux, D. (2006). A systematic review of faculty development initiatives designed to improve teaching effectiveness in medical education: BEME guide no. 8. *Medical Teacher, 28,* 497-526.

Stes, A., Min-Leliveld, M., Gijbels, D., & Van Petegem, P. (2010). The impact of instructional development in higher education: The state-of-the-art of the research. *Educational Research Review, 5*, 25-49.

Theall, M., & Franklin, J. (2001). Looking for Bias in All the Wrong Places: A Search for Truth or a Witch Hunt in Student Ratings of Instruction?. In Theall, M. P. Abrami, & Mets, L. (Eds.), *The Student Ratings Debate: Are they Valid? How Can We Best use Them? New Directions for Institutional Research, 109,* 45-56. San Francisco, CA: Jossey-Bass.

Vorst, H.C.M., & Van Engelenburg, B. (1992). *UVALON UvA-pakket voor onderwijsevaluatie.* Amsterdam: Psychological Methods Department, University of Amsterdam.

Weimer, M., & Lenze, L.F. (1997). Instructional interventions: A review of the literature on efforts to improve instruction. In K. R. Perry, & J. C. Smart (Eds.), *Effective teaching in higher education: Research and practice,* 205-240. New York, NY: Agathon Press.

Nederlandse samenvatting

In deze Nederlandse samenvatting wordt de aanleiding van het huidig onderzoek besproken, gevolgd door de belangrijkste onderzoeksvraag, de opzet van de twee experimenten in dit proefschrift, de voornaamste resultaten en de wetenschappelijke en praktische implicaties daarvan. Voor een uitgebreidere samenvatting, met meer resultaten en discussie, verwijs ik u naar het Engelstalige slothoofdstuk van dit proefschrift.

Aanleiding

Van universitaire docenten (UD's, UHD's en professoren) wordt verwacht dat zij kwalitatief hoogwaardige prestaties leveren, zowel in onderzoek als in onderwijs. Handal (1999) spreekt ook wel van dubbele professionaliteit ('dual professionalism'). Universiteiten dienen daarbij de kwaliteit van onderzoek en onderwijs in toenemende mate te verantwoorden in de vorm van interne en externe kwaliteitzorgsystemen (instellingstoetsen, opleidingsaccreditaties, onderwijscommissies, individuele prestatie monitoring, etc.).

Om te kunnen excelleren in *onderzoek* worden universitaire docenten circa vier jaar opgeleid in de vorm van promotieonderzoek. Tegelijkertijd is de *didactische* scholing van universitaire docenten zeer beperkt. Waar docenten in het primair en voortgezet onderwijs een één- tot vierjarige vooropleiding genieten, leren universitaire docenten het vak voornamelijk in de praktijk, met vallen en opstaan, vaak met weinig middelen en in isolatie.

Als reactie hierop zijn in de jaren '70 en '80 centra opgericht voor training en nascholing van universitaire docenten. In 2008 hebben bovendien de rectores magnifici van alle veertien Nederlandse universiteiten een overeenkomst getekend voor een wederzijdse erkenning van de Basis Kwalificatie Onderwijs (BKO-certificaat). Universitaire docenten dienen sindsdien dit certificaat te behalen bij een vaste aanstelling als UD, UHD of hoogleraar.

Met de opkomst van deze initiatieven, groeit ook het belang van onderzoek naar de effectiviteit van verschillende (na)scholingsinterventies voor universitaire docenten. De kwaliteit en de omvang van dit onderzoek is tot op heden beperkt. In literatuurstudies en metastudies naar de effectiviteit van diverse didactische interventies wordt herhaaldelijk gewezen op de noodzaak van meer onderzoek, en met name meer *experimenteel* onderzoek, in dit veld (zie bijvoorbeeld Stes, Min-Leliveld, Gijbels, & van Petegem, 2010).

Dit proefschrift heeft tot doel een wetenschappelijke bijdrage te leveren aan de kennis over de effectiviteit van didactische interventies voor universitaire docenten. In dit proefschrift is daartoe experimenteel onderzoek uitgevoerd naar twee specifieke interventies;

1. feedback van studenten voor docenten tijdens de cursus, *met* individuele gesprekken met een coach over de studentfeedback en mogelijke verbeteringen in het onderwijs (individuele consultatie), en

2. feedback van studenten voor docenten tijdens de cursus, *zonder* individuele consultatie.

Deze twee interventies zijn toegepast op het hoorcollege onderwijs van universitaire docenten. In de volgende paragraaf wordt nader ingegaan op de keuze voor deze interventies.

Feedback en consultatie

Feedback van studenten wordt volgens eerder onderzoek door zowel docenten als docent-trainers beschouwd als een bruikbaar middel voor het verbeteren van het onderwijs. Tegenwoordig worden wereldwijd vele universitaire cursussen regelmatig geëvalueerd door studenten. Ondanks al deze inspanningen laat eerder onderzoek echter zien dat cursusevaluaties vaak weinig tot geen invloed hebben op de didactische kwaliteiten van de docent. De timing van cursusevaluaties laat niet toe dat docenten al tijdens de cursus hun didactiek kunnen bijsturen en cursusevaluaties komen vaak zonder ondersteunende middelen voor verbetering.

Onderzoeken naar *tussentijdse* evaluaties (tijdens de cursus of het semester) laten wel lichte verbetering zien op cursusevaluaties. De effecten zijn groter bevonden wanneer (tussentijdse) evaluaties worden *gecombineerd met individuele consultatie*. Volgens de literatuur is (collegiale of expert-) consultatie een van de meest gebruikte interventies bij didactische (na)scholing van universitaire docenten, naast formele didactische cursussen en workshops. Echter, de variatie in gevonden effecten van consultatie is groot en ook in dit specifieke onderzoeksgebied dringen auteurs van literatuurstudies en metastudies aan op meer experimenteel onderzoek. In dit proefschrift is daarom gekozen voor grondig onderzoek naar de effecten van regelmatige feedback van studenten in combinatie met een specifieke vorm van consultatie ('collaborative consultation', in Nederland beter bekend als 'coaching') *gedurende* een cursusperiode.

Experimenten in dit proefschrift

In dit proefschrift zijn twee experimenten uitgevoerd. In het eerste experiment zijn vijf-entwintig universitaire psychologiedocenten van de Universiteit van Amsterdam random ingedeeld in een controle conditie en een feedback-plus-consultatie conditie. In het tweede experiment zijn vijfenzeventig universitaire docenten van vijf verschillende faculteiten van de Universiteit van Amsterdam random ingedeeld in drie condities: een controle conditie, een feedback conditie, en een feedback-plus-consultatie conditie. Binnen deze laatste steekproef

bestond een grote variëteit in leeftijd van de docent, functie, didactische kwaliteit, ervaring, cursusniveau en groepsgrootte. De doorsnee cursus van deze docenten had een omvang van acht weken, met tenminste één twee uur durend hoorcollege per week en bijpassende werkgroepen of practica.

In alle condities is een aantal maal aan studenten gevraagd om aan het eind van een hoorcollege een evaluatieformulier (de Instructional Skills Questionnaire, ISQ) in te vullen over de didactische vaardigheden van de docent tijdens het desbetreffende hoorcollege. In het eerste experiment zijn van alle docenten vier hoorcolleges op deze wijze geëvalueerd en in het tweede experiment drie hoorcolleges. Studenten wisten niet dat de docenten ingedeeld waren in verschillende condities.

In de *feedback-plus-consultatie conditie* kregen de docenten binnen enkele dagen na het geëvalueerde hoorcollege een gesprek met een coach, met wie zij de feedback van de studenten doornamen en de mogelijkheden bespraken voor verbeteringen in de daaropvolgende hoorcolleges (vier gesprekken in het eerste experiment en drie gesprekken in het tweede experiment). In de *feedback conditie* ontvingen de docenten de feedback van de studenten steeds binnen enkele dagen na het geëvalueerde hoorcollege per email, zonder consultatiegesprekken. In de *controle conditie* ontvingen de docenten de feedback van de studenten na afloop van de cursus.

Onderzoeksvraag

Door middel van de twee experimenten is getracht de volgende hoofdvraag te beantwoorden:

Wat zijn de effecten van tussentijdse feedback met en zonder consultatie op:

a. de tevredenheid van docenten met de interventies (experiment 2: hoofdstuk 4),
b. leren van docenten met betrekking tot hun hoorcollegeonderwijs, volgens de docenten zelf (experiment 2: hoofdstuk 4),
c. gedrag van docenten tijdens hoorcolleges, volgens de studenten (experiment 1: hoofdstuk 3, experiment 2: hoofdstuk 5), en
d. leren van studenten tijdens hoorcolleges, volgens de studenten (experiment 2: hoofdstuk 5)?

Op basis van de studentdata, verzameld tijdens de hoorcollege-evaluaties, is tevens de psychometrische kwaliteit van het hoorcollege-evaluatie-instrument (de ISQ) onderzocht (hoofdstuk 2). Middels exploratieve analyses is onder andere de samenhang van de effecten met de leeftijd, didactische kwaliteit en groepsgrootte van de docent onderzocht (hoofdstuk 3, 4, en 5).

Resultaten

Psychometrische kwaliteit van de Instructional Skills Questionnaire

In hoofdstuk 2 is de psychometrische kwaliteit onderzocht van de vragenlijst die door studenten tijdens de hoorcolleges is ingevuld (de ISQ). De analyses zijn gebaseerd op 14298 ISQ-formulieren, welke zijn ingevuld door studenten op de drie verschillende meetmomenten (drie hoorcolleges tijdens de cursussen) in het tweede experiment.

De ISQ meet zeven dimensies van de didactische vaardigheid van de universitaire docent tijdens een hoorcollege (*Structure*, *Explication*, *Stimulation*, *Validation*, *Instruction*, *Comprehension* en *Activation*). De resultaten ondersteunen de inhoudsvaliditeit, construct-validiteit, interne structuur en de betrouwbaarheid van de subschalen van de ISQ.

Middels de ISQ is tevens de perceptie van de student gemeten met betrekking tot de leeropbrengst van het hoorcollege voor de student. De scores van de docent op de didactische dimensies voorspellen deze studentperceptie van de leeropbrengst.

Tevredenheid met de interventies en effecten op het leren van de docent (volgens docenten)

Aan alle docenten in het tweede experiment is circa twee weken na hun cursus een uitgebreide vragenlijst toegestuurd. Hiermee is onderzocht of docenten tevreden waren met de interventie die aan hen was toegewezen en wat zij, naar eigen zeggen, hadden geleerd over hun hoorcollegeonderwijs gedurende de cursusperiode. Deze vragenlijst is ingevuld door 70 van de 75 docenten in het tweede experiment.

In alle drie de condities (controle, feedback en feedback-plus-consultatie conditie) waren docenten positief over de hoorcollege-evaluaties. Docenten beschouwden de tijd die de interventies hen kostte als goed besteed en gaven aan dat zij hoorcollege-evaluaties aanbevelen aan hun collega's (met name junior collega's). Docenten in de feedback-plus-consultatie conditie waren gemiddeld het meest tevreden met de hoorcollege-evaluaties en met de gehele interventie. Zij waren gemiddeld zeer tevreden over de consultatie en gaven aan deze aan te bevelen aan zowel hun junior als hun senior collega's.

Met betrekking tot de antwoorden op de vragen over wat docenten hadden geleerd over hun hoorcollegeonderwijs waren er duidelijke verschillen tussen de condities. De feedback-plus-consultatie conditie verschilde positief significant ($p < .01$) van de controle conditie op achttien van de tweeënveertig vragen met betrekking tot leren. De effectgroottes bij deze verschillen waren in alle gevallen groot (Cohen's $d > .80$). De vragen waarop deze verschillen zijn gevonden hadden betrekking op verandering in kennis, aandacht, attitude

en vaardigheden van de docenten ten opzichte van specifieke didactische dimensies, en ten opzichte van het voorbereiden, uitvoeren en evalueren van hun hoorcolleges. Daarentegen verschilde de feedback conditie slechts op één afhankelijke variabele positief significant van de controle conditie met een medium effectgrootte (meer kennis over hoe studenten de hoorcolleges ervaren). Daarnaast maakten docenten naar eigen zeggen in de feedback-plus-consultatie conditie significant meer plannen voor verbetering van de hoorcolleges naar aanleiding van de eerste twee hoorcollege-evaluaties, in vergelijking met docenten in de feedback conditie. Samengevat was het leereffect volgens de docenten aanzienlijk groter in de feedback-plus-consultatie conditie dan in de feedback conditie.

Effecten op docentgedrag en de leeropbrengst van hoorcolleges (volgens studenten)

In hoofdstuk 3 (experiment 1) en 5 (experiment 2) is onderzocht of de hoorcollege-evaluatiescores (ISQ scores) gedurende de cursus vooruitgaan in de feedback conditie en/of de feedback-plus-consultatie conditie, ten opzichte van de controle conditie. Op deze wijze zijn de effecten van de interventies op (de studentperceptie van) het gedrag van de docent tijdens hoorcolleges en op (de studentperceptie van) de leeropbrengst van de hoorcolleges onderzocht.

Er zijn geen significante effecten gevonden van de feedback conditie in vergelijking met de controle conditie, op zowel de studentperceptie van het docentgedrag als de studentperceptie van de leeropbrengst van het hoorcollege (experiment 2). In de feedback-plus-consultatie conditie zijn wel significante verschillen gevonden met de controle conditie, op zowel studentperceptie van docentgedrag als studentperceptie van de leeropbrengst (experiment 1 en 2). In beide experimenten gingen docenten in de feedback-plus-consultatie conditie, in vergelijking met de controle conditie, significant vooruit gedurende de cursus op de totaalscore van de ISQ (*Total Instructional Skills*) en op diverse specifieke gedragsdimensies. Studenten van docenten in de feedback-plus-consultatie conditie gaven tevens aan meer van de hoorcolleges te leren gedurende de cursus, vergeleken met de controle conditie. De effectgroottes op de hoorcollege-evaluatiescores waren echter klein tot middelmatig.

Uit nadere exploratieve analyses volgt dat in beide experimenten met name het eerste consultatiegesprek resulteerde in significante effecten op de hoorcollege-evaluaties. Een tweede en derde consultatiegesprek tijdens de cursus had weinig toegevoegde waarde.

De effecten traden hoofdzakelijk op op didactische dimensies van de ISQ die tijdens de consultatiegesprekken waren geselecteerd voor verbetering. Dit resultaat wijst erop dat het effect van de interventie is toe te schrijven aan de specifieke wijze van consultatie en niet alleen aan de aandacht die de docent ontvangt en de tijd die hij reserveert voor de evaluaties tijdens de gesprekken.

De samenhang van de effecten met de leeftijd, doceerkwaliteit en groepsgrootte van de docent

In het tweede experiment is de samenhang tussen de effecten en de leeftijd van de docent, de didactische kwaliteit van de docent (gemiddelde versus hoge kwaliteit, vastgesteld op basis van eerdere onderwijsevaluaties) en groepsgrootte van de docent onderzocht.

Uit de resultaten volgt dat, op alle afhankelijke variabelen, de effecten niet significant verschillend waren met betrekking tot de leeftijd en de groepsgrootte van de docenten. Er is wel samenhang gevonden tussen de didactische kwaliteit van de docent en de effecten van de interventies op de hoorcollege-evaluatiescores (studentperceptie van docentgedrag). Bij docenten met een hoge doceerkwaliteit in de feedback-plus-consultatie conditie zijn effecten gevonden op meer verschillende didactische gedragsdimensies dan bij docenten met een gemiddelde doceerkwaliteit. Mogelijk treedt dit verschil in effect op doordat docenten met een hoge doceerkwaliteit beter in staat zijn om hun gedrag in een korte tijd succesvol en effectief (zichtbaar voor studenten) aan te passen. De relatief korte periode waarop de effecten van de interventies zijn onderzocht (een cursusperiode van gemiddeld 8 weken) is een beperking van de onderzoeken in dit proefschrift. Een belangrijke suggestie voor vervolgonderzoek is dat ook de lange-termijneffecten van deze interventies worden onderzocht.

Wetenschappelijke bijdrage van dit proefschrift

De resultaten in dit proefschrift geven inzicht in de effecten van tussentijdse feedback en consultatie op diverse afhankelijke variabelen en daarmee in het proces en de toegevoegde waarde van deze interventies. Een tweede belangrijke bijdrage van dit proefschrift is het gebruik van multilevel analyses op de studentdata. Ten tijde van veel eerder onderzoek naar de effecten van feedback en consultatie was deze analysemethode nog niet beschikbaar. Deze analysemethode laat toe dat de resultaten gecorrigeerd worden voor verschillen tussen docenten, tussen studenten binnen een groep en tussen de meetmomenten (hoorcolleges). Zonder deze correcties bleken de effecten op de hoorcollege-evaluatiescores vele malen groter en op meer variabelen aanwezig, dan bij toepassing van de correcties. Dat betekent dat positieve resultaten uit eerdere onderzoeken mogelijk vertekend zijn.

De psychometrische kwaliteit van het hoorcollege-evaluatie-instrument is eveneens met multilevel exploratieve en confirmatieve analyses onderzocht. Hoofdstuk 2 biedt daarmee een voorbeeld van deze techniek bij het onderzoeken van de kwaliteit van onderwijsevaluatie-instrumenten en andere instrumenten met geneste data.

Praktische implicaties van de resultaten

Gezien de resultaten is enkel het aanbieden van tussentijdse feedback voor hoorcollegedocenten niet aan te bevelen; docenten leren, naar eigen zeggen, weinig van de feedback en studenten observeren geen verandering in docentgedrag of de leeropbrengst tijdens de hoorcolleges. Het combineren van tussentijdse feedback met individuele consultatie verdient wel aanbeveling, met name als men voor universitaire docenten omstandigheden wil creëren waarin zij meer kunnen leren van de eigen onderwijspraktijk. Voor zowel docenten van verschillende leeftijd, doceerkwaliteit, faculteit, academische rang en met verschillende grootte groepen is dit een leerzame interventie gebleken. De effecten op de studentperceptie van het docentgedrag en de leeropbrengst voor studenten waren echter niet groot. Eerder onderzoek laat zien dat de effecten op studentperceptie van het docentgedrag groter zijn als individuele consultatie wordt aangeboden in combinatie met andere interventies, zoals een workshop of cursus. Een dergelijke combinatie is daarom aan te bevelen.

Het aanbieden van slechts één consultatiegesprek tijdens een cursus verdient aanbeveling. In vergelijking met resultaten uit eerder onderzoek is daarbij de coachende aanpak aan te bevelen boven een meer descriptieve aanpak (waarbij de docent een interpretatie van studentevaluaties ontvangt en suggesties voor verbetering).

Rest nog de vraag aan welke docenten men een dergelijke interventie het beste kan aanbieden. Op het eerste gezicht ligt het voor de hand om voornamelijk docenten met een aantoonbare lage doceerkwaliteit en/of enkel beginnende docenten tussentijdse feedback met consultatie aan te bieden, met name gezien de kosten die individuele consultatie met zich meebrengt. Echter, universitaire docenten hebben, zoals gezegd, tot op heden slechts zeer beperkte didactische ondersteuning en (na)scholing mogen ontvangen. Vanuit dit perspectief, en op basis van de resultaten in dit onderzoek, is een dergelijke interventie passend voor een diversere groep universitaire docenten (in zowel huidige BKO-trajecten als daarbuiten). Feldman en Paulsen (1999) benadrukken dat een docentondersteunende onderwijscultuur niet alleen noodzakelijk is om de didactische kwaliteiten van minder goede docenten te bevorderen, maar ook om kwalitatief goed onderwijs te behouden en goede docenten te motiveren.

In het huidige onderzoek zijn enkele ervaren docenten getraind om de coaching uit te voeren. Op grotere schaal kan dit mogelijk kosten besparen en daarbij de interne kwaliteitszorg bevorderen. Handal (1999) trekt daarbij een parallel met kwaliteitszorgsystemen binnen de universitaire onderzoekscultuur; naast grondige investering in het opleiden van jonge onderzoekers, bestaat er een voortdurend (peer)reviewsysteem voor onderzoek van zowel kwalitatief goede als minder goede onderzoekers. Het reviewen zelf wordt daarbij beschouwd

als een belangrijke leerzame ervaring en een basisonderdeel van de functie als professioneel onderzoeker.

Gezien het maatschappelijke belang van kwalitatief goed universitair onderwijs, en de bijbehorende hoge eisen aan de docent, verdient de universitaire docent in zijn lesgeven niet minder feedback en ondersteuning op zijn onderwijs dan in zijn onderzoek. Lesgeven is nou eenmaal een vak apart.

Dankwoord

En dan nu eindelijk het meest gelezen onderdeel van proefschriften: het dankwoord. Het is ook het onderdeel wat ik het liefst heb willen schrijven, want mijn dank is groot aan velen. Aan dit project hebben 112 docenten, duizenden studenten, 5 coaches met steun van de afdelingen waar ze werken, 7 onderwijsdirecteuren en secretariaten, 4 technici, 13 onderzoeksassistenten, 1 opperassistent, de facultaire en centrale studentenraad (FSR-FMG en CSR), de afdeling Academische Zaken, het College van Bestuur, 2 promotores, 1 co-promotor, 2 quasi-co-promotores, 2 onderzoeksgroepen, verschillende binnen- en buitenlandse collega-onderzoekers en bijzondere vrienden hun medewerking verleend. Het mag duidelijk zijn dat ik de grootste tegenzin had om de introductie van dit proefschrift in de ik-vorm (ik heb onderzocht…) te moeten schrijven.

In dit dankwoord dank ik graag een groot aantal mensen persoonlijk.

Het College van Bestuur, Universiteit van Amsterdam.
Dit project is gefinancierd door de UvA. In de introductie van dit proefschrift heb ik aangegeven dat onderzoeksgeld voor onderzoek naar het hoger onderwijs zeer beperkt beschikbaar is. Het College van Bestuur heeft, op voorspraak van de toenmalige decaan van de FMG, Dymph van den Boom, besloten dit project te financieren en mij de kans te geven om op dit onderwerp te promoveren. Dit getuigt in mijn ogen van veel betrokkenheid bij de kwaliteit van het onderwijs van de UvA. Het toont ook aan dat het CvB belang hecht aan ondersteuning van de UvA-docenten die jaar in jaar uit de hoorcolleges verzorgen voor grote groepen studenten (vaak alleen). Zeer zeer veel dank hiervoor.

Promotor prof.dr. Han van der Maas, hoofd van de programmagroep Psychologische Methodenleer (PML).
Han, dankzij jou zat ik de afgelopen vijf jaar bij de programmagroep Methodenleer. Ik begreep in het begin weinig van wat er in de labmeetings werd besproken. Inmiddels kijk ik ook geobsedeerd naar computerschermen met R, MLwiN en MPlus, draaien ook mijn berekeningen tenminste een nacht door, vind ik ook dat SPSS zo snel mogelijk moet worden vervangen door R, mopper ik op zwakke methoden van onderzoek in het veld, en geniet ik van een mooi stukje code. De ontwikkeling die ik door heb gemaakt, heb ik aan jou en de groep te danken. Door jou ben ik onderdeel geworden van de programmagroep. Je hebt me regelmatig actief betrokken en je bemoeid met mijn wetenschappelijke en methodologische ontwikkeling en ondersteuning. Je wijze van leidinggeven, kalm, scherp, persoonlijk, vind ik bewonderenswaardig. Dank.

Promotor prof.dr. Jan van Driel, directeur ICLON (Interfacultair Centrum voor Lerarenopleiding, Onderwijsontwikkeling en Nascholing) van de Universiteit Leiden.
Jan, we zijn allebei vele uurtjes afgereisd naar Leiden en Amsterdam. Dank dat je al die jaren zoveel moeite hebt gedaan en zo betrokken bent geweest bij mij en dit project. Ik heb genoten van onze congresreizen samen, naar de ORD, AERA, SRHE en EARLI SIG Higher Education. Daar waren we beiden echt in ons element en vonden elkaar regelmatig in de massa voor een lunch, om alle avonturen te delen. Je bent mijn toegang geweest tot onderwijsonderzoekland. Je hebt me opgenomen in de onderzoeksgroep in Leiden. Dank voor je begeleiding en de bijzondere vriendschap die tussen alle bedrijven door is ontstaan.

Co-promotor, em.prof.dr. Don Mellenbergh (PML).
Don, het is voor mij heel bijzonder om een van je laatste promovendi te mogen zijn. Zeker na de meer dan tachtig proefschriften die je op je naam hebt staan. Ik heb zelden iemand mogen ontmoeten met zoveel kennis en daar mocht ik van leren. Het was een eer om met je te mogen samenwerken.

Prof.dr. Conor Dolan (PML).
Dan wordt het hoog tijd om iets te zeggen over quasi-promotor Conor Dolan. Lezers moeten weten dat de kwaliteit van de onderzoeken in dit proefschrift voor een zeer groot deel te danken is aan deze bijzondere man.
Conor, ontelbare keren ben je bij mij binnengelopen en heb je je bemoeid met mijn werk. "Heb je hier al aan gedacht?" "Je moet het zo doen." "Ik kijk er wel naar." "Hallo, ik zit weer eens jouw werk te doen." "Schrijf toch eens duidelijk." "Hehe, dit is uitstekend." Je hebt me bij de les gehouden en me bij grote druk overeind gehouden. Je hebt me uitgedaagd tot aan het maximum. Ik heb daar geen woorden voor.

Drs. Harrie Vorst (PML).
Nog zo'n wonderlijke man en quasi-promotor.
Harrie, na mijn afstudeerproject met Rachna ben je bij Klaas Visser (onderwijsdirecteur Psychologie) binnengelopen: "je moet die twee meiden een baan aanbieden. Ik wil het anders ook wel zelf betalen." Na maanden kritisch observeren, sprak je je vertrouwen in mij uit en dat ben je in de daarop volgende jaren blijven doen. In de Roeter hebben we de opzet van de onderzoeken tot in detail bediscussieerd. "Nog een espresso alstublieft", en na werktijd "doet u maar een Duvel en een rosé voor de dame." Er bestaat een stapel bierviltjes met tekeningetjes voor het design, de juiste wijze van indeling, analyses, etc. Ik heb werkelijk onvoorstelbaar veel van je geleerd over het opzetten van complexe veldexperimenten. En onderweg heb ik mogen observeren hoe jij studenten op bijzondere wijze motiveert en ze

met plezier harder laat werken dan ze ooit hadden gedacht. Je hebt je ook zeer betrokken gevoeld bij mij persoonlijk. Ik dank je voor je steun in de afgelopen jaren.

Han, Jan, Don, Conor en Harrie. Met elkaar waren we een bont gezelschap. Ik heb me vaak als Alice in Wonderland gevoeld. U blijft mijn beste Heren.

Docenten, studenten, coaches (Anne, Willem, Leendert, Rifka en Marieke), onderzoeksassistenten, medewerkers ondersteuning van de afdelingen, en de onderwijsdirecteuren van de opleidingen, mijn dank voor jullie medewerking is groot. Het vele werk zat me bij tijd en wijle tegen. Ik heb op die momenten vaak gedacht aan iedereen die heeft meegewerkt, dan vond ik weer de motivatie om door te gaan. Dank.

Twee mensen in dit project dank ik in het bijzonder, Rachna en Maarten.
Rachna, onze samenwerking en vriendschap is wonderlijk. We hebben samen het eerste experiment uitgedacht en uitgevoerd. Credits voor de interventies in dit proefschrift gaan ook naar jou.
Maarten, wat een chaos was het geweest als jij niet zo strak alle college-evaluaties had weten te organiseren in het tweede experiment. De vele discussies met jou hebben me scherp gehouden. Dank, allebei, voor een inspirerende tijd samen.

Binnen de afdeling Psychologie dank ik graag (oud-)collega's van de programmagroep PML: Marijke, EJ, Jelte, Laurens, Raoul, Ruud, Marthe, Helen, Dora, Alex, Abe, Sacha, Verena, Sophie, Annemarie, Leendert, Jan, Gunter, Sanne, Josine, Robert, Matthieu, Dingmar, Pieter. In het bijzonder dank ik Ineke voor haar hulp en glimlach iedere dag, Peter, Sanja, Kees-Jan, Charlie en Angels voor hun vriendschap en Denny voor alle inspiratie. Dylan, je bent om nooit meer te vergeten.
Binnen de afdeling dank ik ook Marco en Caspar voor alle technische ondersteuning en hun flexibiliteit. Last, but not least, Klaas, dank voor je adviezen, aanmoediging en ontspanning tussen de bedrijven door. We raken niet uitgediscussieerd over het onderwijs. Ik ga er van uit dat dat zo blijft.

Buiten de UvA heb ik met de onderzoeksgroep van het ICLON in Leiden veel over nieuw onderwijsonderzoek mogen discussiëren. Zij hebben ook regelmatig met mij meegedacht en mij waardevolle feedback op mijn werk gegeven. Dank aan de (oud-)leden Ben, Rosanne, Jacobiene, Nico, Pauline, Jan (v.T.), Dineke, Wilfried, Amanda, Romi, Chantal, Chris, Rikkert, Carlos, Luce, Claire, Wu, Dadi, Nienke, Michiel, Gerda, Mariska, Mirjam, Tamara, Hans, Fred, Peter, Wil, Albert, Sylvia, Monika, en in het bijzonder aan Klaas, Nelleke, Dirk, Christel en Roeland, voor jullie betrokkenheid.

Special thanks to the members of the AERA SIG Faculty Teaching, Evaluation and Development for their feedback on my work and the opportunity to present my work to others. Huub van den Bergh en Sven de Maeyer, multilevel dank voor jullie tijd en advies.

Lieve familie en vrienden. Jullie hebben mij aangemoedigd en met me meegeleefd.
In het bijzonder dank ik ook Martijn, Jan, Gerry en Marieke hiervoor. Jullie zijn van onschatbare waarde geweest. Dank voor alles wat jullie voor me gedaan hebben.
Lieve Germa, Ran, Shirley, Vief, Klaas en Nelleke. In de jaren van dit promotietraject is het privé zwaar weer geweest. Jullie hebben me geholpen om het roer recht te houden. Thank you so…
Pleun en Dirk, de laatste loodjes wegen het zwaarst, dit was me alleen nooit gelukt.
Femke, Bart, Aggie, Henny, Fredo, Jos, Thea en alle andere lieve mensen om me heen. Het is lente, laten we de bloemetjes buiten zetten!

De laatste woorden richt ik graag tot mijn lieve ouders, Henk en Malty, en mijn tantes Dineke en Corry.
Pa, hoe toepasselijk is de titel van het lied van Stef Bos, *Papa, ik lijk steeds meer op jou*. Als je trots bent op mij, ben ik evenzo trots op jou. Mam, ik heb van jou geleerd om door te gaan, wat er ook gebeurt. Ik dank jullie beide voor alle liefde en steun.
Dineke en Corry, al als klein meisje kwam ik graag bij jullie. Jullie zijn in de loop der jaren mijn tweede thuis geworden en dat heeft me ook gevormd. Dank voor jullie liefde en steun door de jaren heen en bij dit werk. We delen bovendien de liefde voor onderwijs. Ik draag dit proefschrift daarom op aan jullie.