A Feedback and Evaluation System That Provokes Minimal Retaliation by Trainees

Keith Baker, M.D., Ph.D., Bishr Haydar, M.D., Shawn Mankad, Ph.D.

ABSTRACT

Background: Grade inflation is pervasive in educational settings in the United States. One driver of grade inflation may be faculty concern that assigning lower clinical performance scores to trainees will cause them to retaliate and assign lower teaching scores to the faculty member. The finding of near-zero retaliation would be important to faculty members who evaluate trainees.

Methods: The authors used a bidirectional confidential evaluation and feedback system to test the hypothesis that faculty members who assign lower clinical performance scores to residents subsequently receive lower clinical teaching scores. From September 1, 2008, to February 15, 2013, 177 faculty members evaluated 188 anesthesia residents (n = 27,561 evaluations), and 188 anesthesia residents evaluated 204 faculty members (n = 25,058 evaluations). The authors analyzed the relationship between clinical performance scores assigned by faculty members and the clinical teaching scores received using linear regression. The authors used complete dyads between faculty members and resident pairs to conduct a mixed effects model analysis. All analyses were repeated for three different epochs, each with different administrative attributes that might influence retaliation.

Results: There was no relationship between mean clinical performance scores assigned by faculty members and mean clinical teaching scores received in any epoch ($P \geq 0.45$). Using only complete dyads, the authors' mixed effects model analysis demonstrated a very small retaliation effect in each epoch (effect sizes of 0.10, 0.06, and 0.12; $P \leq 0.01$).

Conclusions: These results imply that faculty members can provide confidential evaluations and written feedback to trainees with near-zero impact on their mean teaching scores. (ANESTHESIOLOGY 2017; 126:327-37)

EVALUATION systems are a cornerstone of medical education. Clinical performance evaluations are judgments by educators of a learner's clinical progress. Evaluations ensure that performance standards are being met. Clinical teaching evaluations are judgments by learners of educator's clinical teaching skillfulness. Teaching evaluations are important because they are often used in decisions about the educator's promotion, tenure, access to teaching venues, and merit raises.[1,2]

Grade inflation threatens the validity of evaluations, and in the worst case, faculty members have passed a medical student they felt should have failed a clinical rotation.[3,4] Grade inflation has been documented in high schools,[5] higher education,[6,7] third-year medical school clerkships,[3,8,9] fourth-year medical school subinternships,[4] applications to residencies,[10] and residency.[11] The mechanisms driving grade inflation include faculty members inflating scores in an attempt to receive reciprocal positive evaluations of their teaching skills.[2,3,6,12–14] Conversely, they may avoid assigning a low score to a trainee to avoid a reciprocal low teaching score.[3] This concern has some merit because retaliation (also termed reciprocity) was demonstrated in a general surgical residency when faculty members gave lower clinical scores to residents and the name of the faculty member was known to the trainee.[15]

What We Already Know about This Topic

- Teaching evaluations are important in medical education
- Inflation of student grades is common, and one driver is faculty inflating student evaluations in an attempt to receive reciprocal positive evaluations of their teaching skills, or avoiding giving low scores to avoid a reciprocal low teaching score

What This Article Tells Us That Is New

- In a residency training program, faculty members who assigned lower clinical performance scores to residents did not receive lower clinical teaching scores
- In this institution's residency program, there was little or no retaliatory effect when faculty members gave residents low clinical scores when providing confidential evaluations and written feedback to trainees

The evaluation and feedback system used in our study keeps evaluator name confidential. As such, it may appear to

Submitted for publication March 13, 2016. Accepted for publication November 7, 2016. From the Harvard Medical School, Boston, Massachusetts; Department of Anesthesia, Critical Care and Pain Medicine, Massachusetts General Hospital, Boston, Massachusetts (K.B.); Division of Pediatric Anesthesia, Department of Anesthesiology, University of Michigan, Ann Arbor, Michigan (B.H.); and Samuel Curtis Johnson Graduate School of Management, Cornell University, Ithaca, New York (S.M.).

 Anesthesiology 2017; 126:327–37

<zdoi;10.1097/ALN.0000000000001471>

be impossible for residents to retaliate for low scores or negative feedback. However, characteristics of optimal feedback (*i.e.*, specific, timely, nonjudgmental, and aimed at helping the learner improve[16,17]) will often provide enough information to identify the author of the feedback. Additionally, a significant amount of communication is nonverbal,[18,19] and it is possible that negative evaluation is communicated nonverbally when a faculty member interacts with a resident.[20] Some faculty members have expressed concern to the residency program director (K.B.) about receiving lower clinical teaching scores if they submit a negative evaluation on a resident. This concern is shared by surgical faculty members who are leery of providing poor evaluations to trainees, even when done anonymously, due to concern that residents can identify the faculty member who rendered the poor evaluation.[21]

Our program possesses both faculty member concern with retaliation and grade inflation, so we sought evidence of retaliation using two different approaches. Using our confidential evaluation and feedback system, we addressed macroscopic retaliation by investigating whether faculty members who assigned, on average, lower clinical performance scores to residents were assigned, on average, lower clinical teaching scores by residents. This is termed the leniency hypothesis (teachers who provide higher mean scores to learners are awarded higher mean teaching evaluations).[14] We also addressed microscopic retaliation using dyads of individual faculty member–resident pairs using a mixed effects model. Dyads allowed us to study whether there was direct retaliation between individual faculty member–resident pairs. This is termed the reciprocity hypothesis (a learner will assign a higher teaching evaluation to a faculty member if they had received a higher evaluative score from the teacher).[14] Last, we evaluated sex since it has been shown to influence the assessment of faculty teaching.[22–24]

Materials and Methods

The Massachusetts General Hospital Institutional Review Board (Boston, Massachusetts) waived the need for informed consent and classified this study as exempt (protocol no. 2013P000912, May 21, 2013). Three distinct periods (epochs) were identified during the study period (September 1, 2008, until February 15, 2013). Each epoch was characterized by a unique combination of administrative details pertaining to how evaluation and feedback information was obtained and distributed (table 1). The evaluator's name was kept confidential on all evaluations in all three epochs.

Faculty Member Evaluation of Resident Clinical Performance

Each week, faculty members were assigned to provide numerical evaluation and written feedback on resident clinical performance. Evaluation assignments were based on our anesthesia information management system. This system tracks which faculty members supervised which residents during the previous week. When anesthesia information management system data were not available (intensive care unit, preoperative clinic, recovery room, and pain clinic), we used weekly schedules to determine who worked with whom as previously published.[11] Faculty evaluation of resident clinical performance was based on the peer comparison section of our evaluation form, which had seven elements, each with a Likert score ranging from 1 to 5. We used the mean of the seven subscores to represent the overall score by a faculty member even though the subscores were from a Likert ordinal scale. Our use of means for summarizing ordinal data has been criticized,[25] but the pragmatic use is supported in instances where the sample size is large and the data are approximately normally distributed,[26,27] as we have

Table 1. Administrative Details of Evaluation and Feedback for Each Epoch

	Epoch 1	Epoch 2	Epoch 3
Time period	September 1, 2008, to January 31, 2010	April 1, 2010, to June 30, 2011	September 1, 2011, to February 15, 2013
Months	17	15	18
Faculty are requested to evaluate residents	Weekly, confidentially, electronically	Weekly, confidentially, electronically	Weekly, confidentially, electronically
Mean delay in days for faculty submissions (SD)	23 (26)	20 (18)	18 (13)
Residents receive these evaluations	Every 1–2 wk	Every 2 wk	Every 2 wk
Salient features	Residents see full evaluations with scores and comments	Residents see only aggregated comments	Residents see only aggregated comments
Residents are requested to evaluate faculty	Monthly, confidentially, in writing, after rotation ends	Monthly, confidentially, in writing, after rotation ends	Weekly, confidentially, electronically
Mean delay for resident submissions (SD)	About 4.5 wk	About 4.5 wk	23 (32) d
Faculty receive these evaluations	Every 6 mo	Every 6 mo	Every 6 mo
Salient features	Delayed to preserve resident confidentiality; mean scores and aggregated comments, as well as benchmark departmental mean scores	Delayed to preserve resident confidentiality; mean scores and aggregated comments, as well as benchmark departmental mean scores	Delayed to preserve resident confidentiality; mean scores and aggregated comments, as well as benchmark departmental mean scores

shown to be the case with our data.[11] Importantly, in the peer comparison section of our form, we previously published and defined a score of 3 to mean peer average when compared to other Massachusetts General Hospital anesthesia residents at the same level of training.[11] Thus, the average peer comparison score should not rise as residents advance in the program. The average of the seven elements was used as an overall clinical performance score,[11] and the average was rescaled onto a 0 to 100 range. Each evaluation form also has areas for faculty members to provide written feedback to the resident. The complete evaluation form has been published.[11] Reminder emails were sent at least weekly in response to delinquent evaluations. Details of faculty member evaluation and feedback regarding resident clinical performance for each epoch are found in table 2.

Resident Evaluation of Faculty Member Clinical Teaching

During epochs 1 and 2, residents were assigned to evaluate their faculty member's clinical teaching based on monthly billing data, which enabled us to know which faculty member supervised which resident. Pairings were extracted about 2 to 3 weeks after the completion of each month-long rotation as previously published.[28] During epoch 3, we used the weekly list detailing which faculty member was assigned to evaluate which resident (see Faculty Member Evaluation of Resident Clinical Performance) to then assign residents to evaluate the corresponding faculty members. Thus, in epoch 3, we had a weekly bidirectional evaluation process. Raw teaching scores contained seven clinical teaching subscores, each with a Likert score ranging from 0 to 10; thus, composite teaching scores ranged from 0 to 70.[28] Teaching scores were rescaled onto a 0 to 100 range. Each evaluation form also has areas for residents to provide written feedback to the faculty member. Details of resident evaluation and feedback regarding faculty member clinical teaching for each epoch are found in table 3. Reminders were sent monthly (epochs 1 and 2) or at least weekly (epoch 3) in response to delinquent evaluations.

Macroscopic Assessment of Retaliation

Macroscopic retaliation refers to the process whereby faculty members who assign, on average, lower clinical performance scores to residents receive, in return, lower teaching scores from residents. Detection of macroscopic retaliation amounts to finding lower average teaching scores among faculty members who assign lower resident clinical performance scores. This has been termed the leniency hypothesis.[14] For each epoch, our

Table 2. Faculty Evaluation of Resident Clinical Performance

	Epoch 1	Epoch 2	Epoch 3
Total evaluations	9,540	7,904	10,117
No. of faculty members submitting evaluations	123	133	138
Mean evaluations submitted by each faculty (SD)	78 (53)	59 (47)	73 (57)
No. of residents evaluated	109	103	115
Mean evaluations by each faculty for each resident (SD)	1.9 (1.3)	1.9 (1.2)	1.9 (1.2)
Median (maximum)	1 (13)	2 (10)	1 (10)
Percent of evaluations with comments	51.6	69.1	71.0
Common faculty members between epochs			
1 and 2	107	107	
2 and 3		109	109
1 and 3	89		89
Mean raw score (SD)	3.50 (0.64)	3.48 (0.61)	3.49 (0.62)
Mean scaled score (SD)	62.4 (15.9)	62.1 (15.4)	62.1 (15.6)

Table 3. Resident Evaluation of Faculty Clinical Teaching

	Epoch 1	Epoch 2	Epoch 3
Total evaluations	7,923	4,747	12,388
No. of residents submitting evaluations	106	82	103
Mean evaluations submitted by each resident (SD)	75 (59)	58 (48)	120 (70)
No. of faculty members evaluated	142	140	163
Mean evaluations by each resident for each faculty (SD)	1.6 (0.9)	1.5 (0.8)	2.0 (1.3)
Median (maximum)	1 (8)	1 (6)	2 (10)
Common residents between epochs			
1 and 2	57	57	
2 and 3		46	46
1 and 3	24		24
Mean raw score (SD)	57.4 (10.1)	61.2 (9.0)	59.2 (9.9)
Mean scaled score (SD)	82.0 (14.5)	87.5 (12.9)	84.6 (14.1)

independent measure was the mean resident clinical performance score assigned by each faculty member. This measure determined the level of faculty leniency (a point measure on the hawk–dove continuum[11,29]). We then used the mean teaching score received by each faculty member as our dependent measure of retaliation. Figure 1A displays these interactions. A pairing was null-resident if a faculty member submitted clinical performance scores to a resident and that resident did not submit any clinical teaching scores on that faculty member during an epoch (fig. 1A). A pairing was null-faculty if the resident submitted clinical teaching scores on a faculty member but the faculty member did not submit any clinical performance scores on that resident during an epoch (fig. 1A).

Microscopic Retaliation: Linking Faculty Member and Resident Evaluations to Create Complete Dyads

Microscopic retaliation is a term we use to describe the retaliation effect between a single faculty member and a single resident (a dyad). A pair (dyad) was complete if a faculty member submitted one or more clinical performance scores to a resident and that resident also submitted one or more clinical teaching scores to that faculty member during an epoch. Figure 1B displays these interactions.

Faculty members and residents sometimes evaluated each other more than once in an epoch because they worked together more than once during an epoch. On average, each resident was evaluated twice by a given faculty member during an epoch. On average, each resident evaluated each faculty member once or twice during an epoch. Thus, the most common dyads in an epoch were 1:1 or 2:1. We defined a dyadic interaction within an epoch as the average score that a faculty member assigned to a resident coupled to the average teaching score that the resident assigned to that faculty member.

Timing (Sequencing) of Evaluation Requests and Returns

In order to investigate retaliation, we had to know the sequence of who evaluated whom and when. During all epochs (1, 2, and 3), faculty members were assigned to evaluate residents

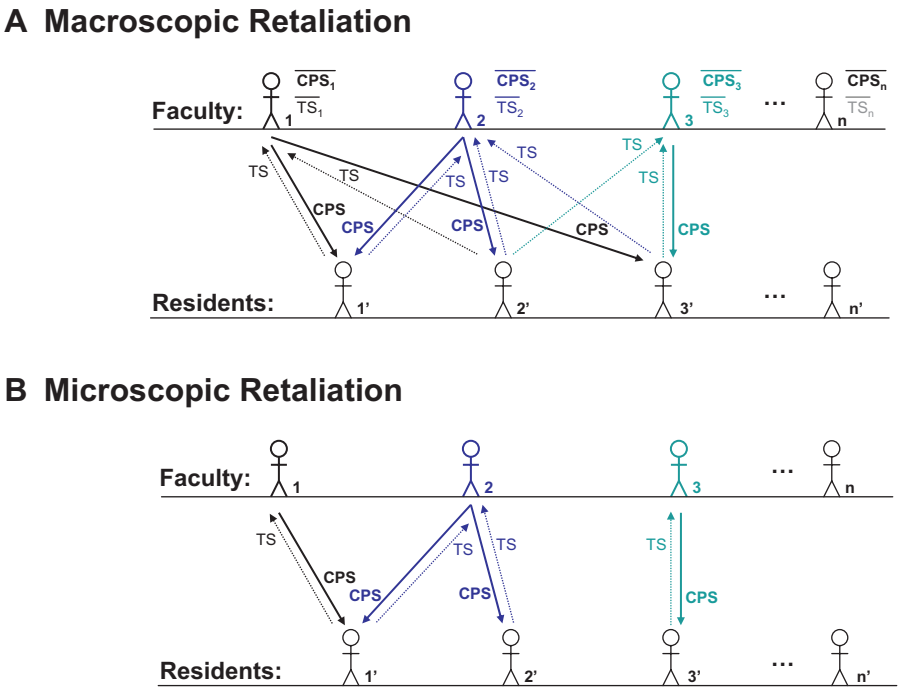

Fig. 1. Interactions between faculty members and residents. (*A*) Model of interactions used in macroscopic retaliation analysis. Schematized faculty members are shown on the top line interacting with schematized residents on the bottom line. Faculty members and residents interact in the perioperative setting and can evaluate each other. Faculty members assign clinical performance scores (CPSs) to many of the residents they work with (*colored solid lines* labeled CPS from faculty member to resident). Sometimes, faculty members do not submit an evaluation (no arrow connecting faculty member to resident). Residents assign clinical teaching scores (TSs) to many of the faculty members they work with (*colored dotted lines* labeled TS from resident to faculty member). Sometimes, residents do not submit an evaluation (no arrow connecting resident to faculty member). Each faculty member has a mean clinical performance score that they assigned to residents (correspondingly colored CPS topped by a line) and a mean clinical teaching score that they received from residents (correspondingly colored TS topped by a line). Individual faculty members and their associated CPSs and TSs are denoted by the same color. When only the faculty member of the pair submits an evaluation (the resident evaluation of the faculty member is absent), this is termed a null-resident interaction (see, for example, faculty member 1 and resident number 3). When only the resident of the pair submits an evaluation (the faculty member evaluation of the resident is absent), this is termed a null-faculty interaction (see, for example, faculty member 1 and resident number 2). (*B*) Model of interactions used in microscopic retaliation analysis. Only complete dyads are used in the microscopic retaliation analysis. A complete dyad is indicated by a faculty member–resident pair where each individual of the pair evaluates the other (see, for example, faculty member 1 and resident number 1).

they had worked with during the previous 7 days. During epochs 1 and 2, after each month-long rotation, residents were assigned to evaluate the faculty members that they had worked with during the previous month. Thus, in epochs 1 and 2, there was a built-in structural delay of at least 2.5 weeks and up to 6.5 weeks (mean, 4.5 weeks) before residents were assigned to evaluate the faculty members they worked with. During epoch 3, residents were assigned to evaluate faculty members they worked with during the previous 7 days. Thus, in epoch 3, all requests to evaluate were essentially synchronized in time. We measured the real delay (in days) for all faculty-based evaluations of resident clinical performance during all three epochs (table 1). We were able to measure the real delay (in days) for all resident-based evaluations of faculty clinical teaching only for epoch 3 (table 1). In epoch 3, residents had a longer delay than faculty members (mean [SD], 23 [32] *vs.* 18 [13] days; $P < 0.001$, unpaired Student's *t* test). Thus, our system was arranged so that, on average, faculty members evaluated residents before residents evaluated the corresponding faculty member. We were not able to determine the actual timing for each dyad, and thus, we expect that some sequencing was synchronous or even inverted in all epochs but especially in epoch 3.

Different Components of the Evaluation and Feedback Form Were Revealed in Each Epoch

During epoch 1, residents were able to see the entirety of each evaluation form that contained both evaluative scores and formative feedback comments (but not the name of the faculty member who submitted the form). Links to view these completed evaluation forms were emailed to residents every 7 to 10 days during epoch 1, and residents were required to sign that they read them. During epochs 2 and 3, residents were only able to see portfolios of aggregated written feedback comments but not the corresponding evaluative scores or names of the faculty members who submitted the comments (table 1). Links to portfolios were emailed to residents and their mentors every 2 weeks. Both residents and mentors were required to sign that they read the portfolios. Residents were more than 98% compliant with signing that they have reviewed their evaluations (epoch 1) and portfolios (epochs 2 and 3). We uncoupled evaluative scores from formative feedback comments during epochs 2 and 3 due to educational research showing that grades (scores) can reduce the motivation to learn.[30–33] Clinical performance scores were processed into *Z* scores and used by the clinical competency committee to determine resident clinical performance and to identify residents who were particularly in need of improvement.[11]

Statistics

In each epoch, we assessed for macroscopic retaliation by performing linear regression between the mean clinical performance score assigned by each faculty member (independent variable) and the mean clinical teaching score that was received by that faculty member (dependent variable). We used all available data to compute each mean clinical performance score and each mean clinical teaching score. We ensured that linear regression was appropriate for our datasets[34] by ensuring that the population errors of each regression model were normally distributed. We also assessed the residuals for heteroscedasticity. The population errors were normally distributed as determined by linear normal probability plots. In addition, none of our linear regression analyses displayed significant heteroscedasticity.

We sought evidence of microscopic retaliation using a mixed effects model using only complete dyads. A complete dyad was composed of two individuals, a faculty member and a resident who evaluated each other. Each dyad had two numerical parts: a mean clinical performance score assigned by the faculty member to the resident (primary independent variable of interest) and a mean clinical teaching score assigned to that faculty member by the evaluated resident (dependent variable). Each dyad was composed of a unique faculty member–resident pairing. Our model took into account epoch (structurally different ways to acquire and distribute evaluations), age of resident in the program (0 to 36 months) since this has been shown to effect assignment of teaching scores,[24,28] the number of evaluations submitted by a resident on a faculty member, and a retaliation effect (how the faculty member's score of a resident's clinical performance affected that resident's clinical teaching score of that same faculty member). A random effects term was included to account for repeated dyads (the same faculty–resident pairing) that occurred in more than one epoch. The mixed effects model coefficients were estimated using reduced maximum likelihood.

Sex effects on faculty member evaluation of residents and of resident evaluation of faculty members were computed by comparing the means of scores assigned by males and females using unpaired Student's *t* tests assuming unequal variance in the measures. We chose Student's *t* tests to compare means for two reasons. First, with larger datasets ($n > 50$), the Student's *t* test is a robust statistic for both normally and nonnormally distributed datasets[34–38] due to the central limit theorem. Second, the Student's *t* test provides additional power to detect differences. Thus, if we did not find a difference using a Student's *t* test, then we were all but certain not to detect a difference using a parametric test such as the Wilcoxon test.

The effect of overall resident clinical performance on resident-assigned faculty clinical teaching scores was determined by performing linear regression between the mean overall resident clinical performance score (Z_{rel} score[11]) and the mean teaching score assigned by that resident for each epoch. Mean Z_{rel} scores were computed using all individual relative to peers[11] Z_{rel} scores for each resident during each epoch. This regression analysis met the criteria of having a linear distribution of errors and no heteroscedasticity.

Statistical results were determined using StatsDirect, Version 2.6.6 (StatsDirect Ltd., United Kingdom), Excel,

Version 2003 (Microsoft, USA), Origin, Version 7.5 SR4 (OriginLab, USA), or SPSS, Version 21 (IBM Corporation, USA). Effect sizes were determined by Cohen *d* and provide a measure of the size of a difference compared to the variation in the data.[39,40] Effect sizes are classified as small (Cohen $d = 0.2$), medium (Cohen $d = 0.5$), or large (Cohen $d = 0.8$).[39,40] *P* values are two sided and determined exactly whenever possible. A $P < 0.05$ was considered statistically significant.

Results

Faculty Members Who Assigned Lower Resident Clinical Performance Scores Did Not Receive Lower Teaching Scores (No Macroscopic Retaliation Effects)

Our faculty members provide confidential evaluations (scores) and feedback (written comments) to our residents. Our residents receive a large number of evaluations of which 52, 69, and 71% contained written comments in epochs 1, 2, and 3, respectively (table 2). We sought evidence that faculty members who assigned lower average clinical performance scores to residents would receive lower average teaching scores from residents. We found no relationship between the average clinical performance score assigned by a faculty member and the average teaching score that residents assigned to that faculty member in any of the three epochs ($P \geq 0.45$; fig. 2 and table 4). In other words, faculty members who assigned lower clinical performance scores to residents did not receive lower clinical teaching scores in return. This broad macroscopic view indicated a lack of retaliation under each of the three different administrative conditions. A *post hoc* power analysis using data from all three epochs demonstrated that we had more than 80% power (with $\alpha = 0.05$) to detect a very small retaliation effect ($r = 0.2$; $d = 0.04$). Thus, we have essentially ruled out the leniency hypothesis for our program using our confidential evaluation and feedback system.

Analysis of Faculty Member–Resident Pairs (Dyads) Reveals a Very Small Retaliation Effect (Microscopic Retaliation Effects)

Since we did not find a macroscopic retaliation effect, we proceeded with the mixed effects model to evaluate specific interactions. Our mixed effects model detected a very small retaliation effect in each epoch (table 5). In epochs 1, 2, and 3, the retaliation effect amounted to 0.09, 0.05, and 0.11 point changes in the faculty teaching score (on a 0 to 100 scale) for every one-point change in the resident performance score (on a 0 to 100 scale; $P < 0.001$, $P = 0.010$, and $P < 0.001$, respectively). Thus, using our confidential evaluation and feedback system, we found support for a very small effect of the retaliation hypothesis.

In contrast to these very small retaliation effects, the seniority of the resident had a much larger effect on assigned teaching scores in some epochs. In epochs 1, 2, and 3, for each additional month that a resident was in the program, the average assigned faculty teaching scores decreased by 0.26 and 0.30 or increased by 0.04, respectively ($P < 0.001$, $P < 0.001$, and $P = 0.039$, respectively). An additional small effect on teaching scores was also found for the number of teaching evaluations that each resident submitted per faculty member in epochs 2 and 3. In epochs 2 and 3, the teaching score was increased by 0.73 and 0.86 for each additional teaching evaluation returned by the same resident on the same faculty member, respectively ($P = 0.03$ and $P < 0.001$).

Faculty Member Sex, Resident Sex, and Resident Clinical Performance Do Not Affect Assigned Scores

Faculty member sex did not influence the assignment of resident clinical performance scores. Male and female faculty members assigned similar resident clinical performance scores across all three epochs (table 6; $P \geq 0.15$). Resident sex did not influence the assignment of faculty member

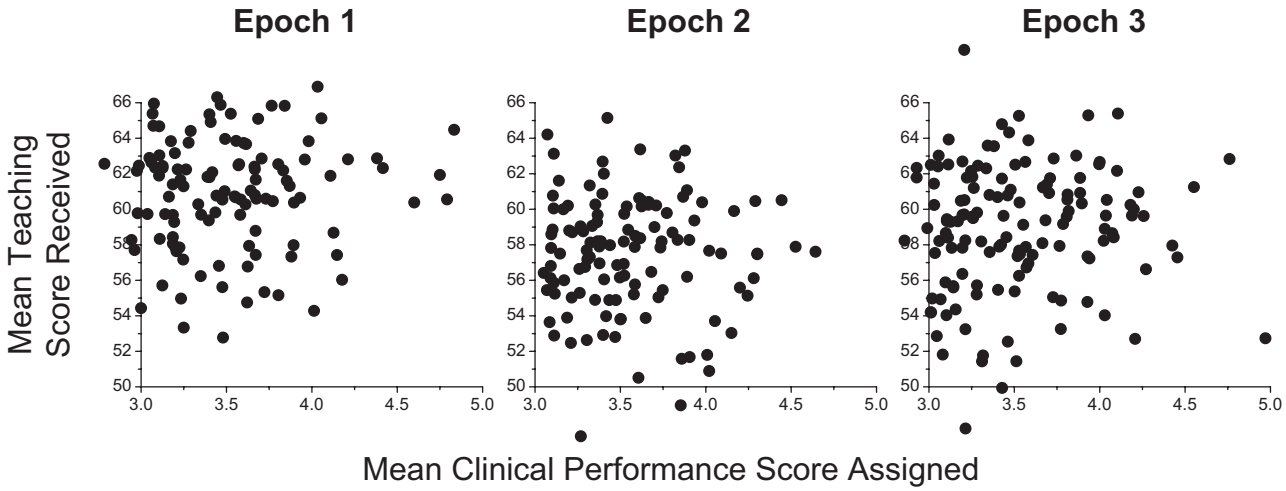

Fig. 2. Faculty clinical teaching scores are unrelated to clinical performance scores assigned by faculty members. Mean raw clinical performance score assigned by each faculty member is plotted (*x*-axis) against the mean raw clinical teaching score received by that faculty member (*y*-axis) during each epoch. There were 119, 115, and 137 faculty members in epochs 1 to 3, respectively. Relationships were not significant in any epoch (all $P \geq 0.45$).

Table 4. Overall Macroscopic Retaliation Effect and Structural Alignment of Dyads

	Epoch 1	Epoch 2	Epoch 3
Faculty members (n)	119	115	137
P value	0.83	0.56	0.45
Total dyads (faculty–resident pairings)	9,540	7,904	10,117
Complete dyads (both faculty evaluation of resident clinical performance present and resident evaluation of faculty clinical teaching present, %)	3,005 (43.8)	1,849 (34.5)	4,023 (54.2)
Null-resident dyads (faculty evaluation of resident clinical performance present but resident evaluation of faculty clinical teaching absent, %)	1,953 (28.5)	2,248 (41.9)	1,231 (16.6)
Null-faculty dyads (faculty evaluation of resident clinical performance evaluation absent but resident evaluation of faculty clinical teaching present, %)	1,900 (27.7)	1,262 (23.5)	2,167 (29.2)

Table 5. Detection of Microscopic Retaliation Using a Mixed Effects Model and Only Complete Dyads

	Epoch 1	Epoch 2	Epoch 3
Retaliation effect			
Change in faculty teaching score, given a unit change in resident clinical performance score	0.09	0.05	0.11
P value	< 0.001	0.01	< 0.001
Change in faculty teaching score, given a 1-SD change in resident clinical performance score	1.4	0.8	1.7
Effect size of a 1-SD change in resident clinical performance score on faculty teaching score (Cohen *d*)	0.10	0.06	0.12
Interpretation of effect size	Very small	Very small	Very small
Time in residency effect			
Change in faculty teaching score for each additional month of residency training (residency is 36 mo)	–0.26	–0.30	0.04
P value	< 0.001	< 0.001	0.039
Difference in teaching score from residents in the final month of training compared to the first month of training	–9.3	–10.7	1.5
Effect size of final *vs.* first month resident assignment of teaching scores (Cohen *d*)	–0.64	–0.83	0.10
Interpretation of effect size	Large	Large	Very small
No. of evaluations submitted effect			
Effect of each additional submitted teaching evaluation on teaching scores	0.03	0.73	0.86
P value	0.91	0.03	< 0.001
Change in teaching score for each additional evaluation submitted	NS	0.73	0.86
Effect size of one additional evaluation on teaching score (Cohen *d*)	NS	0.06	0.06
Interpretation of effect size	NS	Very small	Very small

All tests were performed using 0 to 100 scale.
NS = not significant.

clinical teaching scores. Male and female residents assigned similar faculty member clinical teaching scores across all three epochs (table 6; $P \geq 0.16$). Since different epochs were not likely to inherently influence sex bias, we increased the power to detect an effect by combining the data from all three epochs. The combined dataset had 172 unique residents (67 females and 105 males) who had submitted at least five evaluations. Using this *post hoc* dataset, we were not able to detect an effect of resident sex on the teaching scores they assigned ($P = 0.13$). We were also not able to detect an interaction of sex on assignment of teaching scores. Male residents evaluated male faculty members (n = 105; mean [SD], 84.2 [8.5]) the same as they evaluated female faculty members (n = 98; mean, 80.0 [8.5]; $P = 0.83$). Female residents evaluated male faculty members (n = 67; mean, 81.8 [9.3]) the same as they evaluated female faculty members (n = 63; mean, 82.5 [9.1]; $P = 0.63$). Overall resident clinical performance, as determined using mean Z_{rel} scores, did not influence the assignment of faculty member teaching scores across all three epochs (table 6; $P \geq 0.10$).

Discussion

We Detected Either No or Very Small Retaliation Effects Using Our Evaluation and Feedback System

Our main finding was that faculty members who assigned lower clinical performance scores to residents did not receive lower clinical teaching scores from residents. We, thus, reject the leniency hypothesis using our system. Our results were obtained using three different administrative approaches to evaluation and feedback (epochs 1 to 3), and the results were consistent (fig. 2; table 4). Importantly, the evaluative scores that our faculty members assign to residents are converted to Z_{rel} scores, which correct for individual bias and the unique grade range usage of each faculty member.[11] Average Z_{rel} scores differentiate residents who deliver lower and higher clinical performance,[11]

Table 6. Gender and Resident Clinical Performance Effects

	Epoch 1	Epoch 2	Epoch 3
Faculty evaluation of residents			
Difference in clinical performance scores assigned by male and female faculty members	NS	NS	NS
df	78	49	60
P value	0.15	0.33	0.57
Resident evaluation of faculty			
Difference in teaching scores assigned by male and female residents	NS	NS	NS
df	63	67	79
P value	0.16	0.27	0.22
Relationship between resident clinical performance (*Z* score) and mean teaching score assigned	NS	NS	NS
df	90	68	99
P value	0.32	0.62	0.10

All tests were performed using 0 to 100 scale.
df = degrees of freedom; NS = not significant.

are stable over time, reliably identify low performers, detect improvement in performance when an educational intervention is successful, are related to an external measure of medical knowledge, and identify poor performance due to a wide variety of causes.[11] Z_{rel} scores also are related to American Board of Anesthesiology written (part 1) and oral (part 2) examination scores used to determine board certification.[41] Thus, scores that faculty members assign, and comments they write, contain diagnostic information about resident performance. Our study demonstrates that this information can be conveyed to a clinical competency committee without important resident retaliation toward faculty clinical teaching scores. Our results contrast with those of Gardner and Scott[15] who found a macroscopic retaliation effect, but the name of the faculty member was known to the resident. We believe that our confidential system eliminates the macroscopic retaliation effect.

Our second finding was made using unique faculty member–resident pairings where bidirectional evaluation had occurred to look for microscopic retaliation using a mixed effects model. Our analysis detected a statistically significant but very small retaliation effect in all three epochs. The retaliation effect in epoch 1 was no larger than in any other epoch, and it was the only epoch in which we provided residents with the entire evaluation form (including scores and written comments). The largest, yet still very small, retaliation effect occurred in epoch 3 when residents and faculty members were evaluating each other in the most synchronized manner. A potential mechanism for this finding is described next.

How Can Residents Retaliate When They Do Not Know Who Evaluated Them?

Our system treats evaluator identity as confidential (knowable but not revealed). Ostensibly, our system should be immune to retaliation since evaluator identity remains confidential. However, if a faculty member revealed enough information in their written feedback comments, then the resident would know who wrote the comments. In addition, if a faculty member displayed anger or frustration toward a resident, then the resident may react in a negative and retaliatory manner.[20] Finally, a significant amount of communication is nonverbal,[18,19] and negative evaluations may be nonverbally communicated. Recent evidence demonstrated a universal facial expression that communicates negative judgments.[18] We speculate that nonverbal communication explains the retaliation effect found in epoch 3 when nearly synchronized bidirectional evaluation was occurring. This mechanism would also explain why all three epochs had similarly sized retaliation effects despite structural differences in the evaluation and feedback process.

Contextualizing the Size of the Retaliation Effects

To place these findings in context, we modeled the effects of retaliation on faculty teaching rankings based on mean teaching scores. We modeled having a faculty member decrease the scores they assigned to residents by one full SD and assumed retaliation on all subsequent teaching evaluations by the amount detected using only complete dyads (a very conservative projection). Our model demonstrated a retaliatory change in teaching scores of 1.4, 0.8, and 1.7 (on a 0 to 100 scale) for epochs 1, 2, and 3, respectively. These changes would translate into effect sizes, *d*, of 0.10, 0.06, and 0.12, respectively. A change in teaching scores of this magnitude would only slightly change the rank ordering of faculty members. The effects are shown for a high-scoring (5th percentile), average-scoring (50th percentile), and low-scoring faculty member (95th percentile) for each epoch (fig. 3).

In contrast to the very small retaliation effects, resident seniority had a much larger effect on teaching scores. For example, a senior resident (36 months in the program) would, on average, assign teaching scores that were 9.3 points lower, 10.7 points lower, or 1.5 points higher (on a 0 to 100 scale) than those assigned by a beginning resident (0 months in the program) in epochs 1 to 3, respectively. These effects translate into effect sizes, *d*, of 0.64, 0.83, and 0.10, respectively (table 5). The finding that more senior residents render lower faculty clinical teaching scores has been published[24,28] and appears to be related to increasing discernment of what constitutes effective clinical teaching as residents advance in residency.[24,28] The

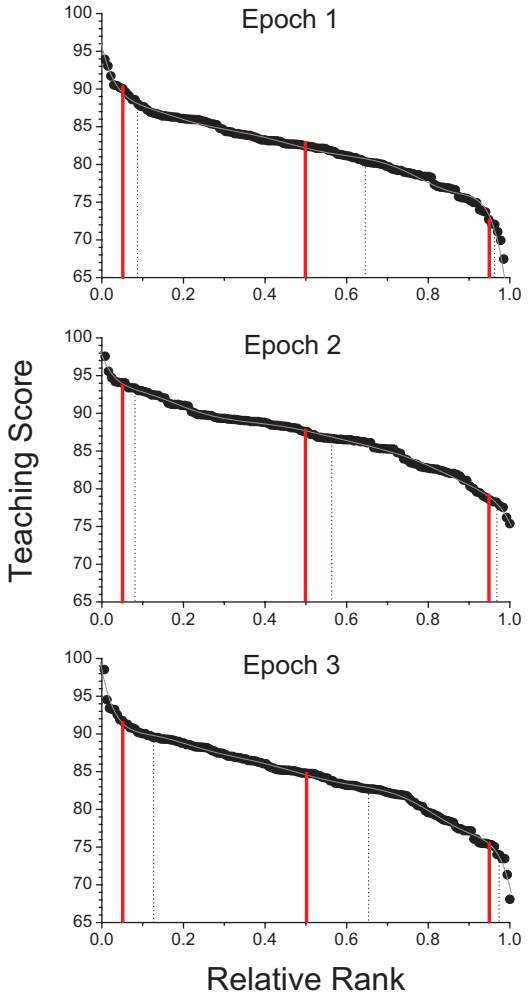

Fig. 3. Rank order based on teaching score is only slightly affected by retaliation. Faculty members are rank ordered (normalized to the total number of faculty members evaluated in each epoch) based on mean scaled (0 to 100) clinical teaching score received in each epoch (*black circles*). *Red vertical lines* denote the 5th, 50th, and 95th percentile rank positions. *Black dotted vertical lines* to the right of each *red line* denote the rank order that would be obtained if the faculty member had reduced their mean resident clinical performance score by one full SD and if retaliation had occurred on all evaluations by the amount detected using only complete dyads in each epoch. The data points are fit by a ninth order polynomial (*gray curve*).

lack of a negative seniority effect (epoch 3) can occur when clinical teaching improves.[28]

Our failure to find a retaliation signal in our macroscopic analysis given a positive microscopic retaliation effect is likely due to incomplete dyads in the macroscopic dataset. In epochs 1, 2, and 3, we had complete dyads for 43.8%, 34.5%, and 54.2% of our evaluations (table 4). The very small retaliation signal was likely confined to these complete dyads and diluted by null-resident and null-faculty evaluations.

It is important to acknowledge that our dataset contains many evaluations per faculty member, which buffer infrequent retaliation events. With a smaller number of evaluations, a single retaliation event would have a larger effect. Prospect theory[42] has shown that people respond to gains (a positive evaluation in this case) and losses (a negative evaluation in this case) asymmetrically such that losses are perceived as far more costly than are equally sized gains. This means that very small retaliation effects may have larger psychologic effects than are justified by the numeric size of the effects.

Lack of Sex Effects on Evaluation

We found no effect of sex on resident assignment of teaching scores to faculty. Previous studies have found mixed results with some demonstrating higher teaching scores for male faculty members[22,23] and others demonstrating higher teaching scores for female faculty members.[24] We analyzed interactions of sex to see if male (or female) residents assigned different scores to male or female faculty members and found no interaction for any combinations.

Grade Inflation Occurs as Residents Progress through Residency but This Is Not Accompanied by Higher Teaching Scores

Our faculty members inflate clinical performance scores of more senior residents.[11] Our relative to peers scoring system defines 3 as peer average at all times during residency. Thus, the average assigned score should remain 3 as a resident advances through residency. During epoch 1, residents saw their actual scores. During this epoch, faculty members assigned increasingly inflated scores to more senior residents, while more senior residents were assigning lower clinical teaching scores to the faculty ($P < 0.001$). Thus, grade inflation did not lead to higher teaching scores.

Limitations of This Study

Our results are from a single residency and may not generalize to other programs. We designed our system to keep evaluator identity confidential; thus, residents did not explicitly know who evaluated them making retaliation more difficult. Our faculty members[11] assign inflated scores, and these scores were seen by residents in epoch 1. Thus, residents may not have perceived a need to retaliate. However, in epochs 2 and 3, we did not reveal scores to residents, and residents still did not retaliate to any important degree. Another limitation is the large number of evaluations we received; consequently, a negative evaluation would be diluted by other evaluations. Our study did not address the difficulties of maintaining confidentiality with a small program. Although we found little evidence of retaliation, we did not actually address whether grade inflation would be reduced if the faculty had this knowledge. Our study also did not analyze the information contained in the comments that residents and faculty members wrote. Comments are important to the process of evaluation and feedback and potentially to the process of retaliation; thus, they will need to be studied in the future.

Conclusions and Practical Implications

Our results provide reassurance to medical educators who worry about the consequences of assigning low clinical performance

scores to residents. We found no relationship between clinical performance scores that faculty members assigned to residents and clinical teaching scores that residents assigned to faculty members using our confidential system. This means that hawks and doves[29] do not receive different teaching scores as a result of their grading characteristics. When we analyzed faculty member–resident dyads using a mixed effects model, we detected only a very small retaliation effect in each epoch. This suggests that programs can use confidential evaluations with written feedback as a strategy to minimize retaliation. The lack of important retaliation documented in our study should encourage faculty members to be more forthright and provide more developmentally useful feedback to residents. Our results should also encourage faculty members to provide appropriate (less grade inflated) evaluations. This will allow the evaluation process to more accurately denote performance level and allow trainees to benefit from more authentic and less inflated evaluations.

Acknowledgments

The authors thank both the faculty members who spent time and effort evaluating residents' clinical skills and the residents who spent time and effort evaluating faculty members' teaching skills. The authors also thank Mary Wright, Ph.D., Director of the Sheridan Center for Teaching and Learning (Providence, Rhode Island) and Adjunct Assistant Professor, Department of Sociology, Brown University (Providence, Rhode Island), for feedback on the manuscript.

Research Support

Support was provided solely from institutional and/or departmental sources.

Competing Interests

The authors declare no competing interests.

Correspondence

Address correspondence to Dr. Baker: Department of Anesthesia, Critical Care and Pain Medicine Massachusetts General Hospital, 55 Fruit Street, Boston, Massachusetts 02114. khbaker@partners.org. Information on purchasing reprints may be found at www.anesthesiology.org or on the masthead page at the beginning of this issue. Anesthesiology's articles are made freely accessible to all readers, for personal use only, 6 months from the cover date of the issue.

References

1. Curran DS, Stalburg CM, Xu X, Dewald SR, Quint EH: Effect of resident evaluations of obstetrics and gynecology faculty on promotion. J Grad Med Educ 2013; 5:620–4
2. Shearin KK: Grade inflation. Science 1976; 191:340
3. Fazio SB, Papp KK, Torre DM, Defer TM: Grade inflation in the internal medicine clerkship: A national survey. Teach Learn Med 2013; 25:71–6
4. Cacamese SM, Elnicki M, Speer AJ: Grade inflation and the internal medicine subinternship: A national survey of clerkship directors. Teach Learn Med 2007; 19:343–6
5. Walsh J: Does high school grade inflation mask a more alarming trend? Science 1979; 203:982
6. Anonymous: Against grade inflation. Nature 2004; 431:723
7. Fighting grade inflation. Science 1994; 264:1255
8. Weaver CS, Humbert AJ, Besinger BR, Graber JA, Brizendine EJ: A more explicit grading scale decreases grade inflation in a clinical clerkship. Acad Emerg Med 2007; 14:283–6
9. Speer AJ, Solomon DJ, Fincher RM: Grade inflation in internal medicine clerkships: Results of a national survey. Teach Learn Med 2000; 12:112–6
10. Love JN, Deiorio NM, Ronan-Bentle S, Howell JM, Doty CI, Lane DR, Hegarty C; SLOR Task Force: Characterization of the Council of Emergency Medicine Residency Directors' standardized letter of recommendation in 2011–2012. Acad Emerg Med 2013; 20:926–32
11. Baker K: Determining resident clinical performance: Getting beyond the noise. Anesthesiology 2011; 115:862–78
12. Redding RE: Students' evaluations of teaching fuel grade inflation. Am Psychol 1998; 53:1227–8
13. Maurer TW: Cognitive dissonance or revenge? Student grades and course evaluations. Teaching Psychol 2006; 33:176–9
14. Clayson DE, Frost TF, Sheffet MJ: Grades and the student evaluation of instruction: A test of the reciprocity effect. Acad Manag Learn Educ 2006; 5:52–85
15. Gardner AK, Scott DJ: Repaying in kind: Examination of the reciprocity effect in faculty and resident evaluations. J Surg Educ 2016; 73:e91–e94
16. Shute VJ: Focus on formative feedback. Rev Educ Res 2008; 78:153–89
17. Archer JC: State of the science in health professional education: Effective feedback. Med Educ 2010; 44:101–8
18. Benitez-Quiroz CF, Wilbur RB, Martinez AM: The not face: A grammaticalization of facial expressions of emotion. Cognition 2016; 150:77–84
19. Keating, CF: The life and times of nonverbal communication theory and research: Past, present, future. In: Matsumoto DR, Hwang HS, Frank MG, eds., American Psychological Association: APA Handbook of Nonverbal Communication, 1st edition. Washington, DC, American Psychological Association, 2016; 17–42
20. Johnson G, Connelly S: Negative emotions in informal feedback: The benefits of disappointment and drawbacks of anger. Hum Rel 2014; 67:1265–90
21. Tarpley JL, Tarpley MJ: The continuing quest for meaningful faculty evaluations of residents. JAMA Surg 2016; 151:31
22. Beckman TJ, Mandrekar JN: The interpersonal, cognitive and efficiency domains of clinical teaching: Construct validity of a multi-dimensional scale. Med Educ 2005; 39:1221–9
23. de Groot J, Brunet A, Kaplan AS, Bagby M: A comparison of evaluations of male and female psychiatry supervisors. Acad Psychiatry 2003; 27:39–43
24. Fluit CR, Feskens R, Bolhuis S, Grol R, Wensing M, Laan R: Understanding resident ratings of teaching in the workplace: A multi-centre study. Adv Health Sci Educ Theory Pract 2015; 20:691–707
25. Watson NC: Likert or not, we are biased. Anesthesiology 2012; 116:1160; author reply 1161–2
26. Norman G: Likert scales, levels of measurement and the "laws" of statistics. Adv Health Sci Educ Theory Pract 2010; 15:625–32
27. Stevens SS: On the theory of scales of measurement. Science 1946; 103:677–80
28. Baker K: Clinical teaching improves with resident evaluation and feedback. Anesthesiology 2010; 113:693–703
29. McManus IC, Thompson M, Mollon J: Assessment of examiner leniency and stringency ('hawk-dove effect') in the MRCP(UK) clinical examination (PACES) using multi-facet Rasch modelling. BMC Med Educ 2006; 6:42
30. White CB, Fantone JC: Pass-fail grading: Laying the foundation for self-regulated learning. Adv Health Sci Educ Theory Pract 2010; 15:469–77

31. Dweck CS: Motivational processes affecting learning. Am Psychol 1986; 41:1040–8
32. Dobrow SR, Smith WK, Posner MA: Managing the grading paradox: Leveraging the power of choice in the classroom. Acad Manag Learn Educ 2011; 10:261–76
33. Lipnevich AA, Smith JK: Effects of differential feedback on students' examination performance. J Exp Psychol Appl 2009; 15:319–33
34. Lumley T, Diehr P, Emerson S, Chen L: The importance of the normality assumption in large public health data sets. Annu Rev Public Health 2002; 23:151–69
35. Barrett JP, Goldsmith L: When is n Sufficiently Large? Am Stat 1976; 30: 67–70
36. Sullivan LM, D'Agostino RB: Robustness of the t test applied to data distorted from normality by floor effects. J Dent Res 1992; 71:1938–43
37. Boneau CA: The effects of violations of assumptions underlying the test. Psychol Bull 1960; 57:49–64
38. Ratcliffe JF: The effect on the t distribution of non-normality in the sampled population. J R Stat Soc Ser C Appl Stat 1968; 17:42–8
39. Cohen J: Statistical Power Analysis for the Behavioral Sciences, 2nd edition. Hillsdale, New Jersey, L. Erlbaum Associates, 1988
40. Cohen J: A power primer. Psychol Bull 1992; 112:155–9
41. Baker K, Sun H, Harman A, Poon KT, Rathmell JP: Clinical performance scores are independently associated with the American Board of Anesthesiology Certification Examination Scores. Anesth Analg 2016; 122:1992–9
42. Kahneman D: A perspective on judgment and choice: Mapping bounded rationality. Am Psychol 2003; 58:697–720

ANESTHESIOLOGY REFLECTIONS FROM THE WOOD LIBRARY-MUSEUM

Jetting after Dr. Douglas Sanders: From "Valley Forge" to New Haven to the Wood Library-Museum

During World War II, a medical inventor and artist named R. Douglas Sanders, M.D. (1906 to 1977), served as Anesthetist-in-Chief at Valley Forge General Hospital in Phoenixville, Pennsylvania. Less than 8 miles from where that facility once operated, I graduated from Ursinus College, Collegeville, Pennsylvania—the year that Dr. Sanders passed away. Fifteen years later in New Haven, Connecticut, as an assistant professor at Yale, I became a second-generation trainee on the Sanders jet injector, learning from a Sanders-trained nurse anesthetist named Michael Johnston. Later, as an honorary curator, I facilitated the acquisition of a watercolor (*above*), Sanders' *Late Blooming*, for the Wood Library-Museum of Anesthesiology. From "Valley Forge" to New Haven to the Wood Library-Museum, I seem to have spent my life chasing after this accomplished anesthesiologist, artist, and inventor, Dr. R. Douglas Sanders. (Copyright © the American Society of Anesthesiologists' Wood Library-Museum of Anesthesiology.)

George S. Bause, M.D., M.P.H., Honorary Curator and Laureate of the History of Anesthesia, Wood Library-Museum of Anesthesiology, Schaumburg, Illinois, and Clinical Associate Professor, Case Western Reserve University, Cleveland, Ohio. UJYC@aol.com.