Anesthesiology 2010; 113:693–703

Clinical Teaching Improves with Resident Evaluation and Feedback

Keith Baker, M.D., Ph.D.*

ABSTRACT

Background: The literature is mixed on whether evaluation and feedback to clinical teachers improves clinical teaching. This study sought to determine whether resident-provided numerical evaluation and written feedback to clinical teachers improved clinical teaching scores.

Methods: Anesthesia residents anonymously provided numerical scores and narrative comments to faculty members who provided clinical teaching. Residents returned 19,306 evaluations between December 2000 and May 2006. Faculty members received a quantitative summary report and all narrative comments every 6 months. Residents also filled out annual residency program evaluations in which they listed the best and worst teachers in the department.

Results: The average teaching score for the entire faculty rose over time and reached a plateau with a time constant of approximately 1 yr. At first, individual faculty members had average teaching scores that were numerically diverse. Over time, the average scores became more homogeneous. Faculty members ranked highest by teaching scores were also most frequently named as the best teachers. Faculty members ranked lowest by teaching scores were most frequently named as the worst teachers. Analysis of ranks, differential improvement in scores, and a decrease in score diversity effectively ruled out simple score inflation as the cause for increased scores. An increase in teaching scores was most likely due to improved teaching.

Conclusions: A combination of evaluation and feedback, including comments on areas for improvement, was related to a substantial improvement in teaching scores. Clinical teachers are able to improve by using feedback from residents.

What We Already Know about This Topic

- Feedback is important to improved teaching, yet there are no long-term studies examining resident feedback on anesthesiologist teaching.

What This Article Tells Us That Is New

- Over a 5-yr period, residents provided qualitative and quantitative anonymous evaluations of teaching faculty.
- Institution of this feedback system was associated with increased teaching scores for the faculty.

RESIDENCY programs aspire to improve clinical teaching provided by clinician educators. One strategy used to improve clinical teaching is to obtain resident evaluations of the teachers. To date, evaluations provided by residents and medical students remain most common. The effect of evaluations on teaching has been mixed. Some studies demonstrate an increase in clinical teaching scores after written feedback,[1–3] whereas feedback in the form of simple numerical ratings results has not improved teaching scores.[4–8] Some studies have been underpowered to find a difference in teaching scores,[7,9,10] whereas others may fail to show improvement because of a ceiling effect.[7,9,11] The literature is also largely silent on the issue of the time needed to improve clinical teaching. Concerns have been raised about the reliability and validity of resident and student evaluations of both clinical and classroom teaching.[12,13] Additional evidence is needed demonstrating that resident evaluation and feedback either does or does not lead to durable improvements in clinical teaching.

Feedback is fundamental to performance improvement.[14] Recent studies have repeatedly shown that self-assessment can be remarkably flawed, with the worst performers most seriously overestimating their skills.[15–18] Claridge *et al.*[19] studied surgeon self-evaluation of teaching and compared it

◆ This article is accompanied by an Editorial View: Please see: Schwartz AJ: Resident/fellow evaluation of clinical teaching: An essential ingredient of effective teacher development and educational planning. ANESTHESIOLOGY 2010; 113:516–7

* Assistant Professor of Anaesthesia, Harvard Medical School, and Assistant Anesthetist, Department of Anesthesia, Critical Care and Pain Medicine, Massachusetts General Hospital, Boston, Massachusetts.

Received from the Department of Anesthesia, Critical Care and Pain Medicine, Massachusetts General Hospital, Boston, Massachusetts. Submitted for publication March 10, 2010. Accepted for publication May 19, 2010. Support was provided solely from institutional and/or departmental sources.

Address correspondence to Dr. Baker: Department of Anesthesia, Critical Care and Pain Medicine, Massachusetts General Hospital, 55 Fruit Street, Jackson 4, Boston, Massachusetts 02114. khbaker@partners.org.

with resident evaluation of their teaching. None of the surgeons who received below average teaching scores self-identified these deficiencies. It is noteworthy that faculty members who declined to engage in self-evaluation had lower teaching scores than those who volunteered to participate. The positive effects of immediate feedback to lecturers was demonstrated by improved teaching scores after second-year medical students provided feedback with numerical ratings and narrative comments.[20]

The current study provides a long-term (5.5 yr) examination of the influence of resident evaluation and feedback on the clinical teaching faculty and strongly supports the conclusion that resident evaluation and feedback increases clinical teaching scores. The report includes data on the kinetics of improvement and a novel form of construct validity relating to teaching scores. The data also effectively rule out the possibility that teaching scores increased as a result of simple grade inflation of the scores given by our residents.

Materials and Methods

Evaluation System

Resident Evaluation and Feedback Regarding Clinical Teaching. We developed an evaluation system to capture anonymous resident feedback regarding faculty members engaged in intraoperative and perioperative clinical teaching. Each month, our computerized billing database determined which resident has worked with which attending physician. For rotations without billing information (Obstetrics, Intensive Care Unit, Pain Rotation, Preadmission Testing Area, and Postanesthesia Care Unit), we used schedules to create the resident-attending physician matches. Each unique resident-attending physician pair results in a request for the resident to anonymously evaluate that faculty member. The paper-based evaluation form lists seven different attributes of teaching: overall, time spent, clinical supervision, quality of teaching, quantity of teaching, role model, and encourages thinking about the science of anesthesia. Each question was rated using a Likert scale ranging from 0 to 10. Zero denotes the worst teaching and 10 denotes the best teaching. Teaching scores are formed by summing up the seven subscores, and thus teaching scores ranged from 0 to 70. Each form requests narrative comments in three areas—strengths, areas that need improvement, and additional comments. Residents were told that whenever they give low scores that they should provide a specific comment regarding what they would like the faculty member to start doing, do more, or stop doing to improve their teaching. During the last 2 months of each 6-month period, residents who had not completed any evaluations were contacted by letter and encouraged to complete and submit their evaluations. Approximately 89% of our clinical rotations occur at the Massachusetts General Hospital. Evaluations pertain only to Massachusetts General Hospital faculty members. The Massachusetts General Hospital Institutional Review Board waived the need for informed consent and classified this study as exempt.

Analysis and Reports of Faculty Member Teaching. Numerical results and verbatim comments from each evaluation were keyed into an electronic database by a single person. Every 6 months, the data were analyzed, and individual reports were prepared for each faculty member who had at least two evaluations. The report provides the faculty member with an average score for each of the seven areas of teaching. They are also provided with an overall composite teaching score, which is the sum of the seven subscores. Reports contain the teaching score of the "average faculty member." The "average faculty member" is represented by the average of all data collected during the 6-month period and includes the average score for each of the seven subscores as well as the overall composite teaching score. Any significant differences between the individual and the average faculty member were highlighted for both subscores and the overall composite teaching score. Reports also contained a graphical comparison of the individual faculty member's composite score compared with all other faculty members. Resident comments pertaining to individual faculty members were included with each individual report.

Each 6-month time window is referred to as a period. Each period was numbered sequentially. Period 1 was our initial or baseline use of this evaluation system and refers to the 6-month time window from December 1, 2000 to May 31, 2001. Period 2 refers to the subsequent 6 months and so forth. During the first six periods, individual reports also contained the explicit rank of each attending physician (*e.g.*, rank 33 of 125). Explicit reporting of rank was stopped after period 6 because scores were so similar that rank differences were largely meaningless. Relative ranks for a period were determined by dividing a rank number by the total number of faculty evaluated in that period. Thus, relative ranks begin near 0 (top ranked person) and progress to 1 (lowest ranked person). Relative ranks allow rank positions to be compared across periods having different numbers of faculty. During the first six periods, faculty members were asked to speak with the chairman if they had both very low scores and negative comments. These few faculty members were encouraged to improve their teaching by meeting with a single senior faculty member who was experienced in faculty development and education. Approximately three to four faculty members per period took advantage of this offer, but we have no formal record of this activity because their involvement was voluntary. Teaching scores and comments were not otherwise used for individual faculty development programs, entitlements (*i.e.*, travel or academic time), or teaching assignments. We did use teaching scores in part to help identify the best teachers to guide annual bonus distribution. The faculty was not made aware of the metrics that went into this decision, and this was not a formalized program. Except for the reports that were distributed every 6 months, faculty members received no further information regarding their teaching. Thus faculty members were provided repeated

Table 1. Summary Table

Period	No. of Evaluations	No. of Attendings	No. of Residents	Percentage Completed	Avg No. of Evals/Attending	All Scores and SD for Each Period: All Teaching Scores Mean	All Teaching Scores SD	All Teaching Scores SEM	Group Scores Mean	Group Scores SD
1	1,206	92	42	39.1	13.1	54.57	13.21	0.38	53.65	8.76
2	1,708	106	61	48.7	16.1	56.05	12.39	0.30	55.07	6.69
3	1,985	117	58	57.5	17.0	57.77	11.26	0.25	56.34	5.41
4	1,449	113	59	40.5	12.8	58.75	11.08	0.29	58.43	6.30
5	1,486	117	54	41.9	12.7	59.06	9.74	0.25	58.45	5.55
6	1,886	112	63	49.3	16.8	60.21	9.13	0.21	60.09	4.47
7	1,579	104	44	42.1	15.2	59.96	9.90	0.25	59.25	4.76
8	1,579	119	64	39.3	13.3	59.32	9.47	0.24	58.75	4.89
9	2,044	109	66	51.9	18.8	61.28	8.86	0.20	60.96	4.47
10	2,022	116	59	47.3	17.4	59.10	9.86	0.22	58.84	4.18
11	2,362	121	62	58.6	19.5	58.97	10.08	0.21	58.73	4.12

The mean of "All Teaching Scores" refers to the average of all teaching scores given during a period. Percentage completed refers to the percentage of all evaluations in each period that were completed. The SD refers to the variation of all teaching scores. The group means were determined by averaging the average teaching score for each attending. The group score SD refers to the variation in mean teaching scores earned by different attending physicians.

rounds of feedback and essentially were allowed to decide for themselves how to interpret the results and how to improve their teaching.

End-of-Year Resident Survey Listing Best and Worst Teachers. Toward the end of each academic year, we anonymously survey our residents regarding a wide variety of issues. Among the questions is a request to list the best and worst teachers in the department. The number of times a faculty member was listed was converted into a frequency histogram and plotted as a function of that person's relative rank as determined using teaching scores over that same academic year. Histogram counts were determined independently from teaching scores.

Statistical Analysis

Scores in different periods are compared by way of unpaired t tests assuming unequal variances. Exponential fits were determined using a Levenberg-Marquardt method. All statistics were determined using StatsDirect (version 2.6.6; StatsDirect Ltd., Cheshire, United Kingdom), Excel 2003 (Microsoft, Redmond, WA), or Origin (version 7.5 SR4; OriginLab Corp., Northampton, MA). Effect sizes were determined by Cohen d values which are calculated as the difference in means divided by the combined SD of the data. Effect sizes provide a measure of the size of a difference compared with the variation in the data. Effect sizes are classified as small (Cohen $d = 0.2$), medium (Cohen $d = 0.5$), or large (Cohen $d = 0.8$).[21] Cronbach α was used to examine reliability between subscores. Rank data are compared using Kendall τ. Kendall τ is used to determine whether two rank orders are the same. When two rank orders are identical, Kendall τ is 1.0; if they are perfectly inversely related, it is -1.0; and if the rank orders bear no relationship to each other, then it is 0.0. P values are two-sided and determined exactly whenever possible. Data points in graphs are means ± SEM unless noted otherwise.

Results

Teaching Scores Increased after Implementing an Evaluation and Feedback Process

During the 5.5 yr of this study, a total of 19,306 evaluations were returned by 194 different residents concerning 197 different faculty members. Table 1 shows the number of evaluations, residents, and faculty members during each 6-month period. The overall Cronbach α measure for internal consistency for all seven subscores over 19,306 evaluations was 0.980. This high Cronbach α strongly suggests that residents generally use each of the subscales in an interchangeable way; thus, it is unlikely that there is enough unique information in the individual subscores to allow meaningful comparisons. Because the subscales were used only to compute a single teaching score, they were not analyzed further.

All individual teaching scores are shown for periods 1 and 7 (fig. 1). Period 1 represents the baseline distribution of scores, and period 7 is representative of all later periods. Teaching scores during period 1 decreased over the first 80% of the faculty (until relative rank 0.8) and then declined more rapidly. In period 7, approximately 3 yr later, teaching scores declined less rapidly as one went down the rank order. The mean teaching score in period 7 was higher than in period 1. The average teaching score increased from period 1 up until approximately period 6 (fig. 2A, solid circles). The overall difference in teaching scores between periods 1 and 6 was significant ($P = 5 \times 10^{-37}$). The effect size, Cohen d, for the change in scores between periods 1 and 6 was 0.50, a medium-sized effect. On a 0–10 scale, this corresponds to a change of 0.8 (from 7.8 up to 8.6). To remove concern that the increase in teaching scores was due to a changing faculty composition, only the scores of the 50 faculty members who were present for all 11 periods were examined. The teaching scores of these 50 persis-

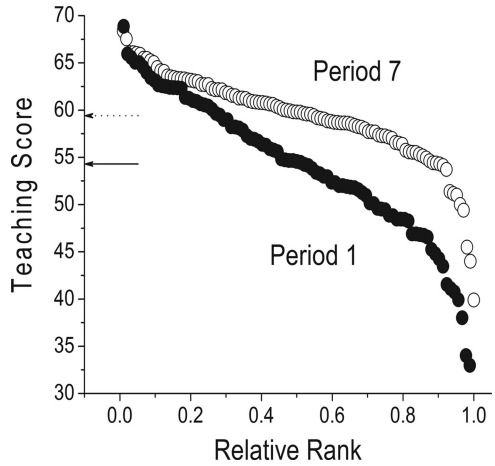

Fig. 1. Teaching scores become higher and more similar after evaluation and feedback. Teaching scores for each faculty member in period 1 (*solid circles*) and period 7 (*open circles*) are plotted as a function of relative rank. The overall average for period 1 is shown by the *solid arrow* and period 7 is shown by the *broken arrow*.

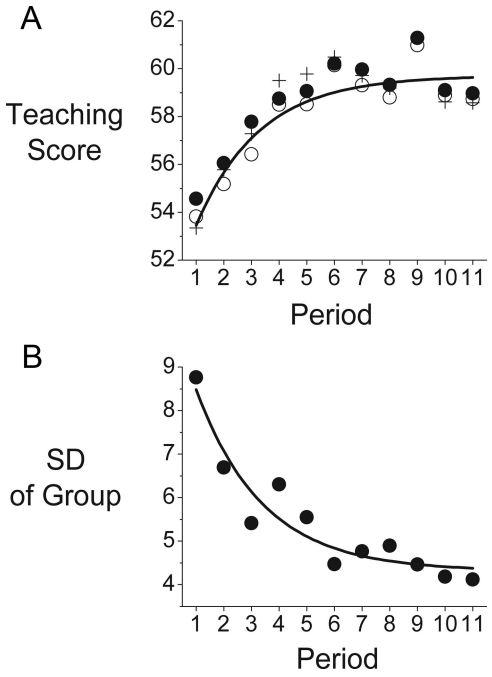

Fig. 2. (*A*) Teaching scores increase over time and reach a plateau. The average teaching score determined from all evaluations for each period is shown by *solid circles*. The average score of all the faculty members evaluated in each period is shown by *open circles*. The average score of the 50 faculty members who were present for all 11 periods is shown as a plus sign. For clarity, only the error bars for the average teaching score using all evaluations are displayed. The overlaid exponential curve was fit to the average teaching score from all evaluations for each period. The best fit parameters included an initial score of 54.11 (after period 1), a final score of 59.95 ± 0.12, and a time constant of 0.94 ± 0.11 years (because each period is 6 months, this time constant is equivalent to 1.87 ± 0.21 periods). The fit has an r^2 value of 0.76. (*B*) Faculty members' teaching scores become more similar over time. The SDs of the group mean scores are shown for each period. The overlaid exponential curve was fit to the SD for each period. The best fit parameters included an initial SD of 8.48 (after period 1), a final SD of 4.31 ± 0.33, and a time constant of 1.21 ± 0.37 years (because each period is 6 months, this time constant is equivalent to 2.41 ± 0.73 periods). The fit has an r^2 value of 0.90.

tent faculty members were very similar to the overall teaching scores for all faculty members (fig. 2A, +).

The Time Course for Improvement

The time frame over which the teaching scores increased was well described by an exponential curve with a time constant of 0.94 ± 0.11 yr (fig. 2A). The change in teaching scores was nearly complete after three time constants, which corresponds to period 6 or 7.

Faculty Members' Scores Became More Homogenous

Over the same time frame that the overall teaching scores were increasing, faculty members' average scores were becoming more similar. In period 1, the teaching scores were broadly distributed and covered a wide range as the relative rank order is descended (fig. 1). In contrast, in period 7, individual average scores occurred over a much narrower range as the relative rank order is descended (fig. 1). The spread in scores was quantified by determining the SD of the average teaching scores of the faculty members in each period. Figure 2B shows the SD of the faculty member's scores as a function of period. The diversity of scores decreases exponentially with a time constant of 1.21 ± 0.37 yr. The change in score diversity (as represented by the group SD) is largely complete by approximately three time constants, which approximately corresponds to period 7. The difference in group score diversity between periods 1 and 9 was significant as determined by an F test on the group variance ($P = 4.9 \times 10^{-11}$). The effect size, Cohen d, for the change between periods 1 and 9 was 0.95, a large effect.

An Independent Determination of Teaching Quality

Our yearly anonymous residency program evaluations ask residents to list (with no limit) the best and worst clinical teachers. This provides an independent and nonnumeric approach to assessing teaching quality. We do not provide this information to the faculty and thus it has no impact on them. The number of times that a faculty member is listed as a "best" or "worst" clinical teacher was counted for each academic year, and these frequency data were plotted against the corresponding relative ranks for these same faculty members over these same time periods. Figure 3 shows that faculty members who had the highest counts for best teacher were also independently ranked the highest using numerical teaching scores. Likewise, faculty members who had the most counts for worst teacher were also ranked lowest based on our numerical teaching scores. The "best" histogram was fit by an exponential function that decayed with a "relative rank" rate of 15.3 ± 0.02%. This means that for every 15.3% reduction in relative rank, the number of times a faculty member was

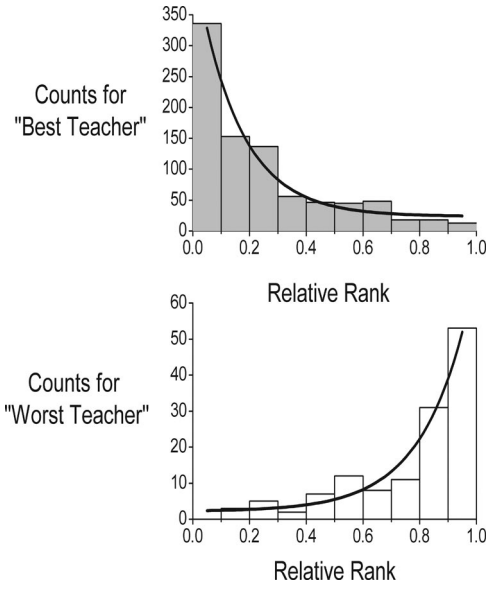

Fig. 3. Counts of "best" and "worst" teacher are highly related to teacher relative rank order based on teaching scores. The "best" and "worst" counts and the relative ranks of the faculty are from 5 academic years. Relative ranks are based on teaching scores and were determined for each corresponding academic year. The faculty ranks were binned every 0.1 (10% of the faculty occurred in each bin width). The "best" and "worst" exponential fits have r^2 values of 0.97 and 0.96, respectively.

labeled as "best" was reduced by 63%. Thus, after three relative rank rates (which encompassed the top 46% of the faculty), it became unlikely that a faculty member was labeled as one of the "best" teachers. The "worst" histogram was fit by an exponential function that decayed with a relative rank rate of 16.6 ± 0.03%. This means that for every 16.6% increase in relative rank, the number of times a faculty member was labeled as "worst" was increased by 63%. Thus, it became more likely that a faculty member was labeled as one of the "worst" teachers as their relative rank increased and especially increased as they fell into the bottom half of the relative ranks. It is noteworthy that even faculty members who were ranked in the lower half of the numerical relative ranks were sometimes listed among our best faculty. The faculty listed as the worst teachers mainly dwell within the lowest 20% of the numerical relative ranks. Overall, the residents listed 870 names as "best" and 132 names as "worst." Thus, the number of teachers listed as "best" was more than 6 times greater than the number listed as "worst."

Did Residents Systematically and Indiscriminately Increase Teaching Scores?

If residents systematically and indiscriminately provided higher teaching scores for any reason (Grade Inflation Model), then as teaching scores increased, the rank order of the faculty would remain the same; scores would increase equally for all faculty members, diversity of scores would remain the same, and baseline scores given by residents would increase.

Rank Order Was Not Preserved over Time

To examine the stability of rank orders over time, the 50 faculty members who were evaluated in all 11 periods were studied. Their teaching scores are representative of the entire faculty (fig. 2A, +). They were ranked from 1 to 50 for each of the periods 1–11 based on their teaching scores during each period. When the rank order from period 1 was compared with any other later rank order (periods 2–11), the average Kendall τ was 0.42. When comparing the rank order of period 1 to any other later rank order, the maximum Kendall τ was only 0.61 (period 1 *vs.* period 11), and the mean upper 95% confidence interval for the Kendall τ was 0.57. Kendall τ never reached 1, which implies that later ranks had significant differences from the baseline rank order of period 1. Thus, rank order significantly changed as teaching scores increased over time.

The teaching score distributions shown in figure 1 reveal that teaching scores are not linearly distributed over the entire rank order. Teaching scores are disproportionately high and low in the top and bottom quarter of the rank list. Thus the faculty in the top and bottom quarter appeared separate from the middle faculty. Teaching scores from the middle 50% of the ranks is quite linear (period 1, rank region 0.25–0.75 $r^2 = 0.99$, $P < 0.0000001$; period 7, rank region 0.25–0.75 $r^2 = 0.99$, $P < 0.0000001$). We used this finding to divide the 50 persistently present faculty members into three categories: top 25%, middle 50%, and bottom 25%, which corresponds to the top 12, middle 26, and bottom 12 faculty members. When the top and bottom groups were analyzed for stability of rank order, they were notably better preserved than were the ranks of the middle 26 (fig. 4A). In particular, the faculty who initially occupied either the top 12 or the bottom 12 ranks in period 1 retained much of their rank order into periods 8, 9, 10, and 11. A significant relationship was found between the early rank order (period 1) and each of the later rank orders (period 8, 9, 10, or 11). The average Kendall τ for these rank-order comparisons was 0.51 ($P < 1 \times 10^{-7}$), and the 95% confidence interval did not include either 0 or 1 (fig. 4B). This contrasts with the complete mixing of rank orders for the middle 26 faculty members between period 1 and periods 8–11 (fig. 4A). The average Kendall's τ for these rank-order comparisons for the middle 26 faculty members was −0.064 and was not different from 0 ($P = 0.34475$) (fig. 4B). This analysis reveals that the middle 26 faculty members changed ranks to the extent that the initial rank order had no relationship to later rank orders.

Teaching Scores Increased the Most for Lower Ranked Faculty

The change in teaching scores between periods 1 and 9 for the top 12, middle 26, and bottom 12 ranked faculty members were computed. The scores increased the least for the top 12 ranked faculty members, moderately for the middle 26 ranked faculty members and the most for the bottom 12 ranked faculty members (fig. 5). The differences in score

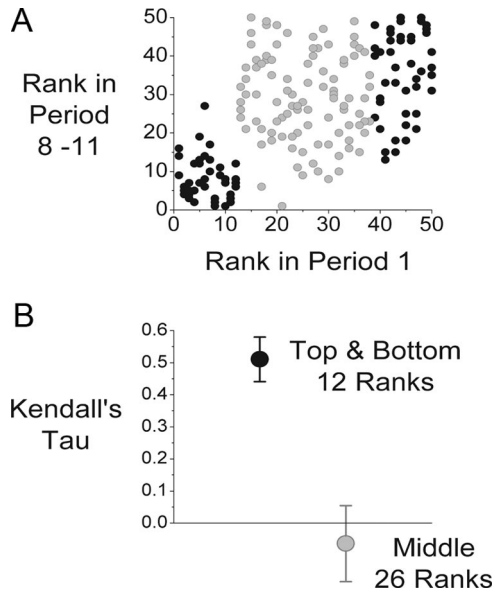

Fig. 4. (*A*) Top- and bottom-ranked faculty members better preserve their rank ordering; middle-ranked faculty members do not retain their rank order. Each symbol represents a single faculty member who was ranked in both period 1 and periods 8–11. Top 12 and bottom 12 faculty members are shown by *black symbols*. The middle 26 faculty members are shown by *gray symbols*. (*B*) Ranks are better preserved for the top and bottom faculty members. Ranks were compared between period 1 and periods 8–11. Kendall τ for the rank ordering of faculty members at the extremes of the ranks (*black symbol*, top 12 and bottom 12 ranks) was 0.51 ($P < 1 \times 10^{-7}$). Kendall τ for the rank ordering of faculty members in the middle relative ranks (gray symbol, middle 26 ranks) was −0.064 ($P = 0.34$). The error bars are the 95% confidence intervals.

increases were all significant (top *vs.* middle, $P = 0.0091$; middle *vs.* bottom, $P = 0.00016$; top *vs.* bottom, $P = 0.0000014$). The increases in teaching scores are thus not equal for all faculty members and instead are rank-related.

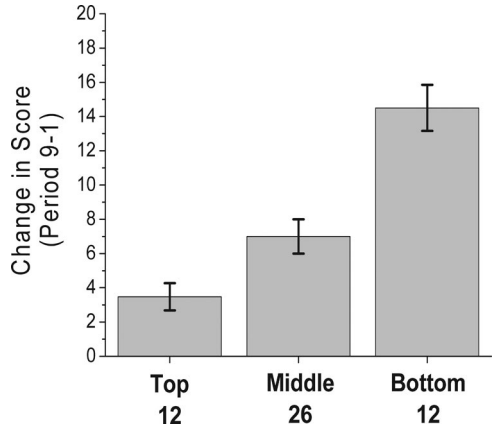

Fig. 5. Teaching scores increased the most for faculty members who were initially ranked the lowest. The 50 faculty members who were present for all 11 periods were grouped according to their relative rank in period 1. The mean teaching score change (difference in teaching scores between periods 9 and 1) is shown for each group.

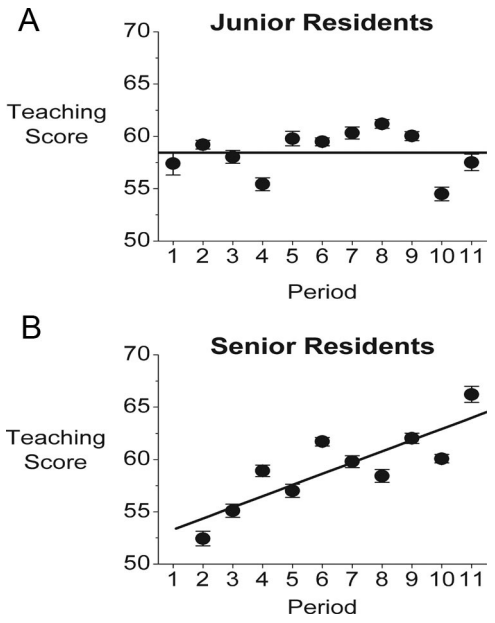

Fig. 6. (*A*) Junior residents give similar teaching scores over time. All teaching scores given by residents who had been in the program between 1 and 4 months were averaged for each period. The average teaching score given by these junior residents is plotted as a function of period. The data do not change over time ($P = 0.99$). (*B*) Teaching scores given by senior residents increase over time**.** All teaching scores given by residents who had been in the program between 24 and 36 months were averaged for each period. The average teaching score given by these senior residents is plotted as a function of period. The scores increase over time (slope of the fitted line is positive, $P = 0.0021$).

Junior Residents' Scores Stayed the Same Whereas Senior Residents' Scores Increased

The teaching scores given by junior residents (1–4 months of residency) and senior residents (24–36 months of residency) are plotted as a function of period (fig. 6). The scores given by junior residents did not change over time ($P = 0.99$, fig. 6A). In contrast, the scores given by senior residents increased over time ($P = 0.0021$, fig. 6B).

Discussion

Did Clinical Teaching Scores Improve?

The data from this study provide evidence that resident-based evaluation and feedback increases the clinical teaching scores of the teaching faculty. The increase in teaching scores was similar whether it was determined using all faculty evaluations in each period or just the scores from the 50 faculty members who were present throughout periods 1–11. This indicates that the increase in scores was not due to a change in faculty composition. Prior studies have found mixed results of the effectiveness of evaluation and feedback on teaching.[1,3,7] The current analysis differs from prior analyses by including a far larger number of evaluations. This study also provides the first measurement of the time needed to improve teaching. Few prior studies have had this longitudinal perspective or quantity of data.

A Novel Form of Construct Validity

We used an independent measure (being listed as "best" or "worst" teacher) to provide a non-numerical assessment of teaching. Our numerical teaching scores strongly indentify the same high and low performers as concurrently determined by counts of "best" and "worst" teacher designation. Overall, the histograms support our numerical evaluation system with concurrent construct validity.

Are Comments Necessary to Improve Teaching?

Although this study did not assess faculty members' interest in receiving feedback from residents, a prior study showed a strong interest of a volunteer community-based faculty in receiving feedback from medical students.[22] The community-based faculty valued student feedback over all other benefits offered to them for their teaching efforts, including money. This implies that at least some faculty members want feedback and may be interested in using it to improve their teaching. It is noteworthy that not all faculty members are interested in resident feedback about their teaching, and this can reach the level of resentment.[10] Our residents provided both quantitative (numeric) as well as qualitative (comments) feedback to the faculty. The comments provide direct constructive feedback for faculty members to use as tools to improve performance. Most faculty evaluation systems that fail to show improvement included only numerical ratings of the faculty without formative comments to help faculty improve.[3–6] In contrast, most faculty evaluation systems showing improvement in teaching scores, including the current study, included comments detailing strengths and areas for improvement.[1–3,20] The longest previous study of teaching scores did not show improvement over its 9-yr duration.[5] However, this study provided faculty only with numerical ratings and lacked formative comments. It is noteworthy that the addition of a comment section was associated with an improvement in scores within 1 yr.[2] Thus formative comments about strengths and weaknesses seem to be a key component enabling faculty to identify areas to improve. It is noteworthy that a study that provided only comments on areas of strength (with no mention of areas for improvement) showed no improvement over a 5-yr period.[8] Thus, our data suggest that areas of weakness need to be specifically identified to achieve increased teaching scores. Only when the teacher knows what areas to improve can they target the areas. Self-evaluation has proved to be remarkably inaccurate,[16,17] and thus external feedback provided by resident comments is likely to help identify areas that need improvement.

The impact of numerical-rating feedback *versus* comments-based feedback was determined using the studies with sufficient information (table 2). An overview of the studies in table 2 reveals that eight studies, including the current one, were designed to look at teaching scores over time. Four studies showed improvement in faculty teaching scores and four studies showed no improvement. All four studies showing an improvement included comments in the feedback material. All four studies demonstrating no improvement included only numerical feedback. Among these eight studies, a chi-square test reveals that comments were related to improved scores (Fisher exact test, $P = 0.03$). Thus, comments seem to be the driver for improved teaching, whereas numerical ratings track teaching but do not drive improvement.

The Kinetics of Improved Clinical Teaching Scores

The time course for improving teaching scores was well described by an exponential function with a time constant of approximately 0.94 yr. Approximately 95% of the improvement occurred within three time constants, which corresponds to approximately 2.8 yr in the current study. The change in score diversity showed a very similar temporal change, with a time constant of 1.2 yr.

Our evaluation and feedback process is quite similar to the evaluation and feedback process used by Schum and Yindra[1] in their "feedback" group. Their faculty had received feedback every 2 months for a total of six episodes of feedback.[1] They found improvement in 4 of the 10 traits in the feedback group. The effect size of the findings of Schum and Yindra was 0.22 after 1 yr, although it was not statistically significant. Cohan *et al.*[2] also had a process that was closely mimicked by the current evaluation and feedback process; they used a single annual feedback process and included specific suggestions from residents for ways that faculty could improve. His study found improvement in faculty teaching scores after 1 yr of feedback. Their data show an effect size of 0.31.[2] In the present study, at 1 yr, the faculty was provided with two episodes of feedback, and this resulted in an effect size of 0.26 (Cohen *d* of 0.26). In each case, after 1 yr, the magnitude of the improvement (the effect size) was similar. This suggests that the frequency of feedback may not be the limiting factor. Rather, it seems that faculty may take time to adjust their teaching skills and suggests that behavioral change is the actual rate-limiting step in improvement.

Do Clinical Teaching Scores Reflect Quality of Teaching?

Our teaching scores are believed to reflect actual quality of teaching. This hypothesis is based on the strong relationship between "best" and "worst" teachers and the concurrent teaching scores. When our residents list a faculty member as "best" or "worst," it is highly likely that the numerical scores will be accompanied by a high or low clinical teaching score, respectively (fig. 3). We also allow our residents to choose a "Teacher of the Year," and faculty members who are chosen regularly score in the top tier of our scoring system (data not shown). It is noteworthy that this form of evaluation synchrony is a form of "convergent validity"[23] and adds strength to our use of teaching scores to identify excellent teachers. Despite these relationships, we do not have any externally valid outcome data showing that the "best" teachers actually improve learning in the residents that they teach. This lack of outcome data is common in medical school and residency education.[23] Even where external raters have shown significant agreement with medical student ratings of lecturers,[24]

Table 2. Effects of Feedback on Teaching Scores

Studies	Description	Avg	Likert Span	Normed Score	No. of Faculty	No. of Evaluations	SEM	SD	*d*
Positive Effect									
Tiberius *et al.*, 1989[3]	Control	5.14	7	0.734	—	—	—	—	—
MS + R; CT	Ratings only	5.11	7	0.730	—	—	—	—	—
	Ratings plus comments	5.08	7	0.726	—	—	—	—	—
Schum and Yindra, 1996[1]	Baseline–overall, feedback group	5.65	7	0.807	21	266	0.11	0.49	0.22
MS + R; CT	After ratings + comments	5.76	7	0.823	21	359	0.12	0.54	—
Cohan *et al.*, 1996[2]	″Teaching″ comments + ratings–1993	7.70	10	0.770	40	117	—	1.30	0.31
RR; CT	″Teaching″ comments + ratings–1994	8.10	10	0.810	40	107	—	1.20	—
Baker, 2010 (this study)	Baseline	54.57	70	0.780	92	1,206	—	13.21	—
AR; CT	After three time constants (period 7)	59.96	70	0.857	104	1,579	—	9.90	0.41
	Plateau (last four periods)	59.67	70	0.852	—	8,007	—	—	—
No Effect									
Tortolani *et al.*, 1991[6]	No comments–no improvement in 1 yr	—	—	—	—	—	—	—	—
SR; CT		—	—	—	—	—	—	—	—
Risucci *et al.*, 1992[4]	Overall	3.76	5	0.752	64	—	—	—	—
R; CT	No comments–no improvement in 1 yr	—	—	—	—	—	—	—	—
Cohen *et al.*, 1996[5]	No comments–no improvement in 9 yr	—	—	—	—	3,750	—	—	—
MS; CT	Baseline (1985/86)	15.97	20	0.799	43	—	—	1.97	−0.06
	Ending year (1993/94)	15.85	20	0.793	43	—	—	1.72	—
Cox *et al.*, 2002[8]	Segments of data missing	—	—	—	—	—	—	—	—
SR + CT		—	—	—	—	—	—	—	—
Insufficient Information									
Sall *et al.*, 1976[30]	No statistics provided, insufficient information	—	—	—	—	—	—	—	—
MS; CT	Unclear if comments provided, volunteer faculty	—	—	—	—	—	—	—	—
Irby and Rakestraw, 1981[31]	Overall	3.92	5	0.784	230	1,567	—	—	—
MS; CT	Comments, improvement not assessed	—	—	—	—	—	—	—	—
Stillman *et al.*, 1983[20]	1981	4.07	5	0.814	—	—	—	0.45	—
MS; LR	1982	4.23	5	0.846	—	—	—	0.56	—
Skeff, 1983[7]	Baseline	4.09	5	0.818	16	—	—	0.34	—
MS + R; CT	After intensive feedback (videotape, etc.)	4.17	5	0.834	—	—	—	0.34	0.24
Nonsignificant ($P > 0.05$)	Baseline	4.19	5	0.838	16	—	—	0.42	—
	After rating feedback (scores only)	4.31	5	0.862	—	—	—	0.41	0.29
Fallon *et al.*, 1987[9]	Self-selection bias	—	—	—	—	—	—	—	—
MS; CT	—	—	—	—	—	—	—	—	—
Donnelly and Woolliscroft, 1989[32]	Improvement not assessed	5.58	7	0.797	90	218	—	—	—

(*continued*)

Table 2. Continued

Studies	Description	Avg	Likert Span	Normed Score	No. of Faculty	No. of Evaluations	SEM	SD	*d*
MS; CT	—	—	—	—	—	—	—	—	—
Blue *et al.*, 1999[28]	Improvement not assessed–no comments	—	—	—	—	—	—	—	—
MS	—	—	—	—	—	—	—	—	—
Copeland and Hewson, 2000[11]	Improvement not assessed	4.12	5	0.824	711	7,624	—	0.77	—
MS + R + F; CT	—	—	—	—	—	—	—	—	—
Stern *et al.*, 2000[29]	Improvement not assessed	4.14	5	0.828	74	476	—	0.47	
MS; CT		—	—	—	—	—	—	—	—
Griffith *et al.*, 2000[33]	Improvement not assessed	—	—	—	62	291	—	—	—
MS; CT	Best (top 20%)	4.80	5	0.960	12	—	—	0.31	—
	Worst (bottom 20%)	3.62	5	0.724	12	—	—	0.71	—
Williams *et al.*, 2001[34]	Improvement not assessed	—	—	—	—	—	—	—	—
MS + R; CT	Residents evaluate faculty	3.50	5	0.700	129	2,318	—	0.50	—
	Medical students evaluate faculty	4.04	5	0.808	129	4,425	—	0.56	—
Claridge *et al.*, 2003[19]	Improvement not assessed	4.16	5	0.832	23	828	0.03	0.14	—
SR; CT	—	—	—	—	—	—	—	—	—

Normalized teaching scores grouped by whether studies showed a positive effect of feedback on teaching scores, no effect on teaching scores, or contained insufficient information to determine an effect. The average absolute teaching score is listed (Avg) along with the dynamic range of the scoring system (Likert Span). Teaching scores were normalized by dividing absolute teaching scores by the dynamic range of the Likert scale used. The number of faculty members evaluated and the number of completed evaluations are listed. The SD and SEM are based on the data from the original studies. The effect size, *d*, was calculated when sufficient data was available in the original studies but was not reported in any of the original reports.

AR = anesthesia residents; CT = clinical teaching; F = fellows; LR = lecture ratings; MS = medical students; R = residents not otherwise specified; RR = radiology residents; SR = surgical residents.

there are not necessarily data showing that "better" lecturers result in "better" learning. Fortunately, there are examples of teaching scores being related to better student performance.[25] In university settings, student evaluations of teaching are associated with valid forms of achievement.[26,27] Although this study did not demonstrate improved learning with higher teaching scores, a body of literature demonstrates that better learning outcomes occur with higher teaching scores.[25–29]

How Do the Present Teaching Scores Compare to Others in the Literature?

Lectures and clinical teaching are usually evaluated using a Likert scale. Scores can be normalized by dividing the actual score by the dynamic range of the Likert scale. This converts each Likert score into a fraction of the maximum attainable score. Normalized teaching scores were computed from a variety of different teaching venues and found to be remarkably similar, with an overall mean of 0.797 (table 2). The 99% confidence interval for this mean was calculated as 0.774–0.820. During our baseline (period 1), our normalized clinical teaching score was 0.780, which falls within the estimated 99% confidence interval determined from the published studies. After our teaching scores had improved (period 7), our normalized clinical teaching score was 0.857, which is distinctly above and outside the 99% confidence interval for the mean estimated from the literature. Our evaluation and feedback system thus seems to have produced one of the highest normalized teaching scores reported.

Did Teaching Scores Increase As a Result of Simple Grade Inflation?

Simple grade inflation would increase teaching scores, leave the faculty rank order intact, increase scores equally for all faculty members, maintain score diversity, and increase the initial scores early in residency. Our data show that as teaching scores increased, the faculty rank order was not preserved (fig. 4), scores increased disproportionately for those whose ranks were lowest (fig. 5), scores became more homogeneous across faculty members (fig. 2B), and scores at the outset of residency were constant (fig. 6A). These analyses effectively rule out simple grade inflation as the cause for our increased clinical teaching scores.

The rank order changed over time, meaning that some but not all faulty members received higher scores, which in turn caused the rank order to change. The data also revealed

that the top-ranked faculty members improved least, perhaps because of a ceiling effect. The middle- and lowest-ranked faculty members' scores improved more than the top-ranked faculty. In fact, the lower the rank, the more they improved (fig. 5). Although the lowest-ranked faculty members improved the most, their very low initial teaching scores caused them to remain ranked near the bottom. This manifested as increased scores but persistently low rankings. The finding that the lowest ranked faculty members improved the most has been reported before.[1,2] A reduced but persistent gap in performance between top and bottom performers has been reported in a prior study.[2]

Why Did Junior Residents Give Similar Teaching Scores over Time Whereas Senior Residents Gave Higher Scores over Time?

When residents first start in residency, they typically find every interaction with a faculty member educational. Junior residents are typically very pleased with the teaching they receive at the beginning of residency. This may explain why scores of junior residents remain stable and high over time. The consistent teaching scores given by beginning residents argues strongly against simple grade inflation over time.

As residents become more senior, they become better at discriminating various aspects of teaching,[6] which implies that they become more sophisticated "consumers" of clinical teaching. Tortolani *et al.*[6] showed that senior residents used teaching evaluations in a more complex fashion that their more junior counterparts.

In the early periods, the lower scores given by senior residents indicate that they had became progressively less satisfied with the teaching they received as they progressed through residency. As evaluation and feedback affected the faculty and teaching improved, the senior residents became more pleased with the clinical teaching that they received, and the decline of the scores disappeared.

Limitations of This Study

This study's primary limitations are lack of a control group and lack of outcome data showing that better teaching scores translate into better learning outcomes. The lack of a control group means that other variables, including the Hawthorne effect, may have led to improved teaching scores. However, the Hawthorne effect would probably cause simple grade inflation, and our data have ruled that out. Our baseline ranks in period 1 pose another limitation, in that some findings in this study relate to rank orders and whether they change. This presumes a stable baseline, something that was not demonstrated. However, our residency had not undergone any large changes during the few years leading up to period 1. Our initial data (period 1) was acquired, analyzed, and reported back to the faculty after all the evaluations for period 1 had been received. Thus they had no feedback until after the conclusion of period 1. There is no reason to expect that the faculty were acting on data that they had not yet received. Our findings are also limited to the context in which our residents receive clinical education. Our clinical teaching interactions involve a great deal of direct observation and supervision. This allows our residents an excellent opportunity to evaluate the faculty member's clinical teaching. We did ask a small number of the lowest ranked faculty to work with a senior faculty member with an interest in education. We have no formal measurement of the impact of this voluntary intervention because we do not know which faculty members chose to use this resource. The small number of faculty identified and the likely smaller number choosing to use this resource make it unlikely to have strongly influenced the results. Last, the cause for the lack of agreement in the middle parts of the rank order in period 1 in comparison with the rank order in later periods (8–11) could be due to a lack of precision in measuring the teaching scores for each individual. In particular, the ability to create a reliable rank ordering is reduced in the middle ranks, where teaching scores are similar.

The author thanks Eleanor Cotter, A.S. (Education Coordinator, Department of Anesthesia, Critical Care and Pain Medicine, Massachusetts General Hospital, Boston, Massachusetts), for accurately and confidentially transcribing every aspect of evaluation and feedback into electronic form.

References

1. Schum TR, Yindra KJ: Relationship between systematic feedback to faculty and ratings of clinical teaching. Acad Med 1996; 71:1100–2
2. Cohan RH, Dunnick NR, Blane CE, Fitzgerald JT: Improvement of faculty teaching performance: Efficacy of resident evaluations. Acad Radiol 1996; 3:63–7
3. Tiberius RG, Sackin HD, Slingerland JM, Jubas K, Bell M, Matlow A: The influence of student evaluative feedback on the improvement of clinical teaching. J Higher Ed 1989; 60:665–81
4. Risucci DA, Lutsky L, Rosati RJ, Tortolani AJ: Reliability and accuracy of resident evaluations of surgical faculty. Eval Health Prof 1992; 15:313–24
5. Cohen R, MacRae H, Jamieson C: Teaching effectiveness of surgeons. Am J Surg 1996; 171:612–4
6. Tortolani AJ, Risucci DA, Rosati RJ: Resident evaluation of surgical faculty. J Surg Res 1991; 51:186–91
7. Skeff KM: Evaluation of a method for improving the teaching performance of attending physicians. Am J Med 1983; 75:465–70
8. Cox SC, Swanson MS: Identification of teaching excellence in operating room and clinic settings. Am J Surg 2002; 183:251–5
9. Fallon SM, Croen LG, Shelov SP: Teachers' and students' ratings of clinical teaching and teachers' opinions of use of student evaluations. J Med Ed 1987; 62:435–8
10. Guyatt GH, Nishikawa J, Willan A, McIlroy W, Cook D, Gibson J, Kerigan A, Neville A: A measurement process for evaluating clinical teachers in internal medicine. CMAJ 1993; 149:1097–102
11. Copeland HL, Hewson MG: Developing and testing an instrument to measure the effectiveness of clinical teaching in an academic medical center. Acad Med 2000; 75: 161–6
12. Jones RF, Froom JD: Faculty and administrative views of problems in faculty evaluation. Acad Med 1994; 69: 476–83
13. Abrami PC, Leventhal L, Perry RP: Educational seduction. Rev Ed Res 1982; 52:446–64

14. Ericsson KA: Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. Acad Med 2004; 79:S70–81
15. Kruger J, Dunning D: Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. J Pers Soc Psych 1999; 77:1121–34
16. Dunning D, Heath C, Suls JM: Flawed self-assessment: Implications for health, education and the workplace. Psych Sci Public Interest 2004; 5:69–106
17. Eva KW, Regehr G: Self-assessment in the health professions: A reformulation and research agenda. Acad Med 2005; 80:S46–54
18. Davis DA, Mazmanian PE, Fordis M, Harrsion RV, Thorpe KE, Perrier L: Accuracy of physician self-assessment compared with observed measures of competence – A systematic review. JAMA 2006; 296:1094–102
19. 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
20. Stillman PL, Gillers MA, Heins M, Nicholson G, Sabers DL: Effect of immediate student evaluations on a multi-instructor course. J Med Ed 1983; 58:172–8
21. Cohen J: A power primer. Psych Bull 1992; 112:155–9
22. Dent MM, Boltri J, Okosun IS: Do volunteer community-based preceptors value students' feedback? Acad Med 2004; 79:1103–7
23. Beckman TJ, Lee MC, Mandrekar JN: A comparison of clinical teaching evaluations by resident and peer physicians. Med Teach 2004; 26:321–5
24. Albanese MA, Schuldt SS, Case DE, Brown D: The validity of lecturer ratings by students and trained observers. Acad Med 1991; 66:26–8
25. Anderson DC, Harris IB, Allen S, Satran L, Bland CJ, Davis-Feickert JA, Poland GA, Miller WJ: Comparing students' feedback about clinical instruction with their performances. Acad Med 1991; 66:29–34
26. Marsh HW, Roche LA: Making students' evaluations of teaching effectiveness effective. Am Psychol 1997; 52: 1187–97
27. Cohen PA: Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. Rev Educ Res 1981; 51:281–309
28. Blue AV, Griffith CH, Wilson J, Sloan DA, Schwartz RW: Surgical teaching quality makes a difference. Am J Surg 1999; 177:86–9
29. Stern DT, Williams BC, Gill A, Gruppen LD, Woolliscraft JO, Grum CM: Is there a relationship between attending physicians' and residents' teaching skill and students' examination scores? Acad Med 2000; 75:1144–6
30. Sall S, Gromisch DS, Rubin SH, Stone ML: Improvement of faculty teaching performance in a department of obstetrics and gynecology by student evaluation. Am J Obstet Gynecol 1976; 124:217–21
31. Irby D, Rakestraw P: Evaluating clinical teaching in medicine. J Med Ed 1981; 56:181–6
32. Donnelly MB, Woolliscroft JO: Evaluation of clinical instruction by third-year medical students. Acad Med 1989; 64:159–64
33. Griffith CH, Georgesen JC, Wilson JF: Specialty choices of students who actually have choices: The influence of excellent clinical teachers. Acad Med 2000; 75:278–82
34. Williams BC, Pillsbury MS, Stern DT, Grum CM: Comparison of resident and medical student evaluation of faculty teaching. Eval Health Prof 2001; 24:53–60