International Journal for Academic Development
Vol. 12, No. 2, November 2007, pp. 73–85

Quality of Instruction Improved by Evaluation and Consultation of Instructors

Heiner Rindermann[a]*, Jürgen Kohler[b] and Gerhard Meisenberg[c]
[a]*University of Magdeburg, Germany;* [b]*Hochschule für Heilpädagogik, Zurich, Switzerland;* [c]*Ross University Medical School, Dominica*

One aim of student evaluation of instruction is the improvement of teaching quality, but there is little evidence that student assessment of instruction alone improves teaching. This study tried to improve the effects of evaluation by combining evaluation with individual counselling in an institutional development approach. Evaluation was conducted in a private school for speech therapy (logopedia) in 35 classes (N = 16 teachers, N = 592 students). Evaluation was done twice within a period of three to twelve months using a standardized questionnaire (HILVE-II) developed for evaluation of university courses. The intervention effect on teaching quality was more than half a standard deviation on the teacher scales. Despite the fact that the counsellor had no pedagogical training, the quality of teaching not only improved quantitatively and qualitatively but also became more homogeneous although the relative rank listing of teachers did not change.

Ein Ziel der studentischen Lehrevaluation ist die Verbesserung von Lehrqualität. Es gibt aber wenig empirische Evidenz dafür, dass studentische Lehrevaluation allein Lehre verbessern kann. Diese Studie versuchte, die Effekte von Veranstaltungskritik in einem Ansatz, der Lehrevaluation mit Beratung verbindet, zu erhöhen. Lehrveranstaltungsevaluation wurde in einem privaten Ausbildungsinstitut für Logopädie in 35 Kursen von 16 Dozenten mit 592 Studierenden durchgeführt. Veranstaltungskritik wurde zweimal praktiziert binnen eines Zeitabschnitts von drei bis zwölf Monaten anhand eines Fragebogens (HILVE-II), der für die Evaluation von Hochschulveranstaltungen entwickelt wurde. Der Verbesserungseffekt auf Lehrqualität war in den dozentenbezogenen Skalen größer als eine halbe Standardabweichung. Obwohl der Berater keine pädagogische Ausbildung hatte, verbesserte sich die Lehrqualität nicht nur in den quantitativen und qualitativen Maßen, sondern wurde auch homogener. Die relative Rangreihung innerhalb der Dozenten blieb aber stabil.

Un des objectifs de l'évaluation, par les étudiants, de l'enseignement est l'amélioration de la qualité de ce dernier. Pourtant, les preuves que l'évaluation de l'enseignement par les étudiants améliore celui-ci sont limitées. Cette étude a tenté d'améliorer les effets de l'évaluation en combinant l'évaluation avec du conseil individuel au sein d'une approche de développement institutionnel. L'évaluation a été effectuée dans 35 classes (N = 16 enseignants, N = 592 étudiants) d'une école privée d'orthophonie (logopédie). L'évaluation a été effectuée deux fois sur une période variant de trois à seize mois en employant un questionnaire standard

*Corresponding author. Institut für Psychologie, Otto-von-Guericke-Universität Magdeburg, PO Box 4120, D-39016 Magdeburg, Germany. Email: heiner.rindermann@ovgu.de

ISSN 1360–144X (print)/ISSN 1470–1324 (online)/07/020073–13

DOI 10.1080/13601440701604849

(HILVE-II) développé pour l'évaluation des cours universitaires. L'effet relatif à la qualité de l'enseignement associé à l'intervention représentait plus d'un demi écart-type sur l'échelle des enseignants. En dépit du fait que le conseiller n'avait pas de formation pédagogique, la qualité de l'enseignement s'est améliorée à la fois au plan quantitatif et au plan qualitatif, en plus de devenir davantage homogène, bien que le positionnement relatif des enseignants n'ait pas changé.

Aims of Student Evaluation of Instruction

Teaching assessment at universities should help to improve the quality of instruction. However, the effectiveness of student evaluation of instruction has—as far as we know—not yet been demonstrated at European universities. The only exception is a thesis by Gijselaers (1988) published in Dutch. There are many anecdotal references, but no systematic tests with repeated measurements. The question remains as to whether evaluation can actually improve the quality of instruction. Teichmann (1999) comments on the lack of European research on "the efficiency of evaluation for teaching and learning". The author demands that in future evaluations "systematic effectiveness-measurements should be introduced …, it has to be researched, what successes have resulted" (p. 247). The aim of the present study is to test the usefulness of student evaluation of instruction accompanied by an individual counselling session and integrated into an institutional development programme. We distinguish institutional and academic development in the following way: academic development deals only with individual improvement of teachers by consultation or training; institutional development includes, in addition, improvement of the organisation and its goals (e.g., appreciation of and orientation towards teaching quality). After a presentation of the four different evaluation models we draw upon—*awareness*, *feedback*, *discussion*, *consultation*—we describe our approach and its results.

Evaluation Models and Their Effectiveness

In the *awareness model*, the implementation of evaluation of instruction is itself conceived as an effective intervention: promoting "quality-awareness" results in improved teaching. In the *feedback model* the instructor is offered an individual report of the results. This feedback provides information about specific weaknesses and strengths of a course (informational function), and is assumed to motivate efforts at improved teaching if there is any target discrepancy (cf. Marsh, 1987, p. 342ff; Balk, 2000). One specification of the feedback hypothesis stresses that feedback is only effective if it informs the teacher about his or her improvable behaviour. However, studies in Anglo-American, Australian (e.g., Marsh, 1987), and German-speaking universities (Rindermann, 2001) have found little or no improvement through mere feedback; further, it has been suggested that negative feedback without counselling and training appears insufficient to motivate improvement (McKeachie, 1997). Webler (1992) proposed a *discussion model*, which incorporates a discussion between students and lecturers. In such a discussion, results should be reported, new information beyond the feedback should be gathered, and possible changes should be discussed and initiated. But Webler (1996) speaks of "processes of petering out", provoked by considerable resistance or indifference. On the whole, no improvements of teaching could be documented. Alean-Kirkpatrick, Hänni, and Lutz (1997) suggested that such discussion should be moderated by higher education counsellors. More recently, Gediga *et al.* (2000) in Osnabrück examined the discussion model at

universities, but measurements in the middle and at the end of the semester in six courses (with 95 students) resulted in no positive changes, except for "massive improvements" in only one course where, in addition to discussion, "a colleague of the project participated as a higher-education-counsellor to help in solving problems" (pp. 62, 77).

In three studies at Heidelberg University in courses of psychology, social sciences, languages, and medicine, the effectiveness of student evaluation of instruction combined with different models was also examined (Rindermann, 1996):

1. Evaluation in the middle and at the end of the semester in the same course, feedback only after the second measurement—*awareness-model*.
2. Evaluation in the middle and at the end of the semester in the same course, feedback after the first and the second measurement—*feedback model*.
3. Evaluation in the middle and at the end of the semester in the same course, feedback after the first and the second measurement, discussion of the first feedback with the students in the course at the lecturer's own discretion, without any external support—*discussion model*.
4. Finally a comparison beyond the three studies: medium-term changes over several semesters with the same lecturers in the same courses (mix of *feedback* and *discussion model*).

The analyses with different sets of data, assessors, and measures revealed no significant improvement. The effect sizes ($d = \frac{\bar{x}_2 - \bar{x}_1}{s}$) averaged 0.04, with a range from –0.03 to 0.18 in teacher scales (Rindermann, 2001). The comparison between courses discussing feedback and no discussions showed no significant effects. However, the discussion or non-discussion of evaluation results was related to teaching competence and commitment. Lecturers whose teaching was judged by students as competent and committed were more likely to discuss evaluation results and courses with their students. This could simply mean that instructors prefer discussing good results. Or else it could indicate that instructors who are committed to teaching not only offer better instruction but are also more interested in discussing evaluation results with their students. Also external evaluators using the same questionnaire (HILVE, described below) with minor rephrasing, who were not aware of whether feedback had been discussed, assessed courses with teacher–student discussion of the results more positively.

To improve the quality of instruction, most authors, as already noted, recommend that student evaluation of university instruction be followed by counselling, with explanation of feedback, suggestions, motivational encouragement, and further training (Marsh, 1987; Penny & Coe, 2004; Piccinin, 1999; Wilson, 1986). In other words, in addition to information about their strengths and weaknesses, instructors also need advice on how to think about their teaching differently and how to change their behaviour.

According to such *consultation models*: (1) lecturers should be informed in a counselling session about the results (which they have been given beforehand); (2) their perception of problems should be called into question (to amplify and to confront); (3) single modifiable behaviour units should be targeted (to simplify and to highlight problems); (4) dysfunctional patterns of attribution should be interrupted (for example to reattribute bad results); (5) concrete suggestions for changes should be worked out (to suggest and to model); and (6) instructors should be alerted to the need for behaviour modifications (appreciation of teaching goals and values), and should be supported emotionally to initiate them (e.g., Wilson, 1986).

Studies about consultation that employed repeated measurements and control groups without counselling have shown significant effects for this model (cf. the meta-analyses by Cohen, 1980, feedback $d = 0.20$, feedback and counselling $d = 0.64$; by Menges & Brinko, 1986, feedback $d = 0.22$, feedback and counselling $d = 1.10$; and by Penny & Coe, 2004, counselling $d = 0.58$). Even improved performance on final exams could be observed ($d = 0.44$; Hampton & Reiser, 2004). Since most of these studies were done in North American universities, their generalisability is uncertain. According to Penny and Coe (2004, p. 248), "more research on the effects of consultative feedback in settings other than North America is sorely needed".

It seems that student evaluation at universities that leads to positive change can be implemented successfully only if accompanied by counselling. The effectiveness of this approach is tested here in a European context.

Reichenau School for Speech Therapy

The School for Speech Therapy on Reichenau Island (Lake Constance) is a privately maintained institution with permanent and part-time instructors (permanent about 35%, part-time instructors 65%) that offers students a 3-year programme to become speech therapists. The programme consists of at least 1740 lessons of "theoretical-practical instruction" and at least 2100 hours of "practical training". Courses—usually seminars without student presentations—are attended by about 20 students who study as a cohort. The students are young adults, most of whom have the *Abitur* (general qualification for university entrance).

The theoretical-practical instruction includes medicine, psychology, and pedagogy, but the main emphasis is on speech therapy itself. Here, application-oriented treatment models of speech therapy are presented by full-time speech therapists. Courses in medicine and social sciences are taught exclusively by part-time instructors; they have gained their specialised subject knowledge through university education and professional experience.

Conditions at this private school with a clear management structure differ from German state universities but resemble those at many international universities: instruction represents a central quality criterion; students pay for their courses; lecturers cannot be dismissed easily, either because they have a permanent post or—in the case of part-time lecturers—because there are hardly any other instructors to hire in the region. However, persistent negative evaluations in the central area of responsibility could lead to termination of a lecturer's employment or to a change in area of responsibility. Therefore, monitoring the quality of instruction is an important administrative concern, particularly since it has an impact on the institution's reputation. There is no academic development unit to support teaching improvement.

Method

Sample

Forty-eight courses taught by 27 instructors were evaluated by 15–20 students per course. Some students from the six different cohorts participated in more than one course and thus completed forms for more than one course; however, overall the students in each course were different. Evaluations of an instructor in two different courses from the same subject area—with a consultation in between—were available for 16 instructors. The time between the first and second student evaluation of courses varied from 3 to 12 months.

Instrument

A standardised 51-item questionnaire (HILVE-II) was used, containing four dimensions that were defined according to theoretical criteria (Rindermann, 2001): *instructor scales* (structure, breadth, depth, teaching competence, engagement/enthusiasm, climate, support, interaction management); *student scales* (student presentation, student participation, student discipline, student self-assessment of competence/ability level); *external conditions* (redundancy, workload/demand/requirement, topic, diligence outside the classroom); and *teaching and course effectiveness* (interestingness of course, learning quantitatively and qualitatively, improved motivation/promotion of interests, general course assessment). Except for the dimensions of workload and redundancy, where a medium value is favourable, high results on the 1 (low)-to-7 (high) scale are optimal. Each scale contains between two and four items. Examples for the most important *instructor scales* include: "The course is well organized" (structure); "Topics are illustrated by examples" (breadth); "The instructor encourages me to think things through" (depth); "The instructor can explain complicated topics" (teaching competence); "The instructor shows enthusiasm in teaching" (engagement); "The instructor is friendly" (climate); "Students are supported outside the course" (support); "The instructor promotes students' questions and comments" (interaction management). There were also three open-ended questions, for example, "What do you like most about this seminar?"

Evaluation Process

1. Evaluation was done in the middle of the semester. Questionnaires were handed out and collected by the instructors themselves.
2. The questionnaires were sent to an external office of evaluation through the school manager.
3. The questionnaires were analysed externally to develop numerical, graphic feedback, and collated verbal feedback. The feedback was sent back to the school manager. The interval between the collection of the questionnaires and the feedback averaged 4 weeks.
4. Feedback was presented individually to the lecturer by the school manager who had no pedagogical training (but 6 years of teaching experience). Positive or negative feedback, the lecturer's self-assessment, didactical problems and interaction patterns, and concrete possibilities for improvement were discussed (see consultation model described below).
5. A second evaluation of another course by the same lecturer (in most cases with another group of students) was done 3–12 months later, following the procedure described above.
6. In an all-institute conference with instructors and students at the end of the academic year (July), the overall results and further steps for institutional improvement were discussed.

There was no control group in the Reichenau study. However, earlier studies at universities had shown that significant improvements could *not* be achieved by evaluation alone or evaluation and feedback (d = –0.03 to 0.22; Cohen, 1980; Menges & Brinko, 1986; Rindermann, 2001, with HILVE d = –0.03 to 0.18). The study design did not permit the isolation of the following factors: evaluation vs. feedback vs. counselling vs. conferences of lecturers and students; part-time vs. permanent lecturers; evaluation and counselling by the school manager vs. someone with pedagogical training vs. an external person; evaluation and counselling at different institutions; a control group without feedback and counselling (given the

institute purpose that the evaluation improve quality of education and increase reputation). But this evaluation study is comparable with earlier studies carried out by the first author where the listed factors were available in other combinations (Rindermann, 2001).

In other words, while the research design does not meet the strict standards of an experimental evaluation (Rindermann, 2002; Shadish, Cook, & Campbell, 2002), it uses a scientifically validated instrument and employs assessments before and after intervention. Research on evaluation is research in applied settings, and is therefore subject to the constraints imposed by these settings. A comparison is possible with earlier HILVE studies at German state universities, although the effects of counselling cannot always be distinguished from the effects of institutional background variables.

Counselling

Due to the results of the previous studies at state universities with the HILVE instrument described above, a *counselling procedure* was chosen. Each lecturer met for a 1-hour individual counselling session with the school manager, who was a qualified psychologist (trained at university in clinical and counselling psychology). Instructors were given descriptions of the dimensions, statistical feedback, and a copy of the hand-written student comments. The statistical feedback included mean values (non-standardised and standardised on a norm-scale of $M = 100$ and $SD = 10$), standard deviations, minima, maxima, number of observations per dimension and per item, and a figure of standardised mean values ($M = 100$ and $SD = 10$) for each dimension. Instructors also received students' comments to the questions "What do you like most about this seminar?", "What do you like least of all?", and "Suggestions for improvement?"

The intent of the counselling session was to increase the relevance of the evaluation. The school manager and the instructor discussed reasons for positive or negative reports, ways of teaching and didactic strategies, and possibilities for improvement. The feedback confronted the instructor's perception of his/her own teaching with perceptions by the students in order to modify stable cognitive and behavioural structures. The counselling session was supportive and served to develop new teaching techniques. Its main focus was to suggest changes in the preparation for instruction and the teaching itself, for example, how to reduce the number of topics or how to use new didactic methods. The instructor was also encouraged to discuss the results in class; about 50% of the teachers did this. The counsellor concluded the talk by stressing his interest in the instructor's further development and his readiness for further face-to-face consultation if wished.

In the institutional context, the evaluation project was an official quality control measure that covered the whole staff. In staff conferences, both before and after evaluation, the evaluation was treated as part of an institutional development process to enhance teaching and learning. In the same way, in the all-institute conference at the end of the year, the aim was for instructors to learn from each other, and for students to identify more strongly with the school by participating in the enhancement of its teaching programme.

Analyses

The analyses were calculated in two variants: (a) all instructors ($n = 16$); and (b) only those instructors whose teaching was marked 6.5 or lower on the 1–7 scale ($n = 13$, the three best

excluded). Intervention effects were expressed as effect sizes, which allow a direct estimation of treatment effects and comparison between studies independent of sample sizes (Cohen, 1988). The effect size *d* is formed by subtracting the first measurement from the second and dividing the difference by a standard deviation ($d = \frac{\bar{x}_2 - \bar{x}_1}{s}$ (1)), usually the standard deviation of the pre-intervention measurement. The conventional formula is only optimal for independent means (Cohen, 1988). It underestimates the effect with dependent means and correlation between measurements.

Therefore, Cohen (1988, p. 48) recommends the following formula for dependent means $d = \frac{\bar{x}_2 - \bar{x}_1}{S_D}\sqrt{2}$. (2) Here, s_D is the standard deviation of the discrepancies $(\bar{x}_2 - \bar{x}_1)$, $\sqrt{2}$ represents the adjustment to the norms by Cohen of $d = 0.2$ as a low effect, 0.5 as a medium effect and 0.8 as a high effect.

Results

Systematic improvements could be observed in the instructor dimensions (Table 1, Figure 1). Effects were greatest for teaching competence ($d = 0.94$), engagement/enthusiasm ($d = 0.78$) and depth ($d = 0.70$), and smallest for support (feedback and support, $d = 0.38$) and climate (instructor's kindness and cooperativeness, $d = 0.47$). Improvements were also observed on the teaching and course effectiveness scales, with greatest improvements in learning quantitatively ($d = 0.76$) and the overall assessment of the course ($d = 0.85$). Student discipline (no noise and disruption in class, few absentees) increased ($d = 0.54$), but diligence outside the classroom (preparation and further study, work needed) remained unchanged and redundancy increased. The latter cannot be interpreted as positive or negative. Redundancy is optimal when it is medium to low, but not zero.

According to Cohen (1988), differences of $d = 0.20$ are regarded as low, from 0.5 as medium, and from 0.8 as high. Most changes we observed were $d = .30$ to .80. Effects on teaching, which was the main target of the intervention, were medium to high. The evaluation of different courses by different students could result in effects being underestimated. In most other studies of this kind, measurements were done twice in one course in one semester (cf. Menges & Brinko, 1986), but in Reichenau courses and students changed. Only the instructor and the subject matter were the same. Unsystematic changes are likely to reduce the measurable effect of a treatment. But the effects here are *not short-term* and they are *not specific to participants of one course* (see similar cross-course approach of Piccinin, Cristi, & McCoy, 1999, and Piccinin & Moore, 2002).

Measurable improvements were hardly possible for the three most highly rated lecturers, and these were therefore excluded from the second analysis (Table 2). Now the improvement was $d = 1.03$ on the scales for teaching and course effectiveness, $d = 0.96$ on the instructor scales, and $d = 1.26$ on the important scale of teaching competence. Thus evaluation and counselling represented a very efficient method to improve teaching for critically and moderately assessed instructors.

The overall improvements could be attributed mainly to improvements of initially less positively assessed lecturers. Changes in the medium and high sectors—mostly positive—were small. Therefore the standard deviations decreased. The teaching staff (N = 16)

Table 1. Difference between first and second measurement in effect size *d*

all instructors

n = 16 instructors repeated, 35 courses, 592 questionnaires, between 12 and 41 questionnaires for each instructor and measurement point (for two instructors: different courses at one measurement point summed), measurement unit is instructor

dimension	**instructor scales**	structure	breadth	depth	teaching competence	engagement	climate	support	interaction mana.
d	**0.65⋆**	0.66[t]	0.64[t]	0.70[t]	0.94⋆	0.78⋆	0.47	0.38	0.59
M_1	**5.08**	5.16	5.32	4.73	4.81	5.12	5.99	4.83	4.67
s_1	**1.00**	1.20	0.89	1.12	1.20	1.02	0.77	1.05	1.19
M_2	**5.38**	5.48	5.56	5.08	5.31	5.48	6.09	5.00	5.01
s_2	**0.62**	0.80	0.61	0.77	0.74	0.57	0.68	0.75	0.80

dimension	**student scales**	student presentation (n=3)	participation	discipline	competence (n = 6)
d	**0.43**	0.35	0.39	0.54	0.12
M_1	**4.84**	4.52	4.86	5.50	4.04
s_1	**0.45**	0.60	0.42	0.74	0.50
M_2	**5.13**	4.77	4.98	5.83	4.08
s_2	**0.34**	0.71	0.47	0.46	0.16

dimension	**external conditions**	redundancy	workload/demand	topic	diligence
d	**0.24**	0.47	0.23	0.21	0.05
M_1	**3.43**	2.13	4.06	5.07	2.48
s_1	**0.31**	0.32	0.56	0.42	0.71
M_2	**3.51**	2.27	4.12	5.15	2.51
s_2	**0.35**	0.41	0.32	0.54	0.56

dimension	**teaching effectiveness**	interestingness	learning quantitative	learning qualitative	motivation improved	general course
d	**0.64[t]**	0.45	0.76⋆	0.60	0.51	0.85⋆
M_1	**4.96**	4.80	4.84	5.40	4.67	5.07
s_1	**0.98**	1.34	0.94	0.69	1.01	1.11
M_2	**5.27**	5.06	5.21	5.66	4.94	5.48
s_2	**0.58**	0.87	0.54	0.41	0.64	0.69

Notes: t $p < 0.1$; ⋆ $p < 0.05$, ⋆⋆ $p < 0.01$; in student presentation and in competence only data from three or six instructors (only courses with at least 10 students' evaluations in these dimensions), mean and standard deviation in an answer scale from 1–7.

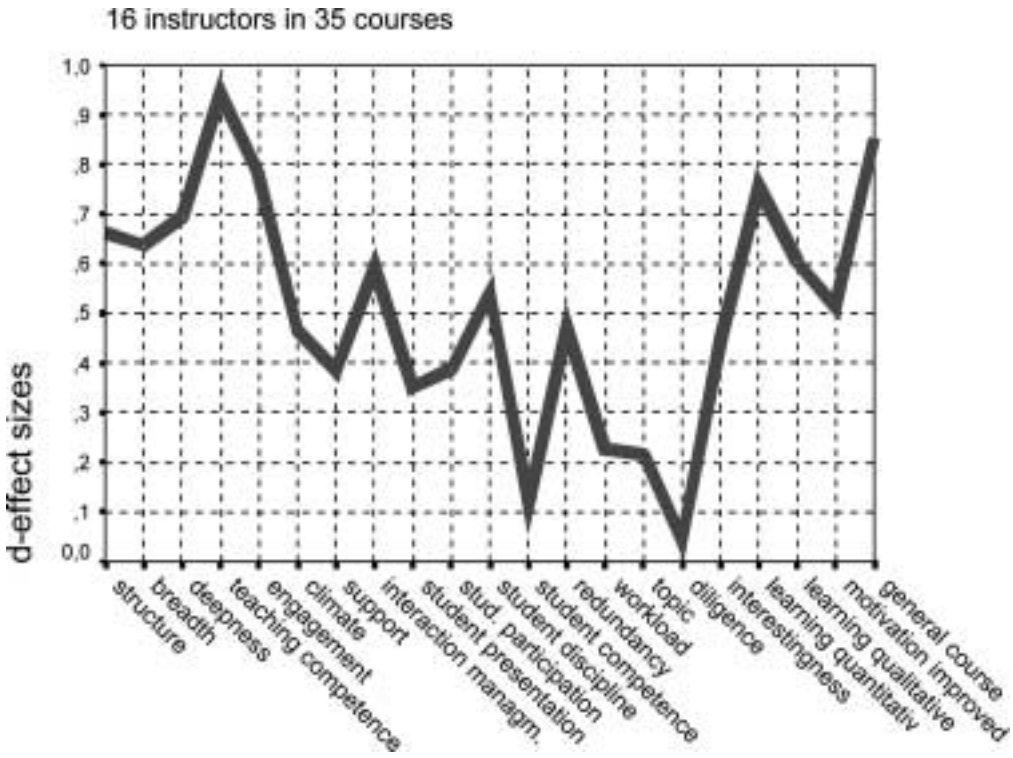

Figure 1. Modification in students' evaluation: *d*-effect sizes for all dimensions

became more uniform regarding their teaching abilities. Figure 2 illustrates this for teaching competence.

The correlation between the initial value and the difference score is $r = -.81$ (N = 16). Nevertheless, the pattern in Figure 2 cannot be explained entirely as regression to the mean because the mean itself changed. Teaching improvement could be demonstrated not only quantitatively, but also qualitatively by students' verbal comments. Examples from the second measurement for students' answers to the question "What is especially good in the course?" were: "Lecturer has been transformed; very well prepared, seems competent and very friendly. Keep it up!" Or: "Lecturer has been very well prepared in the last few weeks. Good planning".

Nevertheless, the majority of critically assessed instructors remained weaker than favourably assessed instructors in the second evaluation. The correlation between first and second measurement in the instructor scales was $r_{tt} = .87$ and in the effectiveness scales $r_{tt} = .79$ (N = 16). For external conditions it was $r_{tt} = .57$ and in the student scales $r_{tt} = .19$.

Discussion and Conclusions for Evaluation Practice

Based on the course ratings, teaching at the institute improved substantially after use of the consultation procedure. The improvements were stable across semesters, even when assessed by different students. This is similar to results by Piccinin and Moore (2002). In the central

Table 2. Difference between first and second measurement in effect size *d*

Without the three best instructors (s_D taken from complete sample used in Table 1)
n = 13 instructors repeated, 29 courses, 485 questionnaires, between 12 and 41 questionnaires for each instructor and measurement point (for two instructors: different courses at one measurement point summed), measurement unit is instructor; only *d* presented

dimension	**instructor scales**	structure	breadth	Depth	teaching competence	engagement	climate	support	interaction mana.
d	**0.96****	0.91*	1.02*	1.06*	1.26**	1.14**	0.69	0.73t	0.87*
dimension	**student scales**	Student presentation (*n* = 2)		participation		discipline		competence (*n* = 5)	
d	**0.63****	0.07		0.56		0.76t		0.69	
dimension	**external conditions**	redundancy		workload/demand		topic		diligence	
d	**0.33**	0.43		0.24		0.28		0.37	
dimension	**teaching effectiveness**	interestingness		learning quantitative		learning qualitative		motivation improved	general course
d	**1.03****	0.87*		1.14**		1.01**		0.93*	1.20**

Notes: t' p < 0.1; * p < 0.05; ** p < 0.01; in student presentation and in competence only data from two or five instructors (only courses with at least 10 students' evaluations in these dimensions); 0.63 in student dimensions is significant at 1%, because for significance testing the reduced s_D-value is used automatically, for comparability reasons the effect sizes were calculated using the total-sample-s_D.

dimension of *teaching competence*, the improvement amounted to approximately one standard deviation. The less favourably assessed instructors showed particular improvement, although they did not achieve the ratings of the more favourably assessed lecturers after counselling. The full potential could probably only be tapped through training approaches (cf. Leitner, 1998). However, not everyone wishes to invest in becoming a better teacher! The main aim should be to avoid negative outliers and to enable all instructors to become more successful teachers (cf. Patton, 1997).

This is the first study demonstrating a positive effect from evaluation and counselling in a German-speaking country. The effect was seen although we did not use a pedagogical specialist for the counselling; a conventional education in clinical and counselling psychology at university seemed to be a good basis. Consultation itself is a powerful intervention; pedagogical specialisation or even length of consultation is less important ("even brief consultation can result in statistically significant and meaningful teaching improvement"—Piccinin, 1999, p. 81). The integration of counselling was conceived as an institutional development to improve teaching and enhance the reputation of the institute. We have been uncertain to what extent the specific conditions in Reichenau are important: many part-time lecturers, many with no permanent status; orientation towards teaching rather than research; private institution; clear management structure with plan for teaching quality, etc. However, recently, an adaptation of the procedure has shown positive results in an institutional development exercise at a university of applied sciences as well (Dresel, Rindermann, & Tinsner, 2007). It is also remarkable that there are *systematic effects on student behaviour and external*

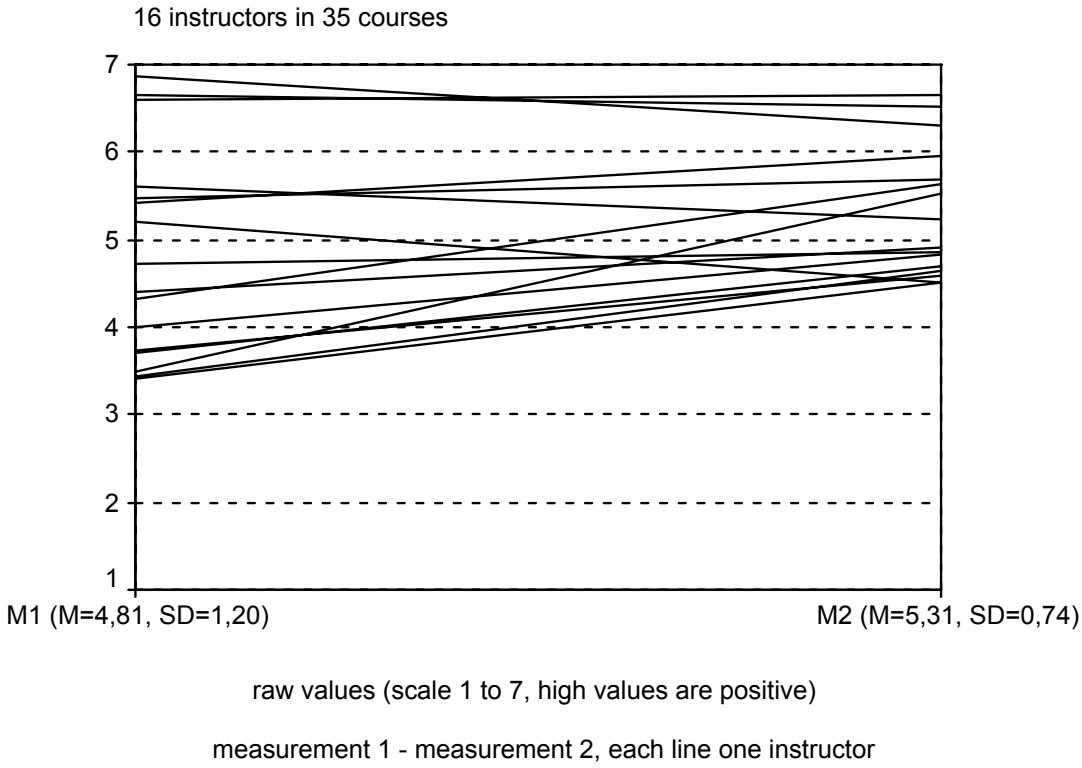

Figure 2. Modification in teaching competence: improvement of mean and reduction of standard deviation in teaching competence

conditions apart from changes in instructor behaviour. Students now report being absent less frequently, less disruptive in class, and more committed to learning. The all-institute conferences at the end of the academic years for instructors and students might have been important for this. One conclusion is that evaluation at universities should place more emphasis on counselling. Evaluation cannot be successful as an isolated administrative routine, but only as a component of an institutional development activity that aims for improved teaching quality leading to enhanced reputation.

Mere feedback about teaching performance without *counselling* in which there are opportunities for *reflection* on teaching and *examples* on how to do it better does not result in measurable improvement (cf. Gray, 1991). Mere counselling without evaluation will not be effective either, because teachers are rarely aware of their own strengths and weaknesses. Studies that compare lecturers' self-assessments with assessments by students and by outside observers show that instructors' views rarely coincide with those of others (colleagues, educationalists, trained former students; $r = .24$ and $r = .06$; Rindermann, 2001). Few lecturers assess quality criteria such as structure and workload adequately. Therefore the perspective of students and external people, such as counsellors, is required.

Counselling not only works, but is preferable for ethical reasons as well. Mere performance measurement without offering assistance for improvement can cause frustration and discouragement for critically assessed instructors. Cranton and Knoop (1991) call this

"professional melancholia" (p. 100). Doyle considers such an approach inhumane and uneconomical:

> I do not think it is humane to open someone to the possibility of a negative evaluation without at the same time providing some meaningful help toward improvement. Business firms know they waste money if they discourage or dismiss potentially productive employees and so they spend large sums of money on intensive training programs. (Doyle, 1991, p. 126)

Similarly Marsh and Roche (1999, p. 517):

> There is an ethically dubious but widespread custom of giving potentially negative feedback to teachers without providing access to cost-effective interventions to assist them to improve their teaching effectiveness. This, perhaps, is the most serious indictment of the current practice.

Universities should be interested in the improvement of teaching quality because of its importance for student learning and engagement (Abrami, d'Apollonia, & Rosenfield, 1997; Hampton & Reiser, 2004). Previous studies suggest this aim is best served through a counselling procedure after evaluation as part of an academic and institutional development programme to enhance the culture of teaching and learning (Hendry & Dean, 2002; Trowler & Bamber, 2005), since evaluation without counselling and acknowledging external conditions is neither effective nor justifiable. However, this study suggests that formally educated pedagogical expertise on the part of the counsellor and the existence of an institutionalised academic development unit may not be as important to achieve this goal as previously thought. But it should also not be forgotten that the rank order (see Figure 2) of the instructors has not changed. While low-rated teachers improved, they are still not in the "good zone" of the possible answer scale (around number 6). A teaching education programme before starting a teaching career (e.g. Piccinin & Moore, 2002)—or during, for instructors with severe problems—could show stronger benefits. Future research should focus on this subgroup.

Acknowledgements

We want to thank all participating students and teachers for their collaboration.

References

Abrami, P. C., d'Apollonia, S., & Rosenfield, S. (1997). The dimensionality of student ratings of instruction: What we know and what we do not. In R. P. Perry & J. C. Smart (Eds.), *Effective teaching in higher education* (pp. 321–367). New York: Agathon Press.

Alean-Kirkpatrick, P., Hänni, H., & Lutz, L. (1997). Internal quality monitoring of the teaching at the ETH, Zürich: Model design and initial impacts. *Quality in Higher Education, 3*(1), 63–71.

Balk, M. (2000). *Die Evaluation von Lehrveranstaltungen.* Frankfurt a.M., Germany: Peter Lang.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Erlbaum.

Cohen, P. A. (1980). Effectiveness of student-rating feedback for improving college instruction: A meta-analysis of findings. *Research in Higher Education, 13*(4), 321–341.

Cranton, P., & Knoop, R. (1991). Incorporating job satisfaction into a model of instructional effectiveness. In M. Theall & J. Franklin (Eds.), *Effective practices for improving teaching (New directions for teaching and learning 48)* (pp. 99–109). San Francisco: Jossey-Bass.

Doyle, K. O. (1991). Report on a trip downtown. In M. Theall & J. Franklin (Eds.), *Effective practices for improving teaching (New directions for teaching and learning 48)* (pp. 123–126). San Francisco: Jossey-Bass.

Dresel, M., Rindermann, H., & Tinsner, K. (2007). Beratung von Lehrenden auf der Grundlage studentischer Veranstaltungsbeurteilungen. In A. Kluge & K. Schüler (Eds.), *Qualitätssicherung und -entwicklung an Hochschulen: Methoden und Ergebnisse.* Lengerich, Germany: Pabst.

Gediga, G., Kannen, K. v., Schnieder, F., Köhne, S., Luck, H., & Schneider, B. (2000). *Kiel: Ein Kommunikations-Instrument für die Evaluation von Lehrveranstaltungen.* Bangor, UK: Methodos.

Gijselaers, W. H. (1988). *Kwaliteit van het onderwijs gemeten.* Doctoral thesis, Limburg, Maastricht.

Gray, P. J. (1991). Using assessment data to improve teaching. In M. Theall & J. Franklin (Eds.), *Effective practices for improving teaching (New directions for teaching and learning 48)* (pp. 53–63). San Francisco: Jossey-Bass.

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*(5), 497–527.

Hendry, G. H., & Dean, S. J. (2002). Accountability, evaluation of teaching and expertise in higher education. *International Journal for Academic Development, 7*(1), 75–82.

Leitner, E. (1998). The pedagogical qualification of the academic teaching staff and the quality of teaching and learning. *Higher Education in Europe, 23*(3), 339–349.

Marsh, H. W. (1987). Students' evaluations of university teaching. *International Journal of Educational Research, 11,* 253–388.

Marsh, H. W., & Roche, L. A. (1999). Reply upon SET research. *American Psychologist, 54*(7), 517–518.

McKeachie, W. J. (1997). Student ratings. *American Psychologist, 52*(11), 1218–1225.

Menges, R. J., & Brinko, K. T. (1986, April). *Effects of student evaluation feedback: A meta-analysis of higher education research.* Paper presented at the 70th meeting of the American Educational Research Association, San Francisco, CA.

Patton, M. Q. (1997). *Utilization-focused evaluation.* Thousand Oaks, CA: Sage.

Penny, A. R., & Coe, R. (2004). Effectiveness of consultation on student ratings feedback: A meta-analysis. *Review of Educational Research, 74*(2), 215–253.

Piccinin, S. (1999). How individual consultation affects teaching. In C. Knapper & S. Piccinin (Eds.), *Using consultants to improve teaching* (*New Directions for Teaching and Learning 79*) (pp. 71–83). San Francisco: Jossey-Bass.

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

Piccinin, S., & Moore, J.-P. (2002). The impact of individual consultation on the teaching of younger versus older faculty. *International Journal for Academic Development, 7,* 123–134.

Rindermann, H. (1996). *Untersuchungen zur Brauchbarkeit studentischer Lehrevaluationen.* Landau, Germany: Empirische Pädagogik.

Rindermann, H. (2001). *Lehrevaluation—Einführung und Überblick zur Forschung und Praxis der Lehrveranstaltungsevaluation an Hochschulen.* Landau, Germany: Empirische Pädagogik.

Rindermann, H. (2002). Evaluation perspectives: An overview of evaluation of communication technologies for education and teaching. In H. H. Adelsberger, B. Collis, & J. M. Pawlowski (Eds.), *Handbook on information technologies for education & training* (pp. 309–329). Berlin, Germany: Springer.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton Mifflin.

Teichmann, S. (1999). Wirksamkeit der Evaluation von Studium und Lehre. In J.-H. Olbertz & P. Pasternack (Eds.), *Profilbildung, Standards, Selbststeuerung* (pp. 241–248). Weinheim, Germany: Deutscher Studienverlag.

Trowler, P., & Bamber, R. (2005). Compulsory higher education teacher training: Joined-up policies, institutional architectures and enhancement cultures. *International Journal for Academic Development, 10,* 79–93.

Webler, W.-D. (1992). Evaluation der Lehre. In D. Grühn & H. Gattwinkel (Eds.), *Evaluation von Lehrveranstaltungen* (pp. 143–161). Berlin, Germany: FU-Dokumentationsreihe.

Webler, W.-D. (1996). Qualitätssicherung in Lehre und Studium an deutschen Hochschulen. *Zeitschrift für Sozialisationsforschung und Erziehungssoziologie, 16*(2), 119–148.

Wilson, R. C. (1986). Improving faculty teaching. Effective use of student evaluations and consultants. *Journal of Higher Education, 57*(2), 196–211.