Evaluating clinicians’ teaching performance

Boerebach, B.C.M.

Citation for published version (APA):
CHAPTER 5

CLINICAL TEACHING PERFORMANCE IMPROVEMENT OF FACULTY IN RESIDENCY TRAINING: A PROSPECTIVE COHORT STUDY

Renée M. van der Leeuw

Benjamin C.M. Boerebach

Kiki M.J.M.H. Lombarts

Maas Jan Heineman

Onyebuchi A. Arah

Submitted
ABSTRACT

The purpose of this study was to investigate the predictive value of residents’ numerical and narrative feedback and faculty’s completion of a self-evaluation on faculty’s teaching performance improvement as perceived by residents. We present a prospective cohort study that was conducted in the Netherlands from September 1, 2008 till February 1, 2013. Nine hundred and one residents and 1,068 faculty of 65 teaching programs in 16 hospitals were invited to annually (self-) evaluate teaching performance using the validated System for Evaluation of Teaching Qualities (SETQ). We used multivariable adjusted generalized estimating equations to analyze the effects of (i) residents’ numerical and (ii) narrative feedback and (iii) whether or not teaching faculty self-evaluated on residents’ perception of faculty’s teaching performance improvement.

The average response rate over three years was 69% for faculty and 81% for residents. Higher numerical feedback scores were associated with improved teaching performance one year following the first measurement (regression coefficient, $b$: 0.077; 95% CI: 0.002 to 0.151; $p=0.045$), but not after the second wave of receiving feedback and evaluating improvement. Receiving more suggestions for improvement was associated with improved teaching performance throughout follow-up. Neither the number of positive comments nor participation in self-evaluation had impact on teaching performance improvement.

In conclusion, teaching performance evaluation systems may be helpful in enhancing faculty’s teaching performance in residency training. Receiving more narrative suggestions for improvement might help faculty to improve their teaching performance in subsequent years.
INTRODUCTION

In residency training, or postgraduate medical education, teaching faculty supervise residents, serve as role models and facilitate learning of residents in the clinical setting.\textsuperscript{1-4} To support faculty in providing high quality training during daily patient care, teaching evaluation systems have been developed.\textsuperscript{5-9} Generally, these evaluation systems provide faculty with performance feedback and often consist of residents’ feedback and a self-evaluation. There is limited evidence on the effect of these systems on faculty’s subsequent teaching performance. Organizational psychology literature and knowledge from clinical performance improvement of physicians suggest that research should focus on \textit{how} individuals improve their performance rather than the question \textit{whether} performance improves through feedback.\textsuperscript{10-12} Therefore, the overall aim of the study is to investigate which aspects of teaching evaluation systems predict improvement of future performance of individual faculty providing residency training.

First, in teaching evaluation systems residents’ feedback is gathered, which generally consists of numerical scores and narrative comments. The predictive value of numerical feedback on subsequent performance was found to be small, but potentially meaningful.\textsuperscript{12,13} However, there is a lack of knowledge on the impact of numerical feedback scores (higher or lower) on subsequent performance.

Second, in addition to numerical feedback, there is increased attention for narrative comments, or written feedback. Narrative feedback is regarded as a rich source of information and shown to add individualized information to the evaluation scores.\textsuperscript{14,15} The hypothesis would therefore be that receiving more narrative feedback stimulates performance improvement.

Third, besides the input of residents’ feedback on teaching performance, a critical self-evaluation by faculty, as is frequently used in multi-source feedback assessments in clinical performance assessment, can complement performance assessment.\textsuperscript{16,17} Comparing both evaluations (residents’ and self) can create direction for improvement: discrepancies between self- and other-ratings can be motivating when over raters receive critical feed-
back. Furthermore, individuals who increase the discrepancy of self-other ratings by aiming for higher ratings may be more motivated to improve their performance. We hypothesize that faculty who complete a self-evaluation and thus create something to compare residents’ feedback to, will subsequently improve their performance more than faculty who do not complete a self-evaluation. It is important to test this hypothesis, because the self-other discrepancy can harm ones self-image or result in decreased motivation when unfavorable feedback is received by someone with low self-esteem.

The aim of this prospective cohort study is to investigate the predictive value of (i) numerical feedback, (ii) number of both positive feedback comments and suggestions for improvement, and (iii) completing a self-evaluation on subsequent teaching performance of faculty who train residents.

**METHODS**

*Study setting and the SETQ-system*

In the Netherlands, the trajectory to become a medical specialist starts after finishing six years of high school. Medical students than proceed through six years of study with two to three years of placements. After graduating as medical doctors, residency training consists of three to six years training in both university medical centers as well as university affiliated community based teaching hospitals. Within residency training, all faculty are responsible for the training of residents while ensuring high quality patient care. In the Netherlands, the development of the System for Evaluation of Teaching Qualities (SETQ system) started in 2005 with a pilot study followed by implementation of a web-based system to carefully assess all aspects of the system. In 2008, the web-based SETQ system was launched to provide easy access to all faculty interested in evaluating and possibly improving their teaching performance. The system has been validated for anesthesiology, obstetrics and gynecology, surgical specialties, and medical specialties. Residents’ evaluation of faculty’s teaching performance and faculty’s
self-completed evaluation both consist of the same 22 generic questions and 2 to 8 specialty specific questions and were executed simultaneously. The questions can be organized into five teaching domains: creating a positive and safe learning climate, keeping a professional attitude towards residents, communicating about learning goals, evaluating residents’ competence and performance, and providing feedback. In addition, residents are encouraged to provide written comments on faculty’s teaching performance: ‘Please provide strengths of faculty’s teaching performance’ and ‘Please provide concrete suggestions for improvement of faculty’s teaching performance’. Residents are free to choose whom to evaluate, based on whom they have worked with. After completion of the measurement instruments, faculty individually receive an automatically generated SETQ feedback report containing all residents’ evaluations, their self-evaluation scores, a numerical comparison with colleagues, a peer group ranking (in groups with 15 or more faculty), and all narrative comments. Previous studies showed that the feedback reports of faculty who received four to six residents’ evaluations could be considered statistically reliable. After the first wave of measuring teaching performance and receiving feedback, subsequent measurements were complemented with a question on improvement of faculty’s teaching performance: ‘If you worked with this faculty during the past year, have you observed any changes in his or her teaching performance as your supervisor (since the last SETQ evaluation), and if so, how would you qualify these changes?’. This question was answered on a 4-point Likert response scale: 1 = declined; 2 = not changed; 3 = somewhat improved; 4 = greatly improved. There was an additional option “I cannot judge”.

Study population

From September 1, 2008 to February 1, 2013, residents and faculty could enroll in a yearly evaluation of teaching performance of individual faculty through the SETQ system. Participants were followed-up for three consecutive years (which included two waves of receiving SETQ feedback and two follow-up measurements on teaching performance improvement). Faculty’s identification codes were matched. In total, 1068 faculty and 901 residents from 65 residency training programs, representing four university-based and
twelve community-based hospitals, were invited. No financial compensation was provided for participation.

**Outcome measure and predictors**

The main outcome of all the models was the teaching performance improvement of faculty rated by the residents one year after the feedback was provided. Residents’ ratings of improvement on the above mentioned 4-point Likert scale was aggregated to faculty level and used as outcome in the analyses. The mean of all numerical SETQ-items was calculated per evaluation. Subsequently, these mean scores were aggregated per faculty, resulting in the predictor variable ‘numerical feedback’. The ‘narrative feedback’ was coded and counted before included in the analyses. We coded the narratives on their direction: ‘positive comments’ or ‘suggestions for improvement’ and counted the number of these comments. Counting the number of comments in a specific category was found to be possible with an inter-rater reliability of 0.98. Consequently, one researcher (RvdL) could individually code the direction and number of narrative feedback comments. The number of positive comments and suggestions for improvement were predictor variables in the models. Finally, we also included a binary variable on whether or not each faculty completed a self-evaluation of their own teaching performance concurrently during the wave in which residents evaluated them.

**Data analysis**

Prior to analysis, all variables were examined for accuracy of data entry, missing values, and meeting the assumptions of multivariable analysis. We handled missing data using the well-established techniques of multiple imputation. We created five imputed databases which were then analyzed in parallel; their results were pooled using methods that accounted for the uncertainty of the imputation process. The data were aggregated to the faculty level. We analyzed our data using linear generalized estimating equations (GEE) that allowed for the clustering of the outcome data on faculty within hospitals and specialties. Given our longitudinal data, we specified our regression models as structural nested mean models, the parameters of which were estimated using sequential g-estimation. G-estimation is
a sophisticated analytical technique developed for use in longitudinal data settings where some predictors can be mediators and confounders of subsequent predictor-outcome relationships, as in this study. We fit four final models altogether. The first model analyzed the influence of residents’ feedback, namely i) their numerical feedback score, ii) the number of positive comments provided, and iii) the number of suggestions for improvement provided in the first SETQ wave as predictors of perceived performance improvement measured in the second SETQ wave. That is, the outcome was faculty’s teaching performance improvement, as rated by the residents one year later (at wave 2). The second model was equivalent to model 1, but further adjusted for residents’ feedback provided in the second SETQ wave and using faculty’s teaching performance improvement one year after (that is, at wave 3) as the outcome. Models 3 and 4 added, to models 1 and 2 respectively, the variable on whether faculty performed self-evaluation. All models were additionally adjusted for these covariates: hospital, specialty, faculty’s age, and faculty’s gender. We conducted all analyses using SPSS 20.0 for Windows OS (IBM SPSS Inc., Chicago, 2012).

Ethical consideration

We consulted the institutional ethical review board of the Academic Medical Center of the University of Amsterdam (AMC). They confirmed that the Medical Research Involving Human Subjects Act (WMO) did not apply to this study and thus waived ethical approval.

RESULTS

In total, residents completed 17,049 evaluations of faculty’s teaching performance in three years. Over these three years, residents’ response rates were 91%, 80% and 72% respectively. After data imputation, data was available for 992 faculty for the first wave and 847 faculty for the second wave. Participants’ characteristics are shown in table 1. As shown in table 2, faculty received an average numerical feedback score of 3.82 over the years. The narrative feedback per faculty consisted of approximately 10 positive com-
ments and 4 suggestions for improvement. Most faculty completed a self-evaluation (812 in the first year and 597 in the second year). When categorizing residents’ perception of faculty’s improvement, most faculty somewhat improved their teaching performance or did not change (see table 2).

The results of the multilevel regression models indicate that faculty who received higher numerical feedback scores were more likely to improve (b=0.077, p=0.045 in model 1 and b=0.077, p=0.042 in model 3) in SETQ wave 1, however, no effect of numerical feedback scores was found in SETQ wave 2. The amount of positive narrative comments was not associated with improvement in both SETQ wave 1 and wave 2. Faculty who received more suggestions for improvement were more likely to improve their teaching performance (first SETQ wave: b=0.010, p=0.056 for model 1, and b=0.011, p=0.050 for model 3; second SETQ wave: b=0.009, p=0.006 for model 2 and b=0.009, p=0.006 for model 4). Whether or not faculty self-evaluated did not impact their improvement. All results are presented in table 3.

### Table 1 Characteristics of faculty and residents who participated

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of faculty who participated (response rate)</td>
<td>992 (93%)</td>
</tr>
<tr>
<td>Number of residents invited to evaluate faculty’s teaching performance during the first SETQ wave (response rate)</td>
<td>901 (91%)</td>
</tr>
<tr>
<td>Number of residents invited to evaluate faculty’s improvement after the first SETQ wave and faculty’s teaching performance during the second SETQ wave (response rate)</td>
<td>816 (80%)</td>
</tr>
<tr>
<td>Number of residents invited to evaluate faculty’s improvement of teaching performance after the second SETQ wave (response rate)</td>
<td>862 (72%)</td>
</tr>
<tr>
<td>Faculty from medical specialties</td>
<td>66.1%</td>
</tr>
<tr>
<td>Faculty from surgical specialties</td>
<td>33.9%</td>
</tr>
<tr>
<td>Faculty from university teaching hospitals</td>
<td>68.9%</td>
</tr>
<tr>
<td>Faculty from community-based, university affiliated hospitals</td>
<td>31.1%</td>
</tr>
<tr>
<td>Percentage of male faculty</td>
<td>60.1%</td>
</tr>
<tr>
<td>Mean age of faculty (standard deviation)</td>
<td>46.75 years (8.50)</td>
</tr>
</tbody>
</table>
DISCUSSION

Main findings

This study provides empirical data on the effect of residents’ feedback and faculty’s self-evaluation on subsequent teaching performance of faculty working in residency training. Receiving higher numerical feedback scores increased performance, but only after the first feedback wave. The positive effect of the narrative suggestions for improvement on teaching performance was seen in both the first and second feedback waves. Finally, completing a self-evaluation or receiving more positive narrative feedback had no effect on subsequent improvement.

Explanation and interpretation of results

Our aim was to investigate the effect of different aspects of teaching performance evaluation on teaching performance improvement as perceived by residents.

First, numerical feedback was expected to yield small effect sizes, which was confirmed by this study. Even small effect sizes might represent meaningful performance improvement.\(^{10}\) Surprisingly, we found that a higher numerical feedback score had a positive effect on improvement in the first year but no effect in the second year. Residents may become more critical of their ‘good’ teachers by increased attention to teaching performance through participation in the SETQ system. This could take away the effect that was found in the first SETQ wave.

Second, the narrative feedback showed a sustained effect over the years, but only for the suggestions for improvement. Faculty’s improved performance in response to suggestions for improvement may be explained by Finkelstein and colleagues’ theory on the response to positive and negative feedback in relation to a person’s level of expertise.\(^{18}\) They state that novices seek and respond more to positive feedback and experts favor negative feedback as it increases a sense of urgency that their goals are (not yet) attained. In terms of our study findings, it suggests that most faculty are or perceive themselves
Table 2 Descriptive statistics of key variables for the first and second SETQ waves

<table>
<thead>
<tr>
<th></th>
<th>1&lt;sup&gt;st&lt;/sup&gt; SETQ wave</th>
<th>2&lt;sup&gt;nd&lt;/sup&gt; SETQ wave</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Evaluation characteristics</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of faculty who received a feedback report</td>
<td>992</td>
<td>847</td>
</tr>
<tr>
<td>Median number of residents’ that evaluated faculty’s teaching performance (±IQR)</td>
<td>5 (3-8)</td>
<td>5 (3-7)</td>
</tr>
<tr>
<td><strong>Predictor variables</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Numerical feedback or the mean SETQ score (±SD)</td>
<td>3.82 (0.47)</td>
<td>3.82 (0.49)</td>
</tr>
<tr>
<td>Median number of positive comments (±IQR)</td>
<td>10 (4-17)</td>
<td>9 (5-19)</td>
</tr>
<tr>
<td>Median number of suggestions for improvement (±IQR)</td>
<td>4 (1-7)</td>
<td>4 (1-7)</td>
</tr>
<tr>
<td>Number of faculty who self-evaluated</td>
<td>812</td>
<td>597</td>
</tr>
<tr>
<td><strong>Outcome variable</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean teaching performance improvement ` (SD)</td>
<td>2.23 (0.40)</td>
<td>2.21 (0.36)</td>
</tr>
<tr>
<td>Declined: 1 - 1.9 (percentage)</td>
<td>10.5 %</td>
<td>9.1 %</td>
</tr>
<tr>
<td>Not changed: 1.9 – 2.1 (percentage)</td>
<td>39.4 %</td>
<td>45.9 %</td>
</tr>
<tr>
<td>Somewhat improved: 2.1 – 2.9 (percentage)</td>
<td>42.2 %</td>
<td>39.0 %</td>
</tr>
<tr>
<td>Greatly improved: 2.9 – 4 (percentage)</td>
<td>7.8 %</td>
<td>6.0 %</td>
</tr>
</tbody>
</table>

* A score of 2 indicated “no improvement nor decline” of teaching performance
Table 3 Sequential g-estimates of the effects of residents’ feedback and faculty’s self-evaluation on faculty’s teaching performance improvement

<table>
<thead>
<tr>
<th>Wave 1 predictors for wave 2 outcome</th>
<th>Wave 2 predictors for wave 3 outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Regression coefficient (b)</td>
</tr>
<tr>
<td><strong>Model 1</strong></td>
<td></td>
</tr>
<tr>
<td>Numerical feedback</td>
<td>0.077</td>
</tr>
<tr>
<td>Number of positive comments</td>
<td>-0.005</td>
</tr>
<tr>
<td>Number of suggestions for</td>
<td>0.010</td>
</tr>
<tr>
<td>improvement</td>
<td></td>
</tr>
<tr>
<td><strong>Model 3</strong></td>
<td></td>
</tr>
<tr>
<td>Numerical feedback</td>
<td>0.077</td>
</tr>
<tr>
<td>Number of positive comments</td>
<td>-0.005</td>
</tr>
<tr>
<td>Number of suggestions for</td>
<td>0.011</td>
</tr>
<tr>
<td>improvement</td>
<td></td>
</tr>
<tr>
<td>Self-evaluation (reference = not</td>
<td>-0.079</td>
</tr>
<tr>
<td>self-evaluated)</td>
<td></td>
</tr>
</tbody>
</table>

All models were adjusted for hospital, specialty, faculty’s sex and faculty’s age.
as experts, as they responded to the suggestions for improvement (negative feedback) in relation to performance improvement.

We found no effect of positive comments on faculty’s teaching performance improvement. This may be the result of the above-described effect of experts’ seeking and responding to negative feedback. Although the positive comments did not predict performance improvement, they may be vital in supporting faculty’s motivation when receiving negative feedback and preventing a detrimental effect of negative feedback on performance.

Third, completing a self-evaluation did not predict performance improvement. Whether or not faculty complete a self-evaluation might capture something that is not of influence on residents’ perception of teaching performance. In other words, the internal reflective process that may have been initiated by completing a self-evaluation may not be within reach of residents’ perception and hence not be captured with our improvement question. This does not make residents’ perception of improvement an incorrect outcome measure; external assessment is more accurate than self-reported outcomes.

**Strengths and limitations**

The strengths of this study are its longitudinal observation of a study population consisting of a large number of participants across multiple specialties from multiple hospital and medical training sites. Furthermore, the statistical analyses were conducted using sophisticated statistical methods. Multiple imputations were used to deal with missing data in an appropriate way, and sequential g-estimation was implemented to study predictors of changes in participants’ teaching performance over time.

At the start of the study, we considered the difference in SETQ-mean score between two time points as a potential outcome measure. Considering the literature on the effect of feedback on performance, we expected small but relevant effects. Although the SETQ instruments were validated to discriminate between individual faculty at a certain point in time, they have not been tested for their ability to measure changes in teaching performance within individual faculty over a period of time. Furthermore, we assumed that,
for faculty who already performed well (scoring at the top of the scale), potential changes in teaching performance would not be visible in the SETQ scores at different time points. The positively skewed SETQ scores (mean 3.82 on a 5-point Likert scale) combined with the limited sensitivity of the 5-point Likert scale could lead to the inability of the measure to reflect improvement of high-performing faculty validly based on differences in SETQ scores. Therefore, we chose to include the additional question to specifically enquire residents about their perception of faculty’s teaching performance improvement. Again, we emphasize that this measure should be interpreted as residents’ perception of whether faculty improved their teaching performance. Previous literature indicates that perception of residents is preferred over faculty’s self-perceived improvement. Additional measures such as observation of teaching practice could increase the strength of the main outcome measure in future research.

**Implications for practice**

The results of this study show that even faculty who received high numerical feedback scores (high performers) could still improve their teaching performance after one year. Even more, the high performers were more likely to improve their teaching performance after the first feedback wave compared to the low performers. Additionally, the suggestions for improvement as phrased by residents are effective aspects of the feedback system. A concrete implication for faculty’s teaching practice is thus to pay attention to these suggestions for improvement. Furthermore, residents should provide suggestions for improvement in order to aid faculty’s teaching performance improvement. Finally, longitudinal participation appears to be beneficial, as noted earlier in clinical performance feedback, it seems to stimulate participants to take action when they know there will be another evaluation next year.

**Suggestions for future research**

Many performance assessment systems include the professionals’ self-evaluation. Based on this study’s findings that self-assessment does not seem to influence performance improvement, exclusion of self-evaluations
from performance assessment systems could be suggested. However, further research is needed to investigate the possible internal processes that self-evaluation may stimulate, as has been initiated by Sargeant et al.\textsuperscript{17,36} As noted above, residents’ perception of faculty’s teaching performance improvement might not capture internal processes. Since the understanding of improvement of teaching performance is still in its infancy, it deserves further research. Additionally, other contextual characteristics should be investigated to increase our understanding of performance and improvement. For example, coping strategies of faculty in dealing with feedback or the way feedback is discussed, individually or in a group. Future work should also elaborate on how to measure improvement in other ways than through residents’ perception.

**ACKNOWLEDGEMENTS**

The authors would like to thank all faculty and residents who generously participated in our study. We also thank MEDOX.nl for developing and maintaining the web application. This study is part of the research project “Quality of clinical teachers and residency training programs” which is co-financed by the Dutch Ministry of Health, the Academic Medical Center, Amsterdam, and the Faculty of Health and Life Sciences of the University of Maastricht. Prof. Onyebuchi A. Arah was supported by Veni career grant number 916.96.059 awarded by the Netherlands Organization for Scientific Research (NWO). The funders had no role in the design and conduct of the study; collection, management, analysis and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication. R.M. van der Leeuw and B.C.M. Boerebach had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.
REFERENCES

(1) ACGME. Common Program Requirements. 2011


