Evaluating clinicians’ teaching performance
Boerebach, B.C.M.

Citation for published version (APA):
CHAPTER 4

THE IMPACT OF RESIDENT- AND SELF-EVALUATIONS ON SURGEON’S SUBSEQUENT TEACHING PERFORMANCE

Benjamin C. M. Boerebach

Onyebuchi A. Arah

Maas Jan Heineman

Olivier R. C. Busch

Kiki M.J.M.H. Lombarts

ABSTRACT

Background: This study evaluates how residents’ evaluations and self-evaluations of surgeon’s teaching performance evolve after two cycles of evaluation, reporting and feedback. Furthermore, the influence of over- and underestimating own performance on subsequent teaching performance was investigated.

Methods: In a multicenter cohort study, 351 surgeons evaluated themselves and were also evaluated by residents during annual evaluation periods for three subsequent years. At the end of each evaluation period, surgeons received a personal report summarizing the residents’ feedback. Changes in each surgeon’s teaching performance evaluated on a 5-point scale were studied using growth models. The effect of surgeons over- or underestimating their own performance on the improvement of teaching performance was studied using adjusted multivariable regressions.

Results: Compared to the first (median score: 3.83, 20th – 80th percentile score: 3.46 – 4.16) and second (median: 3.82, 20th – 80th: 3.46 – 4.14) evaluation period, residents evaluated surgeon’s teaching performance higher during the third evaluation period (median: 3.91, 20th – 80th: 3.59 – 4.27), p<0.001. Surgeons did not alter self-evaluation scores over the three periods. Surgeons who overestimated their teaching performance received lower subsequent performance scores by residents (regression coefficient $b=-0.08$, 95% confidence limits (CL) = -0.18, 0.02) and self ($b=-0.12$, 95% CL= -0.21, -0.02). Surgeons who underestimated their performance subsequently scored themselves higher ($b=0.10$, 95% CL= 0.03, 0.16), but were evaluated equally by residents.

Conclusions: Residents’ evaluation of surgeon’s teaching performance was enhanced after two cycles of evaluation, reporting and feedback. Overestimating own teaching performance could impede subsequent performance.
INTRODUCTION

Training residents is a key task of surgeons in teaching hospitals. Gaining insights into the strengths and weaknesses of surgeons’ teaching performance is crucial for the maintenance and enhancement of high quality training programs. There is evidence suggesting that unguided (isolated) self-evaluation of performance does not provide sufficient information for adequate performance enhancement.\textsuperscript{1,2} In response to these findings, a process of informed self-evaluation, including both external and internal data to self-evaluate performance, has been suggested as a valuable alternative.\textsuperscript{3} There are several feedback sources that can provide an external view on surgeons’ teaching performance, including feedback of residents in training. Previous research has shown that surgeons found residents’ feedback valuable, especially when they combined it with a self-evaluation of their performance to enhance their self-awareness.\textsuperscript{4} Robust performance evaluation systems are now available to guide residents in the process of collecting and feeding back surgeons’ performance data for the purpose of informing surgeons about their teaching performance.\textsuperscript{5-7} However, the effects of such evaluation systems on surgeons’ subsequent teaching performance are unknown. Therefore, this study evaluates how surgeons’ teaching performance evolves after two cycles of evaluation, reporting and feedback.

In the process of informed self-evaluation, internal and external data sources are integrated to provide a comprehensive overview of surgeons’ performance.\textsuperscript{3} However, combining and comparing such data sources as resident-evaluations and self-evaluations of surgeons’ teaching performance, can result in tensions on behalf of surgeons, thereby leading to delay or even dismissal of self-improvement actions.\textsuperscript{8-11} Especially surgeons who reveal a discrepancy between self- and external evaluations of their performance, may develop (emotional) reactions that can impact their reaction towards their performance feedback and subsequently their actual performance improvement.\textsuperscript{8-13} Besides, psychological studies show that discrepancies between self and other perceptions of one’s performance can be perceived as unsatisfactory and suggest that overestimating can impede subsequent performance, while underestimating is usually harmless for performance.\textsuperscript{14-20}
Consequently, surgeons may aim to minimize the discrepancy by either attempting to influence resident-evaluations or by adjusting their self-evaluations. This study evaluates the influence of over- or underestimation on subsequent teaching performance. This study has two main aims. First, we will explore how resident-evaluations and self-evaluations of surgeons’ teaching performance evolve after two cycles of evaluation, reporting and feedback. Second, we will explore if over- or underestimating of surgeons’ own performance influences resident- and self-evaluations of surgeons’ subsequent performance.

**MATERIALS AND METHODS**

*Setting and Study Population*

This study was conducted at 29 surgical teaching programs in 13 hospitals, including general surgery (10), obstetrics & gynecology (10), ophthalmology (3), orthopedic surgery (2), otorhinolaryngology (1), urology (1), neurosurgery (1) and plastic surgery (1). Teaching programs could participate voluntarily by approaching the project leaders. In the Netherlands, postgraduate medical training is organized in eight geographical regions, each coordinated by an academic medical center. All larger (>5 residents) surgical training programs that were based at or coordinated from the project leaders’ academic medical center participated in this study (24 out of the 29 programs included in this study). Additionally, five training programs from other regions in the Netherlands participated. Data was collected from September 2008 till May 2013 and occurred during annual evaluation periods lasting one month. Residents could choose which and how many surgeons to evaluate, based on whose teaching performance the resident believed he/she could evaluate accurately. For each residency training program, data from three subsequent evaluation periods were included, which represent two full cycles of evaluation, feeding back, follow-up and re-evaluation. In total, 351 surgeons were invited to participate in this study. Only surgeons who participated during the first evaluation period at their training program were included in this study; none of the surgeons could enter during a later evaluation periods. All residents were
asked to provide feedback. Overall, 299 residents were invited to evaluate surgeons’ teaching performance during the first, 346 during the second and 341 during the third evaluation period. Participants were invited to participate via email, stressing the formative purpose and use of the evaluations and the confidential and voluntary character of participation.

System for Evaluation of Teaching Qualities (SETQ)

We used the System for Evaluation of Teaching Qualities (SETQ), which provides surgeons with reliable and valid evaluations of and feedback on their teaching performance in order to improve the quality of teaching in residency training. The SETQ items are theory-based and extensively tested. The items are listed in Appendix table 1. Briefly, the SETQ is composed of two tools (questionnaires), one for surgeons’ self-evaluation and another for resident- evaluation of a surgeon’s teaching performance. The two tools include exactly the same items and were applied web-based. The tools consisted of 26 items. Each item could be rated on a 5 point Likert-scale: 1 = “strongly disagree”, 2 = “disagree”, 3 = “neutral”, 4 = “agree”, 5 = “strongly agree” and there was an additional option “I cannot judge”. The items were statements such as “this surgeon explains why residents are incorrect”. In addition to these numerical items, the tools contained two narrative items: residents could provide “positive attributes of surgeon’s teaching performance” and “suggestions for improvement of surgeons’ teaching performance”. A previous study showed that residents provided surgeons with a median of 11 positive open-text feedback comments and 4 suggestions for improvement per evaluation report. The day after closure of an evaluation period, surgeons received their individual feedback report, summarizing residents’ ratings and narrative comments, along with their self-evaluation. Previous studies indicated that resident-evaluations of surgeon’s teaching performance had high reliability at six to eight resident-evaluations. To preserve the anonymity of the residents, only the number of residents that provided feedback was reported to surgeons. The surgeons were encouraged to discuss their feedback with their peers or program director.
Study Variables

The first variables of interest were surgeons’ self-evaluation and resident-evaluations of surgeons’ teaching performance. To obtain an overall teaching performance score, all SETQ items were averaged. For residents, evaluations were aggregated on surgeon level first. Subsequently, the discrepancy between resident- and self-evaluation was calculated. Previous studies defined the cut-off points for over- and underestimating at half a standard deviation (which corresponds to 0.45-0.50 point across the evaluations in current study).15;19 Although no clear rationalization for this method of selecting cut-off points had been given in the previous studies,15;19 absence of a rationalized alternative led us to adapting this method in current study. Consequently we categorized surgeons who evaluated their performance > 0.5 higher than residents as overestimating, surgeons who evaluated their performance > 0.5 lower than residents as underestimating and as in-agreement if the discrepancy was within +0.5 to -0.5. In addition, a few covariates were included in the analyses, these were: surgeon’s sex, years of experience, teacher training, whether or not surgeons formally discussed the feedback of a previous evaluation, training programs’ specialty and training programs’ hospital.

Analytical Strategies

Initially, we calculated appropriate descriptive statistics. Subsequently, missing data were imputed using multiple imputations (mice package in R statistics).24 We used generalized linear mixed effects growth models to explore how the evaluation scores changed over the three subsequent evaluation periods.25;26 The mixed models framework allowed for adjustment of clustering on individual, specialty and hospital levels.

Next, the effect of over- and underestimating performance on subsequent teaching performance was analyzed using regression analysis. More specifically, sequential g-estimation within generalized linear mixed models framework was used (a technique developed to estimate causal effects with time varying exposures in longitudinal studies).27 The first regression model had resident-evaluated subsequent teaching performance as outcome and included whether surgeons over- or underestimated their previous perfor-
mance as predictor. The second model had surgeons’ self-evaluated subsequent teaching performance as outcome and included whether surgeons over- or underestimated previous performance as predictor. Both models were additionally adjusted for previous teaching performance scores, whether surgeons formally discussed their previous evaluation report, surgeon’s sex, experience, teacher training, residency training programs’ specialty and residency training programs’ hospital. Effect heterogeneity by surgeon’s sex and by surgeons who discussed or did not discuss their previous performance was explored and will be reported in an appendix.

Because this cohort study involved surgeons who were lost to follow-up (because they retired, switched jobs, quit teaching or received no residents’ evaluation), sensitivity analysis for this loss-to-follow-up (or selection or censoring) bias were performed. In this sensitivity analysis, the inverse probability of censoring (IPC) weight was calculated for each surgeon based on a surgeon’s background characteristics and his evaluation scores of previous evaluations. Subsequently, all models described above were re-estimated, now weighting each surgeon by their IPC weight to account for the loss-to-follow-up bias. All analyses were performed using IBM SPSS Statistics 21.0 for windows operating system.

RESULTS

Study participants and response

Of the 351 invited surgeons, 347 (99%), 313 (89%) and 288 (82%) received residents’ feedback during the first, second and third evaluation periods respectively. Self-evaluations were completed by 295 (84%), 249 (71%) and 242 (69%) surgeons during the first, second and third evaluation periods respectively. Residents’ response rates were 84%, 74% and 78% respectively during the three subsequent evaluation periods. Characteristics of surgeons and residents are reported in table 1.
Table 1 Study and participants’ characteristics

<table>
<thead>
<tr>
<th>Evaluation characteristics</th>
<th>Evaluation period 1</th>
<th>Evaluation period 2</th>
<th>Evaluation period 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of surgeons that performed a self-evaluation (% of study population)</td>
<td>295 (84%)</td>
<td>249 (71%)</td>
<td>242 (69%)</td>
</tr>
<tr>
<td>Number of residents that performed evaluations (% of study population)</td>
<td>251 (84%)</td>
<td>256 (74%)</td>
<td>266 (78%)</td>
</tr>
<tr>
<td>Number of surgeons who received a feedback report containing residents’ feedback (% of study population)</td>
<td>347 (99%)</td>
<td>313 (89%)</td>
<td>288 (82%)</td>
</tr>
<tr>
<td>Median number of resident-evaluations per feedback report</td>
<td>7</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Percentage of surgeons who attended a formal teacher training course</td>
<td>65%</td>
<td>81%</td>
<td>86%</td>
</tr>
<tr>
<td>Percentage of surgeons who discussed their feedback following the evaluation</td>
<td>72%</td>
<td>71%</td>
<td>-</td>
</tr>
</tbody>
</table>

Surgeons characteristics

| Surgeon’s age (mean ± SD) | 48.1 (8.2) | - | - |
| Number years of experience at current training program (mean ± SD) | 10.0 (8.4) | - | - |
| Percentage of female surgeons | 33% | - | - |

Resident characteristics

| Percentage of female residents | 53% | 51% | 60% |
| Percentage of residents in residency year 1-2 | 49% | 43% | 39% |
| Percentage of residents in residency year 3-4 | 25% | 27% | 27% |
| Percentage of residents in residency year 5-6 | 25% | 30% | 34% |

Findings

The median score of resident-evaluations of surgeons’ teaching performance increased from 3.83 in the first and 3.82 in the second evalu-
ation period to 3.91 in the third evaluation period \((p<0.001)\) (Table 2, Figure 1). Surgeons’ median self-evaluated teaching performance scores did not change over the three subsequent evaluation periods and the growth models indicated no change (Table 2, Figure 1). There were no differences between the unweighted growth models and the IPC weighted models. 

Overestimating teaching performance resulted in lower subsequent teaching performance as evaluated by both residents (regression coefficient \((b) = -0.08\), 95% confidence limits \((CL) = -0.18, 0.02\)) and surgeons themselves \((b= -0.12, 95\%\ CL = -0.21, -0.02)\). Underestimating performance did not impact resident-evaluated teaching performance \((b= 0.01, 95\%\ CL= -0.08, 0.06)\), while it resulted in enhanced self-evaluated performance \((b= 0.10, 95\%\ CL= 0.03, 0.16)\) (Table 3). The IPC weighted models yielded similar effect estimates and are available through Appendix table 2.

**Figure 1** Median teaching performance scores over three subsequent evaluation periods

![Figure 1](image-url)
Table 2 Median, 20th and 80th percentile scores, marginal means and 95% confidence limits (CL) of residents’ evaluations and surgeon’s self-evaluations for the three subsequent evaluation periods

<table>
<thead>
<tr>
<th>Evaluation period 1</th>
<th>Evaluation period 2</th>
<th>Evaluation period 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median teaching performance score of residents-evaluations (20th, 80th percentile score)</td>
<td>3.83 (3.46, 4.16)</td>
<td>3.82 (3.46, 4.14)</td>
</tr>
<tr>
<td>Mean performance score of residents’ evaluations (95% CL)</td>
<td>3.79 (3.75, 3.84)</td>
<td>3.79 (3.74, 3.84)</td>
</tr>
<tr>
<td>Median teaching performance score of surgeon’s self-evaluations (20th, 80th percentile score)</td>
<td>3.70 (3.44, 3.98)</td>
<td>3.72 (3.40, 4.00)</td>
</tr>
<tr>
<td>Mean teaching performance score of surgeon’s self-evaluations (95% CL)</td>
<td>3.69 (3.64, 3.73)</td>
<td>3.70 (3.66, 3.75)</td>
</tr>
</tbody>
</table>

* = the growth models indicated that the mean score of the third evaluation period was higher compared to the mean scores of the first and second evaluation period by p>0.001

Surgeons’ sex was found to modify the relationship between over- and underestimating teaching performance and subsequent performance. Therefore, the models were re-estimated for male and female surgeons separately (Appendix table 3, Appendix figure 1). No modification by discussion of feedback was found.

**DISCUSSION**

This study showed that residents evaluated surgeons’ teaching performance higher after two cycles of evaluation, feeding back, follow-up and re-evaluation. Surgeons’ self-evaluations of their teaching performance did not alter over the years. Surgeons who overestimated received lower scores by residents on their subsequent teaching performance. Surgeons who underesti
mated, self-evaluated their subsequent teaching performance higher, while surgeons who overestimated self-evaluated their subsequent performance lower.

Surgeons’ teaching performance was enhanced after two cycles of feedback, but not after the first feedback cycle. Several reasons such as lack of time to or low prioritization of changing particular behaviors in response to feedback, could have delayed actual changes in behaviors.2330

**Table 3** Unstandardized regression coefficients (b) and 95% confidence limits (CL) for the associations between (residents' and own) evaluation discrepancy and surgeon's subsequent teaching performance

<table>
<thead>
<tr>
<th>Resident-evaluations of surgeon’s subsequent teaching performance</th>
<th>Surgeon’s own evaluation of subsequent teaching performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>b Lower 95% CL Upper 95% CL p-value</td>
<td>b Lower 95% CL Upper 95% CL p-value</td>
</tr>
<tr>
<td>Overestimated teaching performance at previous evaluation (reference = in-agreement with residents)</td>
<td>-0.08 -0.18 0.02 0.116</td>
</tr>
<tr>
<td>Underestimated teaching performance at previous evaluation (reference = in-agreement with residents)</td>
<td>-0.01 -0.08 0.06 0.692</td>
</tr>
</tbody>
</table>

All models were additionally adjusted for previous teaching performance, teacher training, number of years experience at current training program, residency training programs’ specialty and residency training programs’ hospital
Besides, there may be some distrust in the validity and usefulness of a recently developed evaluation system and surgeons may perceive discomfort with the new process of receiving residents’ feedback. These factors may have impeded surgeons from changing their behaviors after the first feedback cycle. After the second cycle, surgeons – individually as well as a group - were more familiar with the evaluation system and the process of receiving feedback, and may have prioritized changes higher after receiving particular feedback twice.

Surgeons who overestimated their performance had lower subsequent teaching performance as evaluated by residents. As noted earlier, although the regression coefficients are small, they do have potential clinical relevance. Several managerial and psychological studies found similar negative effects of overestimating one’s own performance. The negative effects may be caused by the perceived inaccuracy of the feedback by overestimating surgeons or by other negative (emotional) reactions evoked by overestimating one’s own performance. An alternative explanation for the negative effects of overestimation may be found in the different background characteristics of over-estimators compared to under- and in-agreement estimators. It was proposed that characteristics as gender, experience and age might influence performance (enhancement) more than the overestimation itself. Previous studies identified that over-estimators tended to be older and more likely to be male compared to under- or in-agreement estimators. The modification by surgeons’ sex, as found in this study, also suggests that female surgeons, who are less likely to be over-estimators, had higher subsequent performance compared to male surgeons. With more females entering surgery, the number of overestimating surgeons may decrease in the near future. Underestimation of performance had no influence on subsequent teaching performance as evaluated by residents. This may not be surprising, since most studies in the psychological literature found little differences in performance between under-estimators and in-agreement estimators.

Surgeons who overestimated their teaching performance self-evaluated lower in subsequent evaluations, while surgeons who underestimated rated themselves higher in follow up evaluations. These findings are in line with
previous research showing that peoples’ most obvious reaction towards external performance evaluations that disagree with self-evaluations of performance, is to converge their self-evaluations in a follow-up evaluation towards the external ratings.\textsuperscript{15;18;20} These findings can be explained by the self-consistency theory, that states that people seek to minimize the discrepancy between self- and external ratings of performance.\textsuperscript{14}

In line with informed self-assessment theory,\textsuperscript{3} the results of self- and external evaluations should be integrated to draw any conclusions about the (enhancement of) performance of individual surgeons. We suggest that, at least, resident- and self-evaluated performance are considered when interpreting the performance of individual surgeons, especially since we know that these two evaluations tend to be complementary, not identical.\textsuperscript{5,34} This study involved all clinician teachers of 29 residency training programs of 13 teaching hospitals. The participation rates were high, loss to follow-up was limited to only 17% over three years and several potential sources of bias (including loss to follow-up bias) were addressed in the data analyses and contributed to the robustness of this study’s findings. The cut-off scores for over- and underestimating applied in this study were arbitrary, although they were similar to previous studies on this topic.\textsuperscript{15;19} Further, there was no uniform procedure for the discussion of the feedback. Therefore, modification by discussion of feedback and adjustment of the regression analyses could only be performed for the variable if the feedback was formally discussed and not how the feedback was discussed. The results of this study suggest that changing performance takes time and therefore, it will be interesting to study if a surgeon’s performance will be even further enhanced after a third, fourth or fifth evaluation cycle. Future studies will explore the effects of evaluation over a longer follow-up period. Because the self-evaluated performance remained stable while resident-evaluated performance was enhanced, fewer surgeons were overestimating their performance after two SETQ cycles. Given the finding that overestimating performance negatively impacted subsequent performance, this trend is probably beneficial for surgeon’s subsequent performance after more than two SETQ cycles.
Knowledge about whether surgeons over- or underestimated their teaching performance can be important to guide the follow-up once the feedback is received. Because surgeons who overestimated their performance were more likely to have lowered subsequent teaching performance, specific guidance and support in the reflection process can probably help these surgeons in their interpretation of, and reactions after receiving, the feedback. For this purpose, structured reflection methods that take surgeon’s individual emotions and the specific content of the feedback into account, may help surgeons in appreciating their performance evaluation feedback.\(^{35}\) However, more research is needed to explore if tailored guidance and support, for surgeons who over-, under- or in-agreement estimated their performance, for male and female surgeon, can enhance subsequent performance.

**ACKNOWLEDGEMENTS**

This project has been supported by the foundation “De Drie Lichten” in The Netherlands. The authors would like to thank all surgeons and residents for their participation in this study. We also want to thank Anne Huitema for her assistance with the data preparation and analysis and MEDOX.nl for the maintenance and optimization of the web-based SETQ application.
REFERENCES


