World J Surg (2014) 38:2761–2769
DOI 10.1007/s00268-014-2655-3

World Journal of Surgery

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

Published online: 28 May 2014

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 five-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 with the first (median score: 3.83, 20th to 80th percentile score: 3.46–4.16) and second (median: 3.82, 20th to 80th: 3.46–4.14) evaluation period, residents evaluated surgeon's teaching performance higher during the third evaluation period (median: 3.91, 20th to 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.

B. C. M. Boerebach (✉) · O. A. Arah ·
M. J. Heineman · K. M. J. M. H. Lombarts
Center for Evidence-Based Education, Academic Medical
Center, University of Amsterdam, Meibergdreef 9,
PO Box 22700, 1100 DE Amsterdam, The Netherlands
e-mail: b.c.boerebach@amc.uva.nl

O. A. Arah
Department of Epidemiology, School of Public Health,
University of California, Los Angeles (UCLA), Los Angeles,
CA, USA

O. A. Arah
UCLA Center for Health Policy Research, Los Angeles, CA,
USA

M. J. Heineman
Academic Medical Center, University of Amsterdam,
Amsterdam, The Netherlands

O. R. C. Busch
Department of Surgery, Academic Medical Center, University
of Amsterdam, Amsterdam, The Netherlands

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 [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 [3]. Several feedback sources 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 [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 [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 [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 in, or even dismissal of, self-improvement actions [8–11]. In particular, 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 [8–13]. Furthermore, 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 [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 explore how resident evaluations and self-evaluations of surgeons' teaching performance evolve after two cycles of evaluation, reporting, and feedback. Second, we explore whether 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 and 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 (more than five residents) surgical training programs that were based at or coordinated from the project leaders' academic medical center participated in this study (24 of the 29 programs included in this study). Additionally, five training programs from other regions in the Netherlands participated. Data were collected from September 2008 until May 2013 and occurred during annual evaluation periods lasting 1 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 period. 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 [5, 6, 21, 22]. The items are listed in Appendix Table 4. 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 via the Internet. The tools consisted of 26 items [5, 6]. Each item could be rated on a five-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 "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 four suggestions for improvement per evaluation report [23]. 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 have indicated that resident evaluations of surgeon's teaching performance had high reliability, at six to eight resident evaluations [5, 6]. 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 first aggregated at the surgeon level. Subsequently, the discrepancy between resident evaluation 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 the current study) [15, 19]. Although no clear rationalization for this method of selecting cut-off points was given in the previous studies [15, 19], absence of a rationalized alternative led us to adapt this method in the 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: 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 a 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 the outcome and included whether surgeons over- or underestimated their previous performance as predictor. The second model had surgeons' self-evaluated subsequent teaching performance as the 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 is 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 resident evaluations), 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/her evaluation scores of previous evaluations [28]. 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 (IBM, Armonk, NY, USA).

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.

Findings

The median score of resident evaluations of surgeons' teaching performance increased from 3.83 in the first and 3.82 in the second evaluation period to 3.91 in the third

Table 1 Study and participant characteristics

	Evaluation period		
	1	2	3
Evaluation characteristics [*n* (%)]			
Surgeons who performed a self-evaluation (% of study population)	295 (84)	249 (71)	242 (69)
Residents who performed evaluations (% of study population)	251 (84)	256 (74)	266 (78)
Surgeons who received a feedback report containing residents' feedback (% of study population)	347 (99)	313 (89)	288 (82)
Resident evaluations per feedback report (median)	7	6	6
Surgeons who attended a formal teacher training course (%)	65	81	86
Surgeons who discussed their feedback following the evaluation (%)	72	71	–
Surgeon characteristics			
Surgeon's age (mean ± SD)	48.1 (8.2)	–	–
Years of experience at current training program (mean ± SD)	10.0 (8.4)	–	–
Female surgeons (%)	33	–	–
Resident characteristics			
Female residents (%)	53	51	60
Residents in residency year 1–2 (%)	49	43	39
Residents in residency year 3–4 (%)	25	27	27
Residents in residency year 5–6 (%)	25	30	34

SD standard deviation

Table 2 Median, 20th and 80th percentile scores, marginal means, and 95 % CL of resident evaluations and surgeon self-evaluations for the three subsequent evaluation periods

	Evaluation period		
	1	2	3
Median teaching performance score of resident evaluations (20th, 80th percentile score)	3.83 (3.46, 4.16)	3.82 (3.46, 4.14)	3.91 (3.59, 4.27)
Mean performance score of residents' evaluations (95 % CL)	3.79 (3.75, 3.84)	3.79 (3.74, 3.84)	3.91 (3.86, 3.96)[a]
Median teaching performance score of surgeon self-evaluations (20th, 80th percentile score)	3.70 (3.44, 3. 98)	3.72 (3.40, 4.00)	3.70 (3.43, 3.99)
Mean teaching performance score of surgeon self-evaluations (95 % CL)	3.69 (3.64, 3.73)	3.70 (3.66, 3.75)	3.70 (3.66, 3.75)

CL confidence limit

[a] The growth models indicated that the mean score of the third evaluation period was higher than the mean scores of the first and second evaluation periods by $p > 0.001$

evaluation period ($p < 0.001$) (Table 2; Fig. 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; Fig. 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 in Appendix Table 5.

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 6; Fig. 2). No modification by discussion of feedback was found.

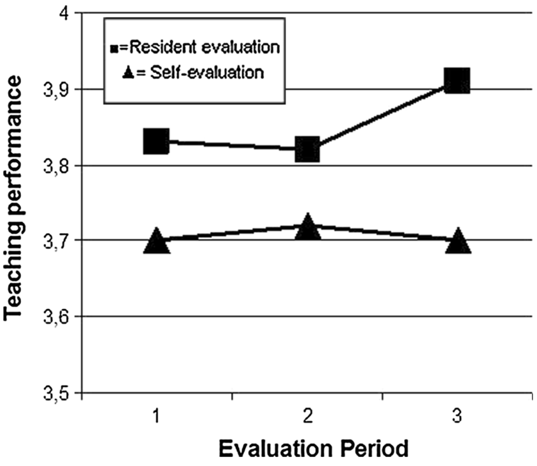

Fig. 1 Median teaching performance scores over three subsequent evaluation periods

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 underestimated, self-evaluated their subsequent teaching performance higher, while surgeons who overestimated self-evaluated their subsequent performance lower.

Surgeons' teaching performance was evaluated higher by residents after two cycles of evaluation, reporting, and feeding back. This finding suggests that feedback can be helpful for teaching performance enhancement. Feedback is often used to guide surgeons' development and to enhance surgeons' performance, and this study provides further empirical evidence that feedback systems can be effective in enhancing performance. Although the changes in performance are limited, a recent Cochrane review concluded that even these small changes have the potential to actually change performance in practice [29]. 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 [23, 30]. Furthermore, 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 [23]. 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 [29]. Several managerial and psychological studies found similar negative effects of overestimating own performance [15–17, 19]. 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 own performance [10, 11, 17, 31, 32]. An alternative explanation for the negative effects of overestimation may be found in the different background characteristics of over-estimators compared with under- and in-agreement estimators [19]. It was proposed that characteristics such as sex, experience, and age might influence performance (enhancement) more than the overestimation itself. Previous studies

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

	Resident evaluations of surgeon's subsequent teaching performance				Surgeon's own evaluation of subsequent teaching performance			
	b	Lower 95 % CL	Upper 95 % CL	*p* value	*b*	Lower 95 % CL	Upper 95 % CL	*p* value
Overestimated teaching performance at previous evaluation (reference = in-agreement with residents)	−0.08	−0.18	0.02	0.116	−0.12	−0.21	−0.02	0.015
Underestimated teaching performance at previous evaluation (reference = in-agreement with residents)	−0.01	−0.08	0.06	0.692	0.10	0.03	0.16	0.003

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

CL confidence limit

identified that over-estimators tended to be older and more likely to be male than are under- or in-agreement estimators [19, 33]. 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 than did 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 [15, 16, 19].

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 [15, 18, 20]. These findings can be explained by the self-consistency theory, which states that people seek to minimize the discrepancy between self- and external ratings of performance [14].

In line with informed self-assessment theory [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 is considered when interpreting the performance of individual surgeons, especially since we know that these two evaluations tend to be complementary, not identical [5, 34].

This study involved all attending surgeons of 29 residency training programs of 13 teaching hospitals. The participation rates were high, loss to follow-up was limited to only 17 % over 3 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 those of previous studies on this topic [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 whether 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 surgeons, can enhance subsequent performance.

Acknowledgments 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 http://medox.nl for the maintenance and optimization of the web-based SETQ application.

Funding This study is part of the research project Quality of Clinical Teachers and Residency Training, 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. Funders had no role in the study design, data collection, data analysis, data interpretation, or writing of the report.

Disclosures None.

Ethical Approval Waiver of ethical approval was provided by the institutional review board of the Academic Medical Center of the University of Amsterdam, Amsterdam, The Netherlands.

Appendix

See Tables 4, 5, 6 and Fig. 2.

Table 4 Scales and items of the SETQ

Item no. *Learning climate*	Scale and items[a]
LC1	Encourages residents to participate actively in discussions
LC2	Stimulates residents to bring up problems
LC3	Motivates residents to study further
LC4	Stimulates residents to keep up with the literature
LC5	Prepares well for teaching presentations and talks
LC6	Creates educational time on the outpatients and surgical department
Professional attitude towards and support of residents	
PA1	Listens attentively to residents
PA2	Is respectful towards residents
PA3	Is easily approachable during on-calls
PA4	Is easily approachable for consultation on the outpatients
Communication of goals	
CG1	States learning goals clearly
CG2	States relevant goals
CG3	Prioritizes learning goals
CG4	Repeats stated learning goals periodically
CG5	Offers to conduct mini-CEX (clinical examination exercise) regularly
Evaluation of residents	
ER1	Evaluates residents' specialty knowledge regularly
ER2	Evaluates residents' analytical abilities regularly
ER3	Evaluates residents' application of knowledge to specific patients regularly
ER4	Evaluates residents' medical skills regularly
ER5	Evaluates residents' surgical skills regularly
Feedback	
FB1	Regularly gives positive feedback to residents
FB2	Gives corrective feedback to residents
FB3	Explains why residents are incorrect
FB4	Offers suggestions for improvement
FB5	Teaches surgical skills in the operating theatre
FB6	Provides constructive criticism about surgical skills during the operation

[a] The items shared the same subject "during my residency in surgery, my attending surgeon generally…" (residents' evaluation of attending surgeons) or "in my role as an attending surgeon, I generally…" (surgeons' self-evaluation)

Table 5 Regression coefficients and 95 % CLs for the associations between the key predictors and subsequent teaching performance, weighted by inverse-probability-of-censoring weights

	Resident evaluations of surgeon's subsequent teaching performance				Surgeon's own evaluation of subsequent teaching performance			
	b	Lower 95 % CL	Upper 95 % CL	*p* value	*b*	Lower 95 % CL	Upper 95 % CL	*p* value
Overestimated teaching performance at previous evaluation (reference = in-agreement with residents)	−0.08	−0.18	0.02	0.110	−0.12	−0.21	−0.02	0.018
Underestimated teaching performance at previous evaluation (reference = in-agreement with residents)	−0.01	−0.08	0.06	0.692	0.10	0.03	0.16	0.003

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

CL confidence limit

Table 6 Regression coefficients and 95 % CLs for the associations between the key predictors and subsequent teaching performance, stratified by gender

	Resident-evaluations of surgeon's subsequent teaching performance				Surgeon's own evaluation of subsequent teaching performance			
	b	Lower 95 % CL	Upper 95 % CL	*p* value	*b*	Lower 95 % CL	Upper 95 % CL	*p* value
Male surgeons								
Overestimated teaching performance at previous evaluation (reference = in-agreement with residents)	−0.07	−0.17	0.04	0.211	−0.15	−0.25	−0.05	0.005
Underestimated teaching performance at previous evaluation (reference = in-agreement with residents)	0.04	−0.06	0.14	0.444	0.07	−0.02	0.16	0.108
Female surgeons								
Overestimated teaching performance at previous evaluation (reference = in-agreement with residents)	−0.10	−0.33	0.14	0.412	−0.03	−0.21	0.14	0.707
Underestimated teaching performance at previous evaluation (reference = in-agreement with residents)	−0.06	−0.18	0.05	0.254	0.12	0.03	0.21	0.012

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

CL confidence limit

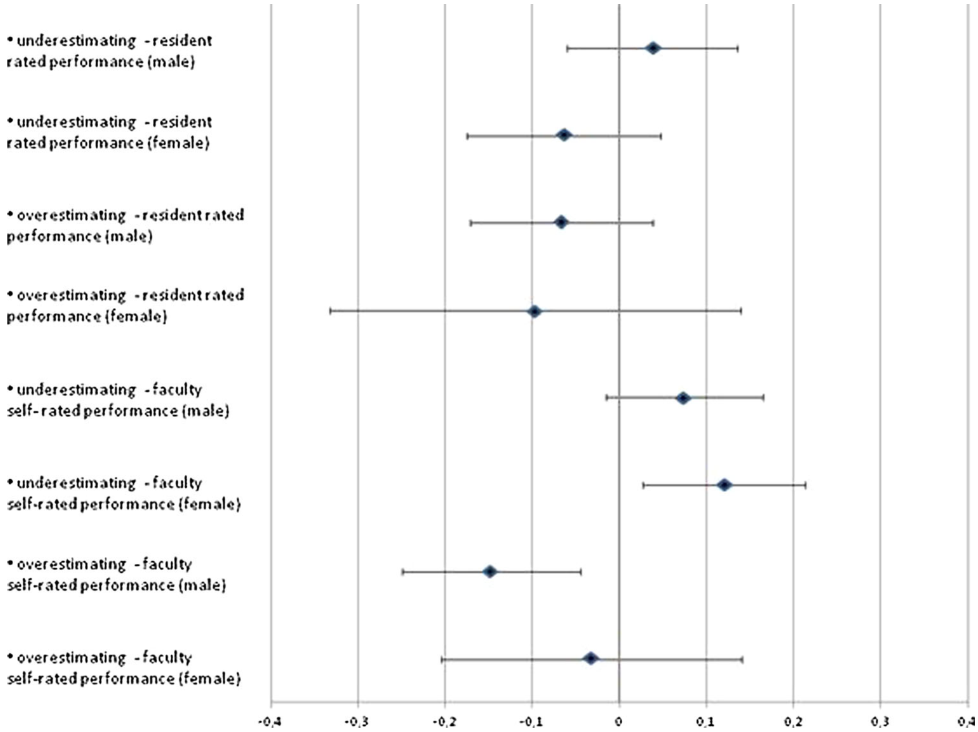

Fig. 2 Sex-specific effects of surgeon under- or overestimating own teaching performance on subsequent resident and own evaluations of teaching performance

References

1. Eva KW, Regehr G (2005) Self-assessment in the health professions: a reformulation and research agenda. Acad Med 80(10 Suppl):S46–S54
2. Eva KW, Regehr G (2008) "I'll never play professional football" and other fallacies of self-assessment. J Contin Educ Health Prof 28(1):14–19
3. Sargeant J, Armson H, Chesluk B et al (2010) The processes and dimensions of informed self-assessment: a conceptual model. Acad Med 85(7):1212–1220
4. Stalmeijer RE, Dolmans DH, Wolfhagen IH et al (2010) Combined student ratings and self-assessment provide useful feedback for clinical teachers. Adv Health Sci Educ Theory Pract 15(3):315–328

5. Boerebach BC, Arah OA, Busch OR et al (2012) Reliable and valid tools for measuring surgeons' teaching performance: residents' vs. self evaluation. J Surg Educ 69(4):511–520
6. van der Leeuw R, Lombarts K, Heineman MJ et al (2011) Systematic evaluation of the teaching qualities of obstetrics and gynecology faculty: reliability and validity of the SETQ tools. PLoS One 6(5):e19142
7. Beckman TJ, Ghosh AK, Cook DA et al (2004) How reliable are assessments of clinical teaching? A review of the published instruments. J Gen Intern Med 19(9):971–977
8. Eva KW, Armson H, Holmboe E et al (2012) Factors influencing responsiveness to feedback: on the interplay between fear, confidence, and reasoning processes. Adv Health Sci Educ Theory Pract 17(1):15–26
9. Mann K, van der Vleuten C, Eva K et al (2011) Tensions in informed self-assessment: how the desire for feedback and reticence to collect and use it can conflict. Acad Med 86(9):1120–1127
10. van der Leeuw R, Slootweg IA, Heineman MJ et al (2013) Explaining how faculty act upon residents' feedback to improve their teaching performance. Med Educ 47:1089–1098
11. van Roermund T, Schreurs ML, Mokkink H et al (2013) Qualitative study about the ways teachers react to feedback from resident evaluations. BMC Med Educ 13(1):98
12. Bandura A (1977) Self-efficacy: toward a unifying theory of behavioral change. Psychol Rev 84(2):191–215
13. Kluger AN, DeNisi A (1996) The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. Psychol Bull 119(2):254–284
14. Korman AK (1970) Toward an hypothesis of work behavior. J Appl Psychol 54(1):31–41
15. Atwater L, Roush P, Fischthal A (1995) The influence of upward feedback on self- and follower ratings of leadership. Pers Psychol 48(1):35–59
16. Atwater LE, Ostroff C, Yammarino FJ, Fleenor JW (1998) Self-other agreement: does it really matter? Pers Psychol 51(3):577–598
17. Brett JF, Atwater L (2001) 360° Feedback: accuracy, reactions, and perceptions of usefulness. J Appl Psychol 86(5):930–942
18. Johnson JW, Ferstl KL (1999) The effects of interrater and self-other agreement on performance improvement following upward feedback. Pers Psychol 52(2):271–303
19. Ostroff C, Atwater LE, Feinberg BJ (2004) Understanding self-other agreement: a look at rater and ratee characteristics, context, and outcomes. Pers Psychol 57(2):333–375
20. Smither JW, London M, Vasilopoulos NL, Reilly RR, Millsap RE, Salvemini N (1995) An examination of the effects of an upward feedback program over time. Pers Psychol 48(1):1–34
21. Arah OA, Hoekstra JBL, Bos AP et al (2011) New tools for systematic evaluation of teaching qualities of medical faculty: results of an ongoing multi-center survey. PLoS One 6(10):e25983
22. Lombarts KM, Bucx MJ, Arah OA (2009) Development of a system for the evaluation of the teaching qualities of anesthesiology faculty. Anesthesiology 111(4):709–716
23. van der Leeuw RM, Overeem K, Arah OA et al (2013) Frequency and determinants of residents' narrative feedback on the teaching performance of faculty: narratives in numbers. Acad Med 88:1324–1331
24. Van Buuren S, Groothuis-Oudshoorn K (2011) MICE: multivariate imputation by chained equations in R. J Stat Softw 45(3):1–67
25. Singer JD (1998) Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. J Educ Behav Stat 23(4):323–355
26. Twisk JWR (2006) Applied multilevel analysis: a practical guide. Cambridge University Press, Cambridge
27. Robins JM, Hernán MA (2009) Estimation of the causal effects of time-varying exposures. Longitudinal data analysis. CRC Press, Boca Raton, pp 553–599
28. Robins JM, Hernan MA, Brumback B (2000) Marginal structural models and causal inference in epidemiology. Epidemiology 11(5):550–560
29. Ivers N, Jamtvedt G, Flottorp S et al (2012) Audit and feedback: effects on professional practice and healthcare outcomes. Cochrane Database Syst Rev 6:CD000259
30. Overeem K, Wollersheim H, Driessen E et al (2009) Doctors' perceptions of why 360° feedback does (not) work: a qualitative study. Med Educ 43(9):874–882
31. Sargeant J, Mann K, Sinclair D et al (2008) Understanding the influence of emotions and reflection upon multi-source feedback acceptance and use. Adv Health Sci Educ Theory Pract 13(3):275–288
32. Sargeant JM, Mann KV, van der Vleuten CP et al (2009) Reflection: a link between receiving and using assessment feedback. Adv Health Sci Educ Theory Pract 14(3):399–410
33. Brutus S, Fleenor JW, McCauley CD (1999) Demographic and personality predictors of congruence in multisource ratings. J Manag Dev 18:417–435
34. Davis DA, Mazmanian PE, Fordis M et al (2006) Accuracy of physician self-assessment compared with observed measures of competence: a systematic review. JAMA 296(9):1094–1102
35. Sargeant J, McNaughton E, Mercer S et al (2011) Providing feedback: exploring a model (emotion, content, outcomes) for facilitating multisource feedback. Med Teach 33(9):744–749