Higher Education (2006) 52: 687–709
DOI 10.1007/s10734-004-8306-0

Quality assurance in higher education – meta-evaluation of multi-stage evaluation procedures in Germany[1]

L. BORNMANN[1,*], S. MITTAG[2] & H.-D. DANIEL[1, 2]
*[1]Swiss Federal Institute of Technology Zurich (ETH Zurich), Zähringerstr. 24, CH-8092 Zurich; [2]Evaluation Office, University of Zurich; *Author for correspondence, E-mail: bornmann@gess.ethz.ch*

Abstract. Systematic procedures for quality assurance and improvement through evaluation have been in place in Western Europe since the mid 1980s and in Germany since the mid 1990s. As studies in Europe and beyond show that multi-stage evaluation procedures as the main quality assurance instrument for evaluation of teaching and learning in higher education institutions have proved reliable and have gained acceptance, in Germany (as well as in other countries) the evaluation of teaching and learning through internal and external evaluations has long come under the fire of criticism. Our results of the first comprehensive and representative investigation of procedures for the evaluation of teaching and learning in Germany show that former participants in the evaluations (reviewers and those reviewed) are satisfied all in all with the multi-stage procedure. They are convinced that the goals of quality assurance and improvement were achieved. Suggestions for improving the procedures target individual aspects, such as, for example, the composition of the review panel. Against this background, it makes sense to perform regular quality assessments of the procedures for quality assurance and improvement.

Keywords: evaluation of study programmes, external evaluation, external quality assurance, follow-up, internal evaluation, meta-evaluation, multi-stage evaluation, peer review, quality assurance, quality improvement.

Introduction

A core component of higher education reform is systematic quality assurance and improvement of higher education institutions (HEIs). The "Communiqué of the Conference of Ministers Responsible for Higher Education in Berlin on 19 September 2003" establishes that the quality of higher education has "proven to be at the heart of the setting up of a European Higher Education Area" (see http://www.bologna-berlin2003.de/pdf/Communique1.pdf, p. 3). Assuring quality in teaching and learning is no longer a matter only for higher education policy

programmes or broad international professional discussion. Quality development and assurance have long since come to play a central role in strategic higher education planning and in the everyday work of HEIs (Hochschulrektorenkonferenz 2003, p. 5).

Systematic procedures for quality assurance and improvement through evaluation have been in place in Western Europe since the mid 1980s. France took up evaluation of universities (institutional evaluation) in 1984; Finland did so in the early 1990s. The evaluation of programmes of study (programme evaluation) was introduced in the Netherlands, the United Kingdom and Denmark in the late 1980s/early 1990s and in Germany in the mid 1990s.[2] For stocktaking of national evaluation systems, analysis of commonalities and differences and the development of common evaluation models, a number of studies have been conducted by researchers on higher education since the mid 1990s in Europe and beyond (for example, Thune et al. 1995; European Commission/Targeted Socio-Economic Research Program 1998; European Network for Quality Assurance in Higher Education 2002; Brennan and Shah 2000; European Training Foundation 2000; Danish Evaluation Institute 2003).

The experiences collected by the studies show unanimously that in all the countries, multi-stage evaluation procedures as the main quality assurance instrument for evaluation of teaching and learning in HEIs have proved reliable and have gained acceptance. In the multi-stage procedure, academic review begins with internal self-assessment, whereby an academic programme or institute conducts its own analysis of strengths and weaknesses for a self-evaluation report. The next step is external evaluation. Here peer reviewers conduct a site visit of the programmes or units under evaluation and prepare an external evaluation report. The follow-up stage entails implementation of the reviewers' recommendations.[3]

Contrary to the positive resonance from the stock taking studies, the evaluation of teaching and learning through internal and external evaluation has long come under the fire of criticism. There are complaints, e.g. in Germany (as well as in other countries), about the supposed high costs and burdens of evaluation, both financial and personnel costs (Hanft 2003, p. 6) and the lack of consequences following evaluation (Berthold 2002, p. 160). In the face of scarce public funding, it is said that evaluation serves merely to supply political decision-makers with information that is used for cost-cutting purposes and for changing the self-determination of professors into external control (Erche 2003a, pp. 3–4). Others object that the high costs of

evaluation reduce the amount of funding available to universities for teaching and learning (Brinck 2003, p. 12) and that continuous evaluations overburden professors, robbing them of the time that they require for their scientific tasks (Erche 2003b, p. 61).

Despite these and other criticisms that have been raised against multi-stage evaluation of teaching and learning in higher education, there has been no comprehensive investigation of the acceptance and success of this evaluation procedure at German HEIs. Certainly, mutual exchange of experiences among higher education experts and evaluation departments across Europe as well as study programme evaluation spanning different countries in the framework of the studies mentioned above have culminated in a common sense regarding the basics of the evaluation procedure. But until recently, no comprehensive empirical findings have been available on a number of points: (i) what elements of the procedure have proved their worth? (ii) What degree of acceptance has the multi-stage evaluation procedure found among reviewers and among the institutions and programmes reviewed? (iii) Have evaluation procedures achieved the goals intended through their implementation?

Upon this background, the *International Centre for Higher Education Research Kassel* (INCHER-Kassel, Germany, formally *Centre for Research on Higher Education and Work)* initiated a project called *Analysis of procedures and effectiveness of the evaluation processes of the Central Evaluation and Accreditation Agency Hannover (Zentrale Evaluations- und Akkreditierungsagentur Hannover, ''ZEvA'') and the Consortium of Universities in Northern Germany (Verbund Norddeutscher Universitäten, ''VNU'') for the evaluation of teaching and learning*. The study is the first comprehensive and representative investigation of procedures for the evaluation of teaching and learning,[4] and it assesses the two most tried-and-tested and best-known evaluation procedures in Germany from multiple perspectives and using multiple methods.[5] VNU – a consortium of six German universities for evaluation of teaching and studies – and ZEvA – a common agency for Lower Saxony HEIs – have employed multi-stage evaluation procedures since the mid 1990s, and they completed the first evaluation cycle in the year 2001.

Following a description of the design of our study in Section Methods, this paper presents the main findings of the study on (i) overall assessment of the evaluation procedures (usefulness and effectiveness of the procedure, satisfaction with the course of the process, achievement of goals, personal benefit, whether results are commensurate to the effort required), (ii) the question of whether a ranking of the participating universities is feared or desired, (iii) the question of

the extent to which evaluation of teaching and learning should be linked with funding, (iv) the composition of the panel of reviewers, (v) judgment of the reviewers' performance and (vi) analysis of the areas addressed by the reviewers' recommendations.

Methods

The study bases on a mail survey by questionnaire of all former external reviewers and members of institutes[6] who participated in evaluations conducted by ZEvA and VNU. A total of 648 returned questionnaires could be included in the analysis, so that a response rate of 41% was achieved. Table 1 shows the different groups of survey respondents in the total population and the sample. As can be seen from the row percentages in the "Sample" column, the response rates range from 28% (reviewer, same discipline, other country) to 80% (reviewer, other discipline).

As our study included two groups of persons with very different perspectives on the evaluation processes of VNU and ZEvA (reviewers and those reviewed) as shown in Table 1, the results of the statistical analyses are reported below separately for each group, and differences between the assessments of reviewers and members of the institutes are calculated. Differences between reviewers and members of the institutes were assessed with Pearson's χ^2 test statistic. As the usual χ^2 test only says whether two variables may or may not be statistically independent, Pearson's Contingency (C) was additionally used as a measure of association. This coefficient is a numerical index summarizing the strength or degree of relationship in a two-dimensional cross-classification (Hays 1981, p. 558; Cohen 1988, pp. 215–271). A coefficient value of approximately $C = 0.28$ indicates a moderately strong association or, equivalently, differences in assessment of reviewers and institute members with medium effect size (Cohen 1988, p. 227).

We also examined the power of the study to detect significant differences by comparing reviewers with institute members, providing a sample size of approximately 500: if the association is "medium" ($C = 0.28$), the degree of freedom is 1 and the significance criterion is 0.01, the power value is greater than 0.995 (Cohen 1988, p. 228).

In addition to the questionnaire survey, we also conducted 33 interviews and examined them using content analysis. The interview participants were university heads and the authorized agents or contact partners at the different HEIs, the spokesperson of VNU, the scientific

Table 1. Breakdown of reviewers and members of institutes in total population and sample (in absolute and relative frequency) by function

Function	Total population		Sample		
	Absolute	%	Absolute	% (column percentages)	% (row percentages)
Members of institute					
Member of evaluation working group	1059	77	305	63	29
Chairperson of evaluation working group	308	23	180	37	58
Total	1367	100	485	100	
Reviewers					
Same discipline, Germany	200	71	118	76	59
Same discipline, other country	32	11	9	6	28
In professional practice	21	7	12	8	57
Student or graduate	24	9	13	8	54
Other discipline	5	2	4	2	80
Total	282	100	156	100	

Note: Chairperson of evaluation working group is also member of evaluation working group.

Table 2. Number of interviews per participant group

Participant group	Absolute frequencies
University heads ZEvA/VNU (three were universities of applied sciences)	14
Agents authorized to organize evaluations (ZEvA) or contact persons (VNU) in head offices at HEIs	11
Spokesperson of VNU/Scientific director at ZEvA	2
Managing directors of VNU and ZEvA	2
Staff members of ZEvA	3
Staff members of VNU	1
Total	33

director at ZEvA as well as the managing directors and staff members of VNU and ZEvA (see Table 2).

In another step, we conducted content analyses of the reviewers' recommendations in 203 external evaluation reports produced by ZEvA and VNU over a period of about 6 years. The focus of the analyses was to ascertain the number of recommendations in the reports and to find out what areas the recommendations address.

Main findings of the analyses of the evaluation procedures

Overall assessment of the evaluation procedures

To tap overall assessments, institute members and reviewers were asked whether the multi-stage evaluation procedure (internal evaluation, external evaluation and implementation of recommendations) proved to be useful and effective, whether the evaluation process achieved the goals of quality assurance and improvement, whether the results of the evaluation were commensurate to the effort required, whether they were satisfied overall with the process and whether participation in the evaluation process proved worthwhile for them personally.

As shown in Table 3, institute members (83%) and reviewers (96%) state that the multi-stage evaluation procedures for evaluating teaching and learning proved to be useful and effective. The majority of participants are satisfied overall with the evaluations conducted at the individual universities (68% of institute members and 95% of reviewers). The majority of respondents also find that the evaluation process achieved the goals of quality assurance and improvement (65% of

Table 3. Overall assessment of the evaluation process by institute members (I) and reviewers (R) (in absolute and relative frequencies; the assessments are sorted by percentage of institute members)

Assessment of evaluation procedures	Institute members		Reviewers	
	Absolute	%	Absolute	%
Multi-stage procedure (internal evaluation, external evaluation, implementation of recommendations) for evaluating teaching and learning proved useful and effective (I: $n=470$, R: $n=142$)	389	83[a]	136	96[a]
Looking back, I am satisfied overall with the evaluation process (I: $n=475$, R: $n=152$)	324	68[b]	144	95[b]
All in all, the evaluation process achieved the goals of quality assurance and improvement of teaching and learning (I: $n=464$, R: $n=126$)	302	65[c]	117	93[c]
All in all, participating in the evaluation process proved worthwhile for me personally (I: $n=479$, R: $n=151$)	291	61[d]	140	93[d]
The results of the evaluation were commensurate to the effort that was required for the evaluation process (I: $n=467$, R: $n=137$)	212	45[e]	112	82[e]

Notes: [a]χ^2 (1, $n=612$) = 15.1, $p<0.001$, $C=0.16$; [b]χ^2 (1, $n=627$) = 42.8, $p<0.001$, $C=0.25$; [c]χ^2 (1, $n=590$) = 37.1, $p<0.001$, $C=0.24$; [d]χ^2 (1, $n=630$) = 54.3, $p<0.001$, $C=0.28$; [e]χ^2 (1, $n=604$) = 56.3, $p<0.001$, $C=0.29$.

institute members and 93% of reviewers) and that participation in the evaluation proved personally rewarding (61% of institute members and 93% of reviewers). Whereas, 82% of the reviewers see the effort entailed as commensurate to the results of the evaluation, the majority of the institute members (55%) find the effort required to be disproportionate to the results.

The results of the χ^2 tests in Table 3 show that there is a statistically significant difference between the assessments, taken together as a whole, by the institute members and reviewers surveyed. Contingency Coefficients (C) of 0.29 and 0.28 indicate assessment differences between

institute members and reviewers of a medium effect size for the appropriateness of the relation between effort required and results and for participation in the evaluation as personally worthwhile. It is safe to assume that these different assessments reflect the comparatively heavy work load that evaluation puts on institute members. As to overall satisfaction with the evaluation process, achievement of quality assurance and improvement goals through the evaluation and usefulness and effectiveness of the multi-stage procedures for the evaluation of teaching and learning, the assessments by institute members and by reviewers differ only slightly ($C < 0.26$).

The interview participants also assess the evaluation procedures overall as positive. The majority of the interviewees believe that the evaluations were necessary and that they proved useful and effective. They emphasize as particular strengths of the evaluations that in their design they are university-independent, self-critical and well structured and organized. On the other hand, some interview participants offer the criticism that the purpose of the evaluation was clarified insufficiently, that the personnel situation in the institute was taken into insufficient account and that the evaluation did not take a sufficiently international orientation. Several interview participants proposed that evaluation, in addition to teaching and learning, should include further areas (such as internationalization strategies, course guidance and counselling or administration) or focus more strongly on particular areas (such as continuing education opportunities or graduate and post-doctoral programmes).

Quality assurance and the issue of ranking

In line with common practice in European higher education (Danish Evaluation Institute 2003, p. 14), the evaluation procedures of ZEvA and VNU do not aim to produce a *ranking of the participating institutions*. An item on the questionnaire survey of institute members and reviewers asked whether they nonetheless feared in the different phases of the evaluation that their institute would be ranked. Our findings show that 62% of institute members and 80% of reviewers did not gain the impression that the evaluation was aiming towards ranking, χ^2 (1, $n = 612$) = 15.2, $p < .001$, $C = .16$.

We asked both questionnaire respondents and interview participants whether they would have found a ranking of the participating HEIs desirable. The clear majority of interview participants spoke against ranking. Nearly three-quarters (72%) of the survey respondents rejected

ranking. Among the 28% that would have found ranking desirable, institute members (31%) were somewhat more strongly represented than reviewers (19%); χ^2 $(1,\ n = 616) = 8.4,\ p < 0.01,\ C = 0.12$. Overall, our findings as to ranking of HEIs participating in evaluation make it clear that the majority of those involved in evaluations neither fear nor are in favour of ranking.

Quality assurance and the issue of linkage to funding

In the discussion on higher education policies in Germany, some voices have considered linking *the amount of funding* allocated to an institute to *the results of evaluations*. This linkage could result in increases or reductions in funding. On the survey questionnaire, we asked the respondents to assess these proposals. Respondents were asked whether evaluation results should be linked to increases of funding only, to reductions of funding only, to both increases and reductions or whether there should be no linkage to funding at all (see Table 4).

Regarding the linkage issue, institute members (42%) voted somewhat more frequently than reviewers (33%) for the alternative, namely no linkage to funding, while reviewers (46%) were somewhat more

Table 4. Linkage of results of evaluation to funding of institutions or programmes by institute member and reviewer (in absolute and relative frequencies; the assessments are sorted in descending order according to percentages among institute members)

Linkage of evaluation to funding	Institute members		Reviewers	
	Absolute	%	Absolute	%
Evaluation results should never be linked to funding	193	42	47	33
Evaluation results should be linked to increases and reductions of funding	169	37	65	46
Evaluation results should be linked to increases in funding only	95	20	27	19
Evaluation results should be linked to reductions in funding only	3	1	3	2
Total	460	100	142	100

Note: As two cells (25%) in the table have expected frequencies less than 5, the χ^2 test was not computed.

frequently than institute members (37%) in favour of linkage in the case of increases or reductions in funding (as two cells in Table 4 have expected frequencies less than 5, the χ^2 test was not computed). Approximately 20% of both groups favour linkage of evaluation results to funding increases only, and only 1% (institute members) and 2% (reviewers) approve of linkage to funding reductions.

The majority of the interview participants see evaluation results linked to funding decisions in future. Most of these interview participants view both positive and negative sanctions as appropriate. Still, a considerable number of the interview participants think that linkage of evaluation to funding decisions is not appropriate.

Selection of peers for the panel of expert reviewers

ZEvA and VNU follow similar guidelines concerning the composition of the panel of expert reviewers. In both cases, members of the institutes to be evaluated have the right to nominate reviewers in order to ensure acceptance of the panels (see, for example, Zentrale Evaluations- und Akkreditierungsagentur Hannover 2003, p. 10). VNU and ZEvA stipulate that the reviewers must not come from the same German *Länder* in which the HEIs under evaluation are located, that they must be capable of evaluating independently and with no conflict of interest and that they must be respected representatives of their disciplines. In addition, the particular areas of expertise of the reviewers on the panel should mirror the range of areas represented within the discipline being evaluated.

We asked questionnaire respondents about the general practices of VNU and ZEvA regarding panel composition and found that 91% of the institute members and 99% of the reviewers think that the panel of experts was made up of respected scholars in their fields; χ^2 (1, $n = 585$) = 9.6, $p < 0.01$, $C = 0.13$. Moreover, 81% of the institute members and 88% of the reviewers said that the subfields within the disciplines being evaluated were mirrored adequately in the panel of experts; χ^2 (1, $n = 574$) = 4.2, $p < 0.05$, $C = 0.09$. The interview participants found, for one, that the right of the institutes to nominate peers for the panel was a strength of the VNU and ZEvA procedures, because it created the necessary trust in the panel of peer reviewers and ensured its acceptance. For another, however, the interview participants saw a problem in too great proximity between peers and institute members. They pointed out the danger that peer reviewers would not act as "critical assessors", but rather as potential "advocates" of the institutes against university management. In this way,

a kind of "comradeship" could develop.[7] Those interview participants that saw proximity between peers and institute members as problematic wanted to see greater distance between the two groups, with the aim to ensure more clear and more critical evaluations.

As to the persons selected for the panel of peer reviewers, the majority of interview participants are in favour of including on the panel – in addition to peers from the same disciplines within Germany – a student or graduate, a peer from another country, and a peer in professional practice. They were not in favour of including non-professorial scientific staff, because due to their status, they might not judge independently.

We also asked the questionnaire survey respondents for their opinions on the composition of the panel of experts. Table 5 shows the percent of institute members and reviewers who are in favour of including a non-professorial scientific staff person, peer within same discipline from another country, peer in professional practice, student, graduate, expert in higher education, scientist from a non-university research institution, peer from another discipline (a person working in a different field of study) and a representative of a professional association in addition to peers in the same discipline in Germany.

As can be seen in Table 5, both institute members and reviewers are most often in favour of including a representative of the non-professorial scientific staff, a peer in the same discipline from another country, a student, a peer in professional practice, and a graduate. Although two χ^2 tests in Table 5 show statistically significant differences between the assessments of institute members and reviewers (for representative of non-professorial scientific staff and expert in higher education), all contingency coefficients in the table indicate that the differences have only small effect sizes. Members of institutes and reviewers thus have similar opinions as to the groups of persons that should be represented on the panel of expert reviewers in addition to peers within the same discipline in Germany: a representative of the non-professorial scientific staff, a peer within the same discipline in another country, a student or graduate and a peer in professional practice.

Work of the panel of experts

During the site visit, the task of the panel of reviewers is essentially to discuss with the members of the institute the self-evaluated strengths, weaknesses and development potentials as presented in the

Table 5. Percent of institute members (I) and reviewers (R) in favour of including various groups of persons on the panel of reviewers in addition to peers working within the same discipline in Germany (in absolute and relative frequencies; the assessments are sorted in descending order by percentage of members of institutes)

Panel should include	Institute members		Reviewers	
	Absolute	%	Absolute	%
Non-professorial scientific staff (I: $n = 445$, R: $n = 141$)	350	79[a]	94	67[a]
Peers, same discipline, from another country (I: $n = 434$, R: $n = 142$)	312	72[b]	108	76[b]
Students (I: $n = 433$, R: $n = 133$)	293	68[c]	82	62[c]
Peers in professional practice (I: $n = 441$, R: $n = 142$)	300	68[d]	86	61[d]
Graduates (I: $n = 435$, R: $n = 136$)	266	61[e]	73	54[e]
Higher education experts (I: $n = 426$, R: $n = 136$)	246	58[f]	60	44[f]
Scientists from non-university research institutes (I: $n = 421$, R: $n = 134$)	197	47[g]	62	46[g]
Peers, other discipline (persons working in other fields of study) (I: $n = 443$, R: $n = 138$)	190	43[h]	61	44[h]
Representatives of professional associations (I: $n = 422$, R: $n = 134$)	133	32[i]	32	24[i]

Notes: [a]χ^2 (1, $n=586$) = 8.4, $p<0.05$, $C=0.12$; [b]χ^2 (1, $n=576$) = 0.9, $p=0.33$, $C=0.04$; [c]χ^2 (1, $n=566$) = 1.7, $p=0.20$, $C=0.05$; [d]χ^2 (1, $n=583$) = 2.7, $p=0.10$, $C=0.07$; [e]χ^2 (1, $n=571$) = 2.4, $p=0.12$, $C=0.07$; [f]χ^2 (1, $n=562$) = 7.7, $p<0.05$, $C=0.12$; [g]χ^2 (1, $n=555$) = 0.01, $p=0.92$, $C=0.01$; [h]χ^2 (1, $n=581$) = 0.1, $p=0.79$, $C=0.01$; [i]χ^2 (1, $n=556$) = 0.28, $p=0.09$, $C=0.07$.

self-evaluation report and to assess them in the light of the goals of the evaluation. In this way the development process initiated through internal evaluation is to be advanced through discussions, assessments and recommendations. The reviewers' assessment should link back to the goals formulated by the institute, and their recommendations should be geared to implementation. According to the Wissenschaftsrat (1996, p. 26), an advisory body to the federal government and the state (*Länder*) governments in Germany, one of the tasks of the panel of reviewers is to perform a critical evaluation of the self-evaluation by the institute and to

point out inconsistencies in the organization of its study programme. The reviewers should be in a position to detect any strategic behaviour in the self-evaluation report and to check/verify the self-report during the site visit interviews. Further, the reviewers are supposed to gather additional information not contained in the self-evaluation report and to subject the goals set by the institute to critical examination.

All in all, the interview participants give high praise to the work of the panel of reviewers on evaluations at the individual universities. They say that the reviewers were very conscientious about examining the strengths and weaknesses reported in the self-evaluation report during the site visits. We also asked the questionnaire respondents to assess several aspects of the work of the panel of reviewers (see Table 6). The percentages in Table 6 for each aspect show that institute members assess the work of the reviewers somewhat more critically than the reviewers themselves do. For example, 28% of institute members state that the panel of reviewers hardly succeeded in facilitating the development process initiated prior to or during the internal evaluation. This assessment is shared by only one in ten of the reviewers. Twenty-one percent of the institute members report that the panel of reviewers hardly considered objections and suggestions by the institute members when forming their judgement. Only 2% of the reviewers share this opinion. About one-fifth of the institute members say that the panel of reviewers hardly represented the up-to-date state of development of the discipline (6% of reviewers shared this assessment).

Although the differences between the institute members' and the reviewers' assessments are statistically significant with one exception, the contingency coefficients ranging from 0.08 to 0.23 indicate that the effect sizes are small. Statistically speaking, this means that institute members assess the work of the panel of reviewers more critically, but not clearly more critically, than the reviewers themselves do.

Content analysis of the recommendations in the panel of reviewers' evaluation reports

In the framework of the ZEvA and VNU evaluations, after completing site visits the reviewers prepare an initial report outlining recommendations for quality assurance and improvement of teaching and learning. This draft report is submitted to the members of the institute to be checked for misunderstandings. Then the reviewers prepare the final external evaluation report.

Table 6. Assessment of the work of the panel of reviewers by members of institute (I) and reviewers (R) (in absolute and relative frequencies; the assessments are sorted in descending order by percent of members of institute)

Aspect of panel's work	Institute member		Reviewer	
	Absolute	%	Absolute	%
The panel of reviewers hardly facilitated the development process begun prior to or during internal evaluation (I: $n = 437$, R: $n = 142$)	122	28[a]	12	9[a]
The panel of reviewers hardly considered the objections and suggestions by institute members when forming their judgement (I: $n = 415$, R: $n = 146$)	87	21[b]	3	2[b]
The panel of reviewers hardly represented the up-to-date state of development of the discipline (I: $n = 426$, R: $n = 136$)	85	20[c]	8	6[c]
The panel of reviewers' evaluation was in part unfair and biased (I: $n = 439$, R: $n = 145$)	73	17[d]	5	3[d]
The panel of reviewers made hardly any action-oriented recommendations (I: $n = 441$, R: $n = 145$)	58	13[e]	0	0[e]
The panel of reviewers hardly considered the self-formulated goals of the institute in its report (I: $n = 426$, R: $n = 148$)	57	13[f]	10	7[f]
The panel of reviewers hardly confronted the institute members with the self-evaluation in the self-evaluation report (I: $n = 436$, R: $n = 148$)	55	13[g]	9	6[g]
The panel of reviewers was frequently not unanimous in its opinions and recommendations (I: $n = 400$, R: $n = 148$)	45	11[h]	3	2[h]
The recommendations were frequently not supported by all reviewers on the panel (I: $n = 389$, R: $n = 147$)	23	6[i]	3	2[i]

Notes: [a]χ^2 (1, $n = 579$) = 22.8, $p < 0.001$, $C = 0.20$; [b]χ^2 (1, $n = 561$) = 28.7, $p < 0.001$, $C = 0.23$; [c]χ^2 (1, $n = 562$) = 14.8, $p < 0.001$, $C = 0.16$; [d]χ^2 (1, $n = 584$) = 16.4, $p < 0.001$, $C = 0.17$; [e]χ^2 (1, $n = 586$) = 21.2, $p < .001$, $C = 0.19$; [f]χ^2 (1, $n = 574$) = 4.7, $p < 0.05$, $C = 0.09$; [g]χ^2 (1, $n = 584$) = 4.8, $p < 0.05$, $C = 0.09$; [h]χ^2 (1, $n = 548$) = 11.5, $p < 0.01$, $C = 0.16$; [i]χ^2 (1, $n = 536$) = 3.5, $p = 0.06$, $C = 0.08$.

We conducted content analyses of the reviewers' recommendations in the total of 203 external reviewers' reports on completed evaluations of 28 HEIs and 25 disciplines produced by ZEvA and VNU up to the end of 2001. The external evaluation reports differed greatly in structure. In some reports, the reviewers' recommendations are scattered throughout the report, in other reports the recommendations are listed together at the end of the document, and in some reports both methods are used. We also found differences with regard to the formulation of individual recommendations. Whereas some recommendations point up problems and inadequacies that the institutes should merely think about or consider resolving, other recommendations strongly urge the institutes to implement specific changes. Also, in addition to very general recommendations (for example, ''optimize the course of teaching and learning''), the reports also contain very specific recommendations (for example, ''offer students more international exchange programmes'').

Similar differences among the recommendations in the reports were found by other investigators. Brennan et al. (1996) concluded: ''Many recommendations are far from straightforward. Many do not indicate a course of action but rather a problem to be investigated or an issue to be reviewed'' (p. 60). One year later Brennan et al. (1997) found: ''[the recommendations] frequently indicated areas of concern, matters to be attended to, problems requiring solution, but avoided the specification of prescribed solutions, seeing these as issues for institutional determination. It follows, therefore, that the implementation of a recommendation will rarely be clear-cut'' (p. 11). Frederiks et al. (1994) evaluated the evaluation system in the Netherlands and found: ''factors that contribute to a negative attitude are that recommendations are often not very precise, reports are sometimes inconsistent'' (p. 195).

The 203 external evaluation reports produced by ZEvA and VNU up to 2001 contain a total of 3452 recommendations, or on average, 17 recommendations per report. In the framework of content analysis, we assigned the recommendations to a total of 11 categories, such as ''planning and organization of teaching and learning'', ''resources'', ''examinations'' and ''forms of teaching and learning''.[8] Table 7 shows the categories and examples of recommendations that were assigned to those categories (see columns 1 and 2). For example, we assigned the recommendation ''optimize structure and sequential course of study programme'' to the category ''planning and organization of teaching and learning''.

Table 7. Frequencies of recommendations by area addressed (in absolute and relative frequencies, $n = 3452$)

Category	Example recommendations in this category	Frequencies	
		Absolute	%
Planning and organization of teaching and learning	Optimize structure and sequential course of study programme, increase or optimize interdisciplinary teaching and learning	1220	35
Resources	Expand scientific staff, increase number of computer work stations	568	16
Examinations	Update/optimize examinations regulations, improve performance control, distribute examination load more equally among teaching staff	258	8
Forms of teaching and learning	Optimize practicum, introduce/ optimize new forms of learning, increase use of new media in teaching	303	9
Course content	Increase relevance and up-to-datedness in teaching, maintain or further develop existing course contents, define course-of-study relevant course contents	263	8
Student guidance and counselling	Improve course guidance and counselling of students, offer and improve introductory courses, publish or improve commentated course prospectus	291	8
Promotion of young academics and scientists	Foster young academics and scientists, create more positions for doctoral and post-doctoral qualifying	145	4
Positioning and development of differentiated profile	Improve positioning and Public Relations work, build and reinforce university or institute differentiating strengths, strengthen positioning of the institute	141	4

Table 7. (Continued)

Category	Example recommendations in this category	Frequencies	
		Absolute	%
Quality assurance and improvement of teaching and learning	Implement and improve regularly instruments for quality improvement of teaching and learning, institutionalise discourse on teaching and learning, introduce or optimize control of teaching load	126	4
Administration and academic self-government	Optimize internal allocation of funds, improve planning and administrative processes in teaching and learning, optimize structures for self-government	68	2
Goals for teaching and learning	Improve computer literacy education, teach (improve) interdisciplinary and transdisciplinary qualifications, assure/maintain the range of the study programme	69	2

Table 7 presents the frequencies of the recommendations per category. Approximately one-half (51%) of all recommendations address only two areas, namely, ''planning and organization of teaching and learning'' (35%) and ''resources'' (16%). The remaining 49% of the recommendations fall under the other nine categories. Only 2% of the recommendations address ''goals for teaching and learning'' and ''administration and academic self-government''.

Findings of content analyses of reviewers' recommendations are available from other European countries as well. Hulpiau and Waeytens (2003) examined external evaluation reports at a Flemish university and found that the reviewers most frequently find fault with pedagogical problems (15%), followed by organizational (8%) and educational conditions (5%). In the assessment of teaching quality in HEIs in England and Northern Ireland – undertaken by the Quality Assessment Division of the Higher Education Funding Council for England (HEFCE) – 90% of the assessors' recommendations (in all, 287 assessment reports contained 1806 recommendations) are related to six aspects: ''Curriculum Design, Content and Organisation; Teaching, Learning

and Assessment; Student Progression and Achievement; Student Support and Guidance; Learning Resources; Quality Assurance and Enhancement ... A few recommendations fell outside the six aspects and most of these were coded under a generic heading of Organisational Context and Policy'' (Brennan et al. 1996, p. 10).

Conclusions

The findings of the previous studies on evaluation of teaching and learning mentioned in Section Introduction confirmed that multi-stage evaluation procedures – with internal evaluation, external evaluation and follow-up – are useful and effective. Our analysis of the evaluation processes of ZEvA and VNU also demonstrate that former participants in the evaluations (reviewers and those reviewed) are satisfied all in all with the multi-stage procedure and believe that the goals of quality assurance and improvement were achieved.[9] Somewhat more than one-half of the members of institutes that we surveyed by questionnaire, however, find that the results of the evaluations do not justify the heavy work burden that the process entails. With other words, concerns remain among a substantial proportion of institute members about the cost-benefit value of the evaluation process.

The majority of the questionnaire respondents and interview participants give high recognition to the evaluation work of the reviewers at the various HEIs. However, when asked about specific aspects of the reviewers' performance, the members of institutes express more criticism of the reviewers' work than the reviewers themselves do.

Concerning the composition of the panel of external reviewers for evaluating teaching and learning, we can derive some general recommendations from the findings of our study. The panel of reviewers should include an expert from another country, an expert working in professional practice, a student or graduate and possibly a representative of the non-professorial scientific staff. Although the German Rectors' Conference (Hochschulrektorenkonferenz 2000, p. 17) does not recommend inclusion of students on the review committee, our findings show that the majority of the people that we surveyed hold a different opinion. They feel that including a student on the review panel increases students' acceptance of the evaluation process.[10]

The 203 external evaluation reports produced in the framework of ZEvA and VNU evaluations contain a total of 3452 reviewers' recommendations (on average, 17 per report). Content analysis of the

recommendations shows clearly that the reviewers' recommendations are very unevenly distributed across the various areas. A large part of the recommendations address the areas "planning and organization of teaching and learning" and "resources", while recommendations considering "goals for teaching and learning" are much more rare. A possible explanation for the latter finding is that reviewers do not find inadequacies in that area. But it is equally possible that reviewers do not ascribe as much importance to this area as they do to other areas. Perhaps it is also more of a challenge to formulate action-oriented, implementable recommendations in that area. Whether the recommendations have been acted upon is a question of a follow-up survey; results will be published in 2006.

To sum up, contrary to the diverse criticisms that are still being raised against evaluations of teaching and learning (see Section Introduction), the findings of our analysis confirm that multi-stage evaluation procedures in Germany enjoy wide acceptance and are seen to be useful and effective. Suggestions for improving the procedures target individual aspects, such as, for example, the composition of the review panel. Upon this background, then, it makes sense to perform regular quality assessments of the procedures for quality assurance and improvement: "One thing is for sure, one never can say that a system for external quality assessment is finished. It always can be improved and it has always to be changed for keeping the academic world alert and to prevent quality assessment becoming a ritual dance" (Vroeijenstijn 2000, p. 66; see also Vroeijenstijn 1995, p. 38; Wissenschaftsrat 1996, p. 37; European Training Foundation 2000, p. 24).

Notes

1. The authors wish to express their gratitude to the Donors' Association for the Promotion of Sciences and Humanities in Germany (Stifterverband für die Deutsche Wissenschaft) for funding our study and two anonymous reviewers for their helpful comments.
2. Overviews of evaluation procedures in different countries are provided by Anderson et al. (2000), Brennan (2001), Hämäläinen et al. (2001), Organisation for Economic Cooperation and Development (2003), Reichert and Tauch (2003) and Thune (1998).
3. Despite agreement on the general course of proceeding, national quality assurance systems differ greatly in the details (Brennan and Shah 2000, pp. 50–69; Billing 2004).
4. The project has so far delivered the following publications (in German): Mittag et al. (2003a), Mittag et al. (2003b), Bornmann et al. (2003) and Daniel et al. (2003).

5. On the evaluation processes implemented by VNU and ZEvA, see Verbund Norddeutscher Universitäten (1999, 2004), Zentrale Evaluations- und Akkreditierungsagentur Hannover (2003) and Hochschulrektorenkonferenz (1998). In all, there are eight evaluation agencies in Germany for systematic assessment of teaching and learning at HEIs.
6. 'Member of institute' refers to all members of an evaluation working group that formed within an institute to be evaluated (including students).
7. On "comradeship" in the peer review process (also called "nepotism", "patronage" or "old boys' network"), see overviews of the literature provided by Bornmann and Daniel (2003) and Cole (1992).
8. Of the 3452 recommendations, 32 were assigned to two categories each instead of one. For example, we assigned the recommendation "create more positions for doctoral and post-doctoral qualifying" to two categories, "promotion of young academics and scientists" and "resources" (sub-category "personnel resources").
9. The authors of this article agree with one of the reviewers that acceptance by the academic profession is a necessary condition for effective evaluations given that professors are the primary instrument for assuring academic quality, academics' satisfaction with the evaluation process is not sufficient. One can certainly argue that ultimately the test of the impact of evaluations of teaching and learning is whether academic standards are strengthened and whether teaching and student learning are improved.
10. Sweden's recently reformed quality assurance system, which has been in implementation since 2002, provides for a student member on the review panel for evaluations of study programmes (Franke 2002, p. 26). In contrast, ZEvA does not include a student on the review panel for follow-up evaluations.

References

Anderson, D.S., Johnson, R. and Milligan, B. (2000). *Quality Assurance and Accreditation in Australian Higher Education: An Assessment of Australian and International Practice*. Canberra: Department of Education, Training and Youth Affairs.

Berthold, C. (2002). 'Von der Evaluation zur strategischen Hochschulentwicklung – 16 Thesen', in Reil, T. and Winter, M. (eds.), *Qualitätssicherung an Hochschulen: Theorie und Praxis*. Bonn: Hochschulrektorenkonferenz (HRK), pp. 160–165.

Billing, G. (2004). 'International comparisons and trends in external quality assurance of higher education: Commonality or diversity?', *Higher Education* 47(1), 113–137.

Bornmann, L. and Daniel, H.-D. (2003). 'Begutachtung durch Fachkollegen in der Wissenschaft. Stand der Forschung zur Reliabilität, Fairness und Validität des Peer-Review-Verfahrens', in Schwarz, S. and Teichler, U. (eds.), *Universität auf dem Prüfstand. Konzepte und Befunde der Hochschulforschung*. Frankfurt: Campus, pp. 211–230.

Bornmann, L., Mittag, S. and Daniel, H.-D. (2003). *Qualitätssicherung an Hochschulen. Empfehlungen zur Durchführung mehrstufiger Evaluationsverfahren in Studium und Lehre* (Reihe Positionen des Stifterverbandes für die Deutsche Wissenschaft). Bonn: Stifterverband für die Deutsche Wissenschaft (ed.).

Brennan, J. (2001). 'Quality management, power and values in european higher education', *Higher Education: Handbook of Theory and Research* 16, 119–145.

Brennan, J. and Shah, T. (2000). *Managing Quality in Higher Education – an International Perspective on Institutional Assessment and Change*. Buckingham: Organisation for Economic Cooperation and Development (OECD).

Brennan, J., Frederiks, M.M.H. and Shah, T. (1997). *Improving the Quality of Education: The Impact of Quality Assessment on Institutions*. London: Higher Education Funding Council for England (HEFCE).

Brennan, J., Shah, T. and Williams, R. (1996). *Quality Assessment and Quality Improvement: An Analysis of the Recommendations Made by HEFCE Assessors*. London: Higher Education Funding Council for England (HEFCE).

Brinck, C. (2003, November 7). 'Euch machen wir mürbe. Hochschulkontrolle: Aufzeichnungen eines Nichtakkreditierten'. *Frankfurter Allgemeine Zeitung*, p. 12.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd Edition). Hillsdale: Lawrence Erlbaum Associates.

Cole, S. (1992). *Making Science. Between Nature and Society*. Cambridge: Harvard University Press.

Daniel, H.-D., Mittag, S. and Bornmann, L. (2003). 'Mehrstufige Evaluationsverfahren für Studium und Lehre. Empfehlungen zur Durchführung', in Berendt, B., Voss, H.-P. and Wildt, J. (eds.), *Neues Handbuch Hochschullehre*. Stuttgart: Raabe, I 2.2.

Danish Evaluation Institute (ed.) (2003). *Quality Procedures in European Higher Education. An ENQA Survey* (ENQA Occasional Papers 5). Helsinki: European Network for Quality Assurance in Higher Education (ENQA).

Erche, B. (2003a). 'Evaluation als politisches Steuerungsinstrument', in Neuser, J. and Urban, R. (eds.), *Evaluation der universitären Lehre in der Medizin. Gegenstände – Methoden – Konsequenzen*. Aachen: Shaker Verlag, pp. 3–7.

Erche, B. (2003b, January 30). 'Evaluation der Evaluation und so weiter. Universitätssysteme im Stress'. *Neue Zürcher Zeitung*, p. 61.

European Commission/ Targeted Socio-Economic Research Program (ed.) (1998). *EVALUE – Evaluation and Self-Evaluation of Universities in Europe*. Luxemburg: European Commission.

European Network for Quality Assurance in Higher Education (ed.) (2002). *Trans-National European Evaluation Project (TEEP 2002). Evaluation Manual*. Helsinki: European Network for Quality Assurance in Higher Education (ENQA).

European Training Foundation (ed.) (2000). *The European University: A Handbook on Institutional Approaches to Strategic Management, Quality Management, European Policy and Academic Recognition*. Turin: European Training Foundation.

Franke, S. (2002). 'From audit to assessment: A national perspective on an international issue', *Quality in Higher Education* 8(1), 23–28.

Frederiks, M.M.H., Westerheijden, D.F. and Weusthof, P.J.M. (1994). 'Effects of quality assessment in Dutch higher education', *European Journal of Education* 29(2), 181–199.

Hämäläinen, K., Pehu-Voima, S. and Wahlén, S. (2001). *Institutional Evaluations in Europe* (ENQA Workshop Reports 1). Helsinki: European Network for Quality Assurance in Higher Education (ENQA).

Hanft, A. (2003). *Evaluation und Organisationsentwicklung* (EvaNet-Positionen 10/2003). Retrieved March 26 2004, from http://evanet.his.de/evanet/forum/pdf-position/HanftPosition.pdf.

Hays, W.L. (1981). *Statistics* (3rd. Edition). New York: Holt, Rinehart & Winston.

Hochschulrektorenkonferenz (ed.) (1998). *Evaluation. State of the Art Report on Quality Assessment and Quality Development in German Universities* (Dokumente & Infor-

mationen, 1/1998). Bonn: Hochschulrektorenkonferenz (HRK), Projekt Qualitätssicherung.

Hochschulrektorenkonferenz (ed.) (2000). *Wegweiser 2000 durch die Qualitätssicherung in Lehre und Studium* (Dokumente & Informationen, 2/2000). Bonn: Hochschulrektorenkonferenz (HRK), Projekt Qualitätssicherung.

Hochschulrektorenkonferenz (ed.) (2003). *Wegweiser 2003. Qualitätssicherung an Hochschulen. Sachstandsbericht und Ergebnisse einer Umfrage des Projektes Qualitätssicherung* (Beiträge zur Hochschulpolitik, 7/2003). Bonn: Hochschulrektorenkonferenz (HRK), Projekt Qualitätssicherung.

Hulpiau, V. and Waeytens, K. (2003). 'Improving the quality of education. What makes it actually work? A case study.' In Prichard, C. and Trowler, P. (eds.), *Realizing Qualitative Research in Higher Education* (Cardiff papers in qualitative research) (pp. 145–169). Aldershot: Ashgate.

Mittag, S., Bornmann, L. and Daniel, H.-D. (2003a). *Evaluation von Studium und Lehre an Hochschulen – Handbuch zur Durchführung mehrstufiger Evaluationsverfahren*. Münster: Waxmann.

Mittag, S., Bornmann, L. and Daniel, H.-D. (2003b). 'Mehrstufige Verfahren für die Evaluation von Studium und Lehre – Eine Zwischenbilanz europäischer Erfahrungen', in Schwarz, S. and Teichler, U. (ed.), *Universität auf dem Prüfstand. Konzepte und Befunde der Hochschulforschung*. Frankfurt: Campus, pp. 187–210.

Organisation for Economic Cooperation and Development (2003). *Education Policy Analysis*. Buckingham: Organisation for Economic Cooperation and Development (OECD).

Reichert, S. and Tauch, C. (2003). *Trends 2003. Progress towards the European Higher Education Area. Bologna four Years after: Steps toward sustainable Reform of Higher Education in Europe*. Brussels: European University Association (EUA).

Thune, C. (1998). *Evaluation of European Higher Education: A Status Report, prepared for the European Commission DG XXII by the Centre for Quality Assurance and Evaluation in Higher Education, Denmark, in Cooperation with Comité National d'Evaluation*. Retrieved December 20 2003, from http://www.enqa.net/docs.lasso?docname=statusreport1.html.

Thune, C., Staropoli, A., Kristoffersen, D., Ottenwaelter, M.-O. and Sursock, A. (1995). *European Report: Final Report for the European Pilot Project*. Brussels: European Commission.

Verbund Norddeutscher Universitäten (ed.) (1999). *Evaluation von Studium und Lehre im Fach Geowissenschaften* (Verbund-Materialien, Band 4). Hamburg: Verbund Norddeutscher Universitäten.

Verbund Norddeutscher Universitäten (ed.) (2004). *Evaluation von Studienfächern – ein Beitrag zur Qualitätssicherung. Projektplan für den zweiten Zyklus ab Frühjahr 2004*. Hamburg: Verbund Norddeutscher Universitäten.

Vroeijenstijn, A.I. (1995). *Improvement and Accountability: Navigating between Scylla and Charybdis. Guide for External Quality Assessment in Higher Education* (Higher Education Policy Series, 30). Melksham, Wiltshire: Cromwell Press.

Vroeijenstijn, A.I. (2000). 'External Quality Assessment (EQA) in the Netherlands. The 3rd Generation 2000–2006', in Hochschulrektorenkonferenz (ed.), *Leitbild der Hochschule – Qualität der Lehre* (Beiträge zur Hochschulpolitik; 2/2000). Bonn: Hochschulrektorenkonferenz (HRK), Projekt Qualitätssicherung, pp. 53–66.

Wissenschaftsrat (ed.) (1996). *Empfehlungen zur Stärkung der Lehre durch Evaluation*. Berlin: Wissenschaftsrat (WR).

Zentrale Evaluations- und Akkreditierungsagentur Hannover (ed.) (2003). *Qualitätssicherung in Lehre und Studium. Erst- und Folgeevaluationen sowie Akkreditierungen* (Handbuch zur Qualitätssicherung in Lehre und Studium, Schriftenreihe "Lehre an Hochschulen", 33/03). Hannover: Zentrale Evaluations- und Akkreditierungsagentur Hannover (ZEvA).