Anesthesiology 2009; 111:709–16

Development of a System for the Evaluation of the Teaching Qualities of Anesthesiology Faculty

Kiki M. J. M. H. Lombarts, Ph.D., M.H.A., Martin J. L. Bucx, M.D., Ph.D.,† Onyebuchi A. Arah, M.D., Ph.D.‡*

GRADUATE medical education is one of the core missions of academic medical centers, wherein medical specialists are responsible for teaching and supervising their future colleagues. However, being a medical specialist is no longer a sufficient qualification or proxy for competence in medical education aimed at training residents. This is particularly true given the modernization requirements for competency-based teaching and training promoted by accreditation institutions in some countries (such as the Accreditation for Council for Graduate Medical Education in the United States).[1] These modernization efforts accelerate faculty development of clinician-educators needed to achieve and maintain the highest standard of postgraduate medical education. An effective faculty development track should include measuring medical teaching effectiveness. This requires valid and reliable instruments, as well as providing the findings in a clear and concise format to faculty. Several studies have found that systematic and constructive feedback can result in improved teaching.[2] There are few published and validated evaluation systems or even instruments aimed at supporting the graduate medical education qualities of clinical faculty. In anesthesiology, there are few published instruments and systems,[3] and the existing ones tend to focus on faculty evaluation by residents only without any self-evaluations by faculty. To ensure actual behavioral change, individuals must usually undergo a stepwise change process. Evaluation insights obtained from feedback should be followed by creating positive intentions to change, trying out new behaviors and integrating them into practice. Supporting this change process has been shown to be effective.[2,4]

To support the specialty-specific evaluation of teaching qualities of anesthesiology faculty in an academic medical center, we developed the System for Evaluation of Teaching Qualities (SETQ) comprising (1) a Web-based self-evaluation by faculty, (2) a Web-based residents' evaluation of faculty, (3) individualized faculty feedback, and (4) individualized faculty follow-up support. This paper has three main objectives: (1) to investigate the psychometric properties of the two instruments underlying the SETQ system, (2) to explore the relationship between residents' evaluation and faculty self-evaluation, and (3) to gauge the feasibility of reliably using residents' evaluation of faculty by estimating the number of such evaluations needed per faculty. We also place these objectives in context by describing SETQ.

SETQ was initially developed in the anesthesiology department of a large academic medical center that has over 7,000 staff (including about 500 faculty and 400 residents) in the Netherlands. It was later expanded to include specialty-specific modules for internal medicine, surgery, and obstetrics and gynecology. At the time of writing, most of the remaining specialties have signed up for SETQ, resulting in more than 90% faculty coverage in 2009. SETQ is receiving nationwide attention.

Materials and Methods

SETQ of Anesthesiology Faculty

Figure 1 provides an overview of the SETQ system for evaluating teaching qualities of anesthesiology faculty. It was conceived as a three-stage, individualized measurement and improvement system. The first stage involved measurements using (1) a Web-based self-evaluation instrument filled in by faculty and (2) another Web-based instrument for evaluation of faculty by residents. The second stage involved individualized faculty feedback in which each participating faculty received detailed reports of the outcomes of the residents' evaluations and, if available, self evaluations, also graphed within the context of the averaged outcomes of their colleagues. The third stage involved individualized faculty follow-up with the aim of discussing the results and finding avenues for improvement, if needed, with each individual faculty and head of department. The research team also provided anonymous overall feedback averaged over all participants to the entire departmental faculty and residents in medical teaching seminars.

Study Population and Setting

Data collection took place in the month of September 2008, when 33 residents who had been in training for at

* Senior Researcher, Department of Quality and Process Innovation, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands; † Anesthesiologist, Department of Anesthesiology, Radboud University Nijmegen Medical Center, Nijmegen, The Netherlands; ‡ Associate Professor, Department of Epidemiology, University of California, Los Angeles (UCLA), School of Public Health, Los Angeles, California.

Received from the Department of Quality and Process Innovation, Academic Medical Center, Amsterdam, The Netherlands. Submitted for publication March 17, 2009. Accepted for publication June 25, 2009. Support was provided solely from institutional and/or departmental sources. The Faculty of Health, Medicine and Life Sciences, Maastricht University, Maastricht, The Netherlands, contributed financially to the development of the System for Evaluation of Teaching Qualities Web application.

Address correspondence to Dr. Lombarts: Department of Quality and Process Innovation, Academic Medical Center, Meibergdreef 9, PO Box 22700, 1100 DE Amsterdam, The Netherlands. m.j.lombarts@amc.uva.nl. Information on purchasing reprints may be found at www.anesthesiology.org or on the masthead page at the beginning of this issue. ANESTHESIOLOGY's articles are made freely accessible to all readers, for personal use only, 6 months from the cover date of the issue.

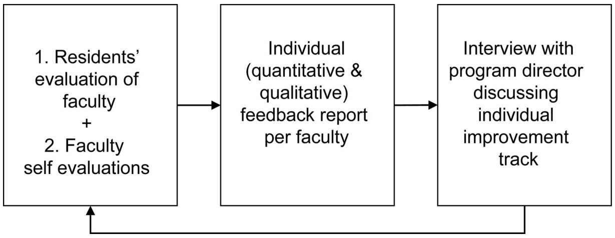
1. Residents' evaluation of faculty
+
2. Faculty self evaluations
Individual (quantitative & qualitative) feedback report per faculty
Interview with program director discussing individual improvement track

Fig. 1. System for Evaluation of Teaching Qualities (SETQ) measurement and improvement system.

least 6 months and 39 faculty in the anesthesiology department were invited *via* e-mail to participate in the evaluations. The invitation assured the formative purpose and use of the evaluations. Participation of faculty and residents remained confidential. Residents' evaluations were anonymous; only the number of residents' evaluations was reported back to the individual faculty members. Each faculty was invited to share and discuss their feedback results with the head of department, but this was not mandatory. The two evaluation instruments were made available electronically *via* a dedicated password-protected SETQ Web portal. Residents chose who to evaluate and could evaluate many faculty. Each faculty could only self-evaluate. Automatic e-mail reminders and the head of department at clinical meetings encouraged both faculty and residents to participate in the evaluation.

Two Instruments: One for Self-Evaluation and Another for Residents' Evaluation of Faculty

Both the self-evaluation and the residents' evaluation instruments were based on the well-known 26-item Stanford Faculty Development Program (SFDP26) instrument, which was developed in the United States.[5–9] It is based on educational and psychological theories of learning and empirical observations of clinical teaching.

In an earlier smaller study, we developed and pilot-tested the SFDP26 instrument for evaluation of anesthesiology faculty in an academic medical center outside the United States. A taskforce of anesthesiology faculty and residents drafted a questionnaire by translating the SFDP26 questionnaire and discussing its completeness, feasibility, and validity for a Dutch residency program. Consensus was reached. After its discussion in separate meetings of anesthesiology residents and faculty, the questionnaire was further edited, tested, and evaluated. We tentatively concluded that the adapted SFDP26 instrument completed by residents could yield reliable and valid evaluation of anesthesiology faculty in an academic medical center outside the United States.[10]

Both the self-evaluation and residents' evaluation instruments shared 24 core items spanning 5 domains of teaching quality, namely learning climate (8 items), professional attitude towards residents (4 items), communication of goals (4 items), evaluation (4 items), and feedback (4 items). Each of the 24 items had a 5-point Likert-type response scale: strongly disagree, disagree, neutral, agree, strongly agree. Each instrument concluded with two global rating items measuring "faculty being seen as a role model" and "faculty's overall teaching qualities," respectively. The global rating "faculty being seen as a role model" had the same response scale as the core items. For the global rating "faculty's overall teaching qualities," the 5-point Likert-type response was 1 = bad, 2 = fair, 3 = average, 4 = good, and 5 = excellent. In addition, the resident instrument had two open questions for narrative feedback on faculty, listing the strong teaching qualities of individual faculty and formulating concrete suggestions for improvement. We also collected data on residents' sex and year of training. For faculty, we collected data on age, sex, number of years in practice since registration as an anesthesiologist, actual time spent on teaching residents, and previous participation in a training program for clinician-educators.

Analytical Strategies

We carried out four main types of analysis. First, we estimated the descriptive statistics (means, proportions) for the response sample to understand the basic characteristics of the participating residents and faculty.

Second, to address the first objective of this study, that is, the psychometric properties of the SETQ instruments for both residents and faculty, we conducted exploratory factor, reliability coefficient, item-total scale correlation, interscale correlation, and scale *versus* global ratings correlation analyses.[11,12] For item reduction or multifactorial structuring of the instruments, we conducted exploratory factor analysis by using the principal components technique with oblique rotation to explore the factor or scale structure of both instruments separately. On the basis of the foregoing results, we calculated the internal consistency reliability coefficient or Cronbach's α for each scale.[13] A Cronbach's α of at least 0.70 was considered satisfactory.[14,15] In addition, we used the residents' instrument to estimate faculty-level reliability coefficients of the intraclass correlation type based on the variance components for each scale.[11] Although no test-retest reliability was conducted in the current study, it was expected that any finding of high levels of interrater reliability would suggest that the intraobserver reliability, hence test-retest reliability, can only be higher.[11] Item-total scale correlations, corrected for item overlap, were used to check for the homogeneity of the scales based on averaging items that loaded strongly on the scales.[11] Furthermore, interscale correlations for residents and faculty separately were used to check for the interpretability of the constructed scales as distinct domains of a related overall construct. An interscale correlation of less than 0.70 was seen as satisfactory and gave credibility to the multidimensional factor or scale structure of the instruments.[12,16] To explore the construct validity of the instruments,[11] the scales were finally correlated with the two global ratings, "faculty being seen as a role model" and "faculty's overall teaching qualities." This approach was an imperfect, opportunistic construct validation testing using global rating items embedded within the instruments for construct hypothesis testing. We emphasize that this construct validation approach was not aimed at providing a final answer but at yielding initial results in an ongoing and cumulative exercise to be improved upon in subsequent research, as is increasingly acknowledged in the modern psychometrics literature.[11] Conceivably, an endless number of related hypotheses could be coined and tested for parts of the instruments over time. We hypothesized or made the informed assumption that faculty who score high on the items/scales should score highly on being seen as a role model and on the singular measure of their overall teaching qualities.[17] In line with the literature, we expected appropriate correlations between the scales and global ratings to fall within the range of 0.40 to 0.80.[11]

Third, to investigate our second objective of exploring the relationship between residents' assessments and faculty self-evaluations, we estimated the mean and SEM of each scale and their related items. Kendall's rank order correlation coefficient τ[18] was used to gauge the correlations between the faculty's teaching qualities rankings based on residents' assessments *versus* those based on faculty self-evaluations. There is no generally accepted cut-point for high rank order correlation; for this study, therefore, the higher the correlation, the better.

Fourth, this study's final objective of investigating the feasibility of reliably using residents' evaluation was analyzed by estimating the number of per-resident evaluations of faculty. This estimation involved solving the equation for the aforementioned reliability coefficient of the intraclass correlation type (using variance components) to determine the number of resident evaluations needed per faculty at any predefined reliability level.[11,19,20] We further triangulated the estimates of the number needed obtained above from solving the variance partitioning of the cross-classified multilevel model equation as follows. It was assumed that for each scale or instrument, the ratio of the sample size (N) to the reliability coefficient (R) would be approximately constant across combinations of sample size and associated reliability coefficients.[11,21] Therefore, the number of residents' evaluations needed (N_{new}) divided by the needed reliability coefficient (R_{new}) would be equal to the observed number of residents' evaluations per faculty (N_{old}) divided by the observed reliability coefficient R_{old}. We already knew N_{old} and R_{old} and could assume different target values for R_{new}; therefore, we easily estimated N_{new} from the assumed equality $N_{\text{new}}/R_{\text{new}} = N_{\text{old}}/R_{\text{old}}$. We repeated the calculations for reliability coefficients (R_{new}) of 0.60, 0.70, 0.80, and 0.90. Reassuringly, both the first but complex and second but simple methods gave similar results within an error margin of no more than ±1. The results of the first method are reported here.

Statistical significance was set at $P < 0.05$ (two-tailed). All analyses were conducted by using the general purpose statistical software SPSS version 16.0.2 (SPSS Inc., Chicago, IL) and Microsoft Office Excel 2003 SP3 (Microsoft Corporation, Redmond, WA).

Table 1. Characteristics of Residents and Faculty who Participated in the Evaluations

	Residents	Faculty
Number invited	33	39
Number of respondents (%)	30 (91%)	36 (92%)
Percentage respondents who are female	66%	33%
Total number of residents' evaluations of faculty or faculty's self-evaluation	611	36
Mean number of evaluations per resident	20.4	n.a.
Mean number of evaluations per faculty member	n.a.	15.7
Percentage of residents per year of residency training		
First year	5.7%	n.a.
Second year	11.5%	n.a.
Third year	27.0%	n.a.
Fourth year	34.0%	n.a.
Fifth year	21.8%	n.a.
Mean number years of practice since first specialist registration as anesthesiologist (SD)	n.a.	12.7 (9.2)
Percentage of faculty who had formal training as educators	n.a.	19%
Percentage of faculty who spend		
Less than 10% of their time on teaching	n.a.	11.1%
From 10% to less than 20% of their time on teaching	n.a.	50.0%
From 20% to less than 30% of their time on teaching	n.a.	25.0%
From 30% to less than 40% of their time on teaching	n.a.	13.9%

n.a. = not applicable.

Results

Study Participants

There were 30 residents and 36 anesthesiology faculty who participated in the study, yielding response rates of 91% and 92%, respectively (table 1). Two-thirds of residents and one-third of faculty participants were female. Residents from all but the last year of training were represented. Residents completed a total of 611 evaluations. There were about 20 evaluations per resident and nearly 16 evaluations per faculty member. Faculty reported being registered anesthesiologists for a mean of 12.7 yr. About 19% of faculty reported having enjoyed a formal training for clinician-educators. The actual time spent on teaching varied substantially among faculty members. Table 1 gives an overview of participant characteristics.

Reliability and Validity of the SETQ Instruments

Explorative factor analysis yielded five teaching domains or scales for both instruments: learning climate, professional attitude towards residents, communication

Table 2. Item and Scale Characteristics, Internal Consistency Reliability, and Item-total Correlations

		Factor Loadings on Primary Scale		Internal Consistency Reliability,† Cronbach's α		Corrected Item-total Correlations	
Item Number	Scale and Items*	Residents	Faculty	Residents	Faculty	Residents	Faculty
Learning climate				0.90 (0.87)	0.57		
Q01	Encourages residents to participate actively in discussions	0.71	0.57			0.74	0.46
Q02	Stimulates residents to bring up problems	0.63	0.38			0.74	0.70
Q03	Teaches residents time management	0.22	−0.27			0.55	−0.12
Q04	Keeps to teaching goals; avoids digressions	0.20	−0.49			0.59	0.18
Q05	Motivates residents to study further	0.71	0.72			0.79	0.43
Q06	Stimulates residents to keep up with the literature	0.74	0.48			0.74	0.36
Q07	Prepares well for teaching presentations and talks	0.77	0.42			0.72	0.30
Q08	Teaches postoperative care in recovery room	0.45	0.14			0.61	0.14
Professional attitude towards residents				0.89 (0.86)	0.73		
Q09	Listens attentively to residents	0.74	0.78			0.76	0.68
Q10	Is respectful towards residents	0.85	0.78			0.77	0.64
Q11	Is easily approachable during on-calls	0.88	0.70			0.74	0.41
Q12	Is easily approachable for discussions during (pain) clinic	0.82	0.67			0.73	0.49
Communication of goals				0.94 (0.91)	0.86		
Q13	States learning goals clearly	0.86	0.75			0.85	0.61
Q14	States relevant goals	0.89	0.92			0.86	0.80
Q15	Prioritizes learning goals	0.90	0.72			0.87	0.74
Q16	Repeats stated learning goals periodically	0.93	0.76			0.84	0.72
Evaluation of residents				0.94 (0.93)	0.85		
Q17	Evaluates residents' specialty knowledge regularly	−0.29	0.31			0.87	0.59
Q18	Evaluates residents' analytical abilities regularly	−0.32	0.76			0.90	0.79
Q19	Evaluates residents' application of knowledge to specific patients regularly	−0.32	0.76			0.90	0.72
Q20	Evaluates residents' medical skills regularly	−0.32	0.55			0.80	0.66
Feedback				0.90 (0.86)	0.85		
Q21	Regularly gives positive feedback to residents	−0.74	−0.35			0.67	0.49
Q22	Gives corrective feedback to residents	−0.74	−0.85			0.74	0.64
Q23	Explains why residents are incorrect	−0.86	−0.87			0.87	0.89
Q24	Offers suggestions for improvement	−0.82	−0.74			0.85	0.75

* The items shared the same subject "During my residency in anesthesiology, my attending generally . . ." (residents' instrument) or "In my role as an attending anesthesiologist/faculty, I generally . . ." (faculty self-evaluation); † Reliability coefficients in parentheses represent faculty-level reliability of residents' evaluation.

of goals, evaluation of residents, and feedback (table 2). Cronbach's α for the internal consistency reliability was high for the residents' instrument ranging from 0.89 for the scale "professional attitude towards residents" to 0.94 for both "communication of goals" and "evaluation of residents." Cronbach's α was lower for the faculty self-evaluation instrument, ranging from 0.57 for "learning climate" to 0.86 for "communication of goals." As a result, all scales except "learning climate" in the faculty instrument achieved reliability coefficients above 0.70. Furthermore, the estimates of the faculty (group) level reliability of the residents' instrument ranged from 0.86 (for "professional attitude towards residents") to 0.93 (for the "evaluation of residents" domain).

The item-total scale correlations were high for most items within their scales and in many cases higher for the residents' instrument than for the faculty instrument (table 2). Nonetheless, three items (Q03, Q04, and Q08) on the faculty instrument displayed low item-total correlations. As shown in table 3, the interscale correlations for the residents' instrument ranged from 0.19 (between "professional attitude towards residents" and "evaluation of residents") to 0.66, $P < 0.01$ (between "communication of goals" and "evaluation of residents"). The faculty instrument displayed similar results, from 0.04 (between "professional attitude towards residents" and "evaluation of residents") to 0.63 (between "learning climate" and "evaluation of residents,"

Table 3. Interscale Correlations for Residents' and Faculty Evaluations Separately

	Learning Climate	Professional Attitude Towards Residents	Communication of Goals	Evaluation of Residents	Feedback
Residents					
Learning climate	1	0.34†	0.56†	0.59†	0.53†
Professional attitude towards residents		1	0.32†	0.19†	0.35†
Communication of goals			1	0.66†	0.55†
Evaluation of residents				1	0.57†
Feedback					1
Faculty					
Learning climate	1	0.34*	0.61†	0.63†	0.54†
Professional attitude towards residents		1	0.17	0.04	0.33*
Communication of goals			1	0.51†	0.30
Evaluation of residents				1	0.48†
Feedback					1

* $P < 0.05$, † $P < 0.01$.

$P < 0.01$). For both instruments, all interscale correlations were less than the 0.70 threshold mentioned in the methods section above.

Table 4 displays the bivariate correlations of each of the five scales with the two global ratings. For the residents' instrument, all scales were significantly and positively correlated with the global ratings. The feedback scale had the highest correlations with global ratings of "faculty being seen as a role model" (0.60, $P < 0.001$) and "faculty's overall teaching qualities" (0.68, $P < 0.001$). Contrastingly, the scale "professional attitude towards residents" had the lowest correlations with the global ratings (0.43 and 0.37, respectively, $P < 0.001$). For the faculty self-evaluation instrument, the "learning climate" scale had the highest correlation (0.66, $P < 0.001$) with the global rating "faculty being seen as a role model." However, the "evaluation of residents" scale had the highest correlation (0.59, $P < 0.001$) with the global rating "faculty's overall teaching qualities." Overall, these correlations between the scales and global ratings tended to fall within the expected moderate range of 0.40 to 0.80, according to the literature.[11]

Table 4. Correlations among Scales and Global Ratings of (1) Faculty Being Seen as a Role Model and (2) Faculty's Overall Teaching Qualities, Estimated Separately for Residents' and Faculty's Evaluations

	Faculty Seen as a Role Model		Faculty's Overall Teaching Qualities	
Scales	Residents	Faculty	Residents	Faculty
Learning climate	0.49‡	0.66‡	0.60‡	0.52‡
Professional attitude towards residents	0.43‡	0.16	0.37‡	0.27
Communication of goals	0.52‡	0.43†	0.66‡	0.46†
Evaluation of residents	0.46‡	0.56‡	0.62‡	0.59‡
Feedback	0.60‡	0.57‡	0.68‡	0.56‡

† $P < 0.01$, ‡ $P < 0.001$.

Relationship between Residents' Assessments and Faculty Self-Evaluation

Table 5 shows that residents' assessments of the teaching qualities of their anesthesiology faculty were positive. On scale of 5, the means of the residents' evaluation scale scores for their faculty ranged from 3.41 for "communication of goals" to 4.15 for "professional attitude towards residents." The faculty evaluated themselves highly, with their mean scale scores ranging from 3.22 for "communication of goals" to 4.13 for "professional attitude towards residents."

Looking at the mean scores across the five scales, there was no clear pattern of whether faculty consistently scored themselves higher than the residents scored them. Yet, three of the five scales ("learning climate," "professional attitude towards residents," and "evaluation of residents") showed low to moderate correlations between the rankings produced by the residents' *versus* faculty self scores. The individual items displayed similar results. The residents scored the faculty higher on the two global ratings than the faculty did themselves. There were no rank correlations between residents' *versus* faculty self scores on the global ratings (table 5).

Feasibility: Number of Residents' Assessments Needed per Faculty

For reliable feedback to faculty by using residents' evaluations, the analysis showed that assuming a reliability coefficient of 0.70 for the entire instrument, at least four completed assessments per faculty would be required (table 6). Applying a stricter reliability coefficient of 0.80 would require as many as seven residents evaluating each faculty.

Discussion

Main Findings

This study demonstrates that the two instruments underlying SETQ seem reliable and valid for the evaluation

Table 5. Mean or Averaged Scores and Rank Order Correlations of Residents' and Faculty's Evaluations for the Five Scales of Teaching Qualities and for the Two Global Ratings

Scale and Items	Residents, Mean Score (SEM)	Faculty, Mean Score (SEM)	Kendall's τ for Rank Correlations of Residents' and Faculty Evaluations	*P* for Rank Order Correlation
Learning climate	3.68 (0.07)	3.63 (0.07)	0.24	0.043
Encourages residents to participate actively in discussions	3.80 (0.08)	3.72 (0.11)	0.48	< 0.001
Stimulates residents to bring up problems	3.82 (0.08)	3.75 (0.12)	0.27	0.042
Teaches residents time management	3.41 (0.06)	3.56 (0.13)	0.31	0.019
Keeps to teaching goals; avoids digressions	3.65 (0.07)	3.17 (0.12)	0.34	0.012
Motivates residents to study further	3.80 (0.09)	4.03 (0.12)	0.38	0.005
Stimulates residents to keep up with the literature	3.61 (0.09)	3.56 (0.13)	0.21	0.110
Prepares well for teaching presentations and talks	3.89 (0.09)	3.94 (0.14)	0.13	0.316
Teaches postoperative care in recovery room	3.56 (0.09)	3.25 (0.17)	0.30	0.022
Professional attitude towards residents	4.15 (0.09)	4.13 (0.10)	0.32	0.009
Listens attentively to residents	4.03 (0.08)	4.08 (0.12)	0.27	0.047
Respectful towards residents	4.13 (0.09)	4.25 (0.11)	0.34	0.013
Easily approachable during on-calls	4.23 (0.09)	4.22 (0.14)	0.20	0.135
Easily approachable for discussions during pain clinic	4.22 (0.07)	3.97 (0.18)	0.29	0.025
Communication of goals	3.41 (0.08)	3.22 (0.11)	0.18	0.158
States learning goals clearly	3.50 (0.08)	3.17 (0.13)	0.24	0.068
States relevant goals	3.50 (0.08)	3.36 (0.17)	0.26	0.045
Prioritizes learning goals	3.30 (0.08)	3.28 (0.13)	0.31	0.019
Repeats stated learning goals periodically	3.31 (0.08)	3.06 (0.12)	0.11	0.430
Evaluation of residents	3.70 (0.08)	3.71 (0.10)	0.42	0.001
Evaluates residents' specialty knowledge regularly	3.69 (0.09)	3.39 (0.11)	0.37	0.006
Evaluates residents' analytical abilities regularly	3.69 (0.08)	3.69 (0.12)	0.35	0.009
Evaluates residents' application of knowledge to specific patients regularly	3.73 (0.08)	3.78 (0.12)	0.28	0.036
Evaluates residents' medical skills regularly	3.76 (0.07)	3.97 (0.12)	0.29	0.029
Feedback	3.76 (0.07)	3.82 (0.10)	0.21	0.093
Regularly gives positive feedback to residents	3.73 (0.08)	3.81 (0.12)	0.27	0.040
Gives corrective feedback to residents	3.84 (0.07)	3.67 (0.11)	0.38	0.006
Explains why residents were incorrect	3.73 (0.07)	3.86 (0.14)	0.18	0.182
Offers suggestions for improvement	3.75 (0.07)	3.94 (0.12)	0.24	0.079
Global ratings				
Faculty seen as a role model	3.56 (0.10)	3.44 (0.12)	0.17	0.219
Faculty's overall teaching qualities	3.61 (0.09)	3.36 (0.12)	0.16	0.237

of the teaching qualities of faculty in an academic medical center. Although there were no large differences in the mean scores from the residents' and self-evaluation of the faculty, the rankings produced by those scores were only lowly to moderately correlated if at all. Finally, we saw that the numbers of residents' assessments per faculty needed—in this case, 4 to 7—were achievable in a typical anesthesiology department such as the one in which this study was conducted.

Table 6. Number of Resident Evaluations Needed per Faculty for Reliable Evaluation

	Reliability Coefficient			
Scales	0.60	0.70	0.80	0.90
Learning climate	3	4	7	16
Professional attitude towards residents	3	5	9	20
Communication of goals	3	4	7	15
Evaluation of residents	2	3	6	13
Feedback	4	6	10	22
Overall–all scales combined	2	4	7	15

Limitations of this Study

Before discussing the meaning and implications of these findings, a few study limitations should be explored. First, the small number of faculty (36) relative to the number of items[24] on the faculty instrument could have contributed to the lower reliability coefficient of 0.57 and poorer factor loadings observed for items Q03, Q04, and Q08 on the "learning climate" scale. However, the lower factor loadings could be expected for our sample size, thus prompting us to retain the items in the scale for now.[22] These psychometric observations could also be reflective of faculty's different perception of teaching compared to residents. Although these observations might be a random finding, the psychometric contribution of the problematic items Q03, Q04, and Q08 should be monitored in future research, especially given that items Q03 and Q04 were originally on a separate scale in the SFDP26.[6,8] Second, the cross-sectional design of this study did not support assessment of intrarater (intraresident or intrafaculty) or test-retest reliability. However, the high levels of interrater reliability found here suggest that the intraobserver reliability can only be higher.[11] Third, we observed that the three faculty members who did not fill out their self-evaluation were scored lower than average on all six domains (results available from authors on request). This raises questions about the impact of nonresponse on the generalizability of the findings especially for the faculty instrument. Fortunately, if residents continue to assess faculty who abscond and the (conditional) probability of faculty participation could be estimated reliably, then it might be possible to use the resulting information to adjust the faculty summary scores if the aim of measurement were to produce generalizable faculty population estimates. Nevertheless, the findings presented here may not be generalizable to residents and faculty in other specialties or (academic) institutions because each residency program and organization has its own structures and cultures. Work is currently being done to replicate the findings of our studies in different settings. Lastly, given the short follow-up period of faculty, this study cannot yet draw any conclusions about the impact of SETQ on the quality of teaching.

Explanation of Results

Globally, anesthesiology residencies are increasingly competency-based.[3,10,23] Our study, like one other recent work,[3] has now developed tools that could be adapted for the systematic evaluation and support of faculty involved in those residency programs. Our findings provide strong empirical support for the reliability and validity of the results obtained from the two resident-completed and self-completed instruments for faculty evaluation. The reliability findings as illustrated by the high faculty-level reliability coefficients estimated in table 2 indicate that residents' instruments can be used to measure and compare the teaching qualities of faculty. The results of the psychometric analysis indicate that we could tap into five domains seen as relevant aspects of teaching by both residents and faculty. We note, however, that the relatively modest interitem correlations among "learning climate," "communication of goals," "evaluation of residents," and "feedback" could indicate that some items within those scales might provide some redundant information. For future measurements, we will further assess the uniqueness of the related items.

The expected modest correlations of each of the five scales with the two global ratings provide an intuitive support for the five teaching domains as part of the phenomenon of clinical teaching. This explanation is premised on the assumption and finding[17,24] that good teachers and good clinicians make good role models for trainees. Moreover, if the five scales measured teaching qualities and the one-item global rating on overall teaching skills did so similarly, then we could expect the scales to correlate at least moderately with the global rating (as indeed was the case). The latter correlations should not, however, be too high (for example, greater than 0.80) because that would point to redundancy of the entire instrument, that is, if it could be reduced to one global item.[11] These findings provide additional evidence of the validity of the SETQ instruments.

The lack of strong correlations between faculty and residents' ratings is consistent with recent research and systematic review that shows that physicians had a limited ability to self-assess accurately.[14,15] These findings neither support nor undercut the validity of the instruments among residents and faculty. The rank order correlations are not intended as a measure of the validity of the instruments but as a measure of the degree of overlap between how residents experience faculty's teaching and how the faculty themselves evaluate their own teaching. It is possible for both instruments to be valid and provide reliable results among residents and faculty, respectively, and yet yield low correlations between faculty and residents if, as was found here and elsewhere,[14,15] faculty's self-evaluation differs from residents' evaluation of faculty. We have no defensible reasons to expect that self-evaluation must perfectly or even moderately reflect other people's evaluation of the same constructs in a specified population (in this case, faculty). In fact, based on celebrated paradoxical results in behavioral psychology demonstrating failures of human perception and reasoning given strong *a priori* expectations and beliefs, we would do well to expect otherwise.[25] Previous work has noted that the weakest performers, as determined by external assessments, self-assessed poorly.[26] We speculate that this might be related to strong *a priori* expectations and beliefs of own qualities not shared by other observers. Nonetheless, the faculty self-evaluation used here parallels recent developments in performance assessment where those who are assessed are increasingly being assessed from different perspectives and sources in what is sometimes termed 360-degree assessment.[27] Obtaining regular feedback from others, such as residents, may help faculty focus their learning activities on legitimate improvement needs.

Our results also suggest that between four and six residents' evaluations per faculty (assuming the standard reliability level of 0.70 seen in the literature) will be required to assess faculty teaching qualities reliably. These numbers seem reasonable for most anesthesiology residency programs. This finding buttresses the feasibility of applying the SETQ instruments routinely.

Implications for Clinical Education, Research, and Policy

SETQ was designed, tested, and implemented for formative purposes. It was never intended as a summative review, although the demonstrated quality of the instruments and the achievement of adequate reliability would allow use in a high stakes context. Developing a system for measuring and, perhaps, improving the quality of faculty members' teaching is indispensible because it is known to affect residents' performance in clinical examinations.[28] Therefore, faculty feedback reporting appears justifiable. The individualized feedback reports implemented in SETQ seemed well-received, and anecdotal reports from faculty suggest that SETQ was raising awareness of effective teaching and sometimes leading to improvement. The reports may help faculty to focus their personal development plans. Feedback and increased insights may, however, not always be sufficient to bring about behavioral change.[4] Therefore, SETQ was designed to include a formative follow-up interview, with the program director aiming at designing individual development tracks. Other studies suggest that this may increase actual performance improvement.[2,29,30] Future assessments need to demonstrate actual improvements.

Conclusions

Our SETQ instruments provided reliable and valid results that could be used in formative support of anesthesiology faculty. This study went further than previous work to include the voice of the faculty to evaluate and encourage self-insight. The instruments are now available, do not require unachievable number of residents' evaluations per faculty, and yield faculty feedback seen as useful in designing individual development tracks.

The authors acknowledge the input in this study of all anesthesiology faculty and residents of the Academic Medical Center (AMC), Amsterdam, The Netherlands. The authors thank Bastiaan Roset, M.Sc., Junior Researcher, AMC, for additional help with data preparation and analysis.

References

1. Tetzlaff JE: Assessment of competency in anesthesiology. Anesthesiology 2007; 106:812–25

2. Seifert CF, Yukl G, McDonald RA: Effects of multisource feedback and a feedback facilitator on the influence behavior of managers toward subordinates. J Appl Psychol 2003; 88:561–9

3. de Oliveira Filho GR, Dal Mago AJ, Garcia JH, Goldschmidt R: An instrument designed for faculty supervision evaluation by anesthesia residents and its psychometric properties. Anesth Analg 2008; 107:1316–22

4. Grol R: Personal paper. Beliefs and evidence in changing clinical practice. BMJ 1997; 315:418–21

5. Skeff KM, Stratos GA, Berman J, Bergen MR: Improving clinical teaching. Evaluation of a national dissemination program. Arch Intern Med 1992; 152: 1156–61

6. Litzelman DK, Stratos GA, Marriott DJ, Skeff KM: Factorial validation of a widely disseminated educational framework for evaluating clinical teachers. Acad Med 1998; 73:688–95

7. Litzelman DK, Westmoreland GR, Skeff KM, Stratos GA: Factorial validation of an educational framework using residents' evaluations of clinician-educators. Acad Med 1999; 74:S25–7

8. Williams BC, Litzelman DK, Babbott SF, Lubitz RM, Hofer TP: Validation of a global measure of faculty's clinical teaching performance. Acad Med 2002; 77:177–80

9. Wong JG, Agisheva K: Developing teaching skills for medical educators in Russia: A cross-cultural faculty development project. Med Educ 2007; 41:318–24

10. Lombarts MJ, Bucx MJ, Rupp I, Keijzers PJ, Kokke SI, Schlack W: [An instrument for the assessment of the training qualities of clinician-educators.] Ned Tijdschr Geneeskd 2007; 151:2004–8

11. Streiner DL, Norman GR: Health Measurement Scales: A Practical Guide to Their Development and Use, 4th Edition. Oxford, Oxford University Press, 2008, pp 5–36, 77–102, 167–206, and 247–74

12. Arah OA, ten Asbroek AH, Delnoij DM, de Koning JS, Stam PJ, Poll AH, Vriens B, Schmidt PF, Klazinga NS: Psychometric properties of the Dutch version of the Hospital-level Consumer Assessment of Health Plans Survey instrument. Health Serv Res 2006; 41:284–301

13. Cronbach LJ: Coefficient alpha and the internal structure of tests. Psychometrika 1951; 16:297–334

14. Claridge JA, Calland JF, Chandrasekhara V, Young JS, Sanfey H, Schirmer BD: Comparing resident measurements to attending surgeon self-perceptions of surgical educators. Am J Surg 2003; 185:323–7

15. Davis DA, Mazmanian PE, Fordis M, Van HR, Thorpe KE, Perrier L: Accuracy of physician self-assessment compared with observed measures of competence: A systematic review. JAMA 2006; 296:1094–102

16. Carey RG, Seibert JH: A patient survey system to measure quality improvement: questionnaire reliability and validity. Med Care 1993; 31:834–45

17. Maker VK, Curtis KD, Donnelly MB: Are you a surgical role model? Curr Surg 2004; 61:111–5

18. Kendall M: A new measure of rank correlation. Biometrika 1938; 30:81–9

19. van der Hem-Stokroos HH, van der Vleuten CPM, Daelmans HE, Haarman HJ, Scherpbier AJ: Reliability of the clinical teaching effectiveness instrument. Med Educ 2005; 39:904–10

20. Raykov T: On multilevel model reliability estimation from the perspective of structural equation modeling. Structural Equation Modeling 2006; 13:130–41

21. Kraemer HC: Ramifications of a population model for *k* as a coefficient of reliability. Psychometrika 1979; 44:461–72

22. Stevens JP: Applied Multivariate Statistics for the Social Sciences, 3rd Edition. Hillsdale, NJ, Lawrence Erlbaum Associates, Inc., 1996, pp 362–92

23. Rashid A, Doger A, Gould G: National survey of College Tutors in the UK regarding training in medical education. Br J Anaesth 2008; 100:42–4

24. Skeff KM, Mutha S: Role models–guiding the future of medicine. N Engl J Med 1998; 339:2015–7

25. Gilovich T: How We Know What Isn't So: The Fallibility of Human Reason in Everyday Life. New York, Free Press, 1993, 9–122

26. Violato C, Lockyer J: Self and peer assessment of pediatricians, psychiatrists and medicine specialists: Implications for self-directed learning. Adv Health Sci Educ Theory Pract 2006; 11:235–44

27. Overeem K, Faber MJ, Arah OA, Elwyn G, Lombarts KM, Wollersheim HC, Grol RP: Doctor performance assessment in daily practise: Does it help doctors or not? A systematic review. Med Educ 2007; 41:1039–49

28. Blue AV, Griffith CH III, Wilson J, Sloan DA, Schwartz RW: Surgical teaching quality makes a difference. Am J Surg 1999; 177:86–9

29. Marvel MK: Improving clinical teaching skills using the parallel process model. Fam Med 1991; 23:279–84

30. Metheny WP, Espey EL, Bienstock J, Cox SM, Erickson SS, Goepfert AR, Hammoud MM, Hartmann DM, Krueger PM, Neutens JJ, Puscheck E: To the point: Medical education reviews evaluation in context: assessing learners, teachers, and training programs. Am J Obstet Gynecol 2005; 192:34–7