Combating Grade Inflation in Nephrology Clinical Rotation Evaluations Using Faculty Education and a 5-Point Centered Rating Scale

Christina M. Yuan, MD
Robert Nee, MD
Kevin C. Abbott, MD, MPH
James D. Oliver III, MD, PhD

ABSTRACT

Background From 2010 to 2011, more than 70% of the clinical rotation competency evaluations for nephrology fellows in our program were rated "superior" using a 9-point Likert scale, suggesting some degree of "grade inflation."

Objective We sought to assess the efficacy of a 5-point centered rotation evaluation in reducing grade inflation.

Methods This retrospective cohort study of the impact of faculty education and a 5-point rotation evaluation on grade inflation was measured by superior item rating frequency and proportion of evaluations without superior ratings. The 5-point evaluation centered performance at the level expected for stage of training. Faculty education began in 2011–2012. The 5-point centered evaluation was introduced in 2012–2013 and used exclusively thereafter. A total of 68 evaluations, using the 9-point Likert scale, and 63 evaluations, using the 5-point centered scale, were performed after first-year fellow clinical rotations. Nine to 12 faculty members participated yearly.

Results Faculty education alone was associated with fewer superior ratings from 2010–2011 to 2011–2012 (70.5% versus 48.3%, $P = .001$), declining further with 5-point centered scale introduction (2012–2013; 48.3% versus 35.6%; $P = .012$). Superior ratings declined with 5-point centered versus 9-point Likert scales (37.3% versus 59.3%, $P = .001$), specifically for medical knowledge, patient care, practice-based learning and improvement, and professionalism. On logistic regression, evaluations without superior scores were more likely for 5-point centered versus 9-point Likert scales (adjusted odds ratio [aOR] = 8.26; 95% CI 1.53–44.64; $P = .014$) and associated with faculty identifier (aOR= 1.18; 95% CI 1.03–1.35; $P = .013$), but not fellow identifier or training year quarter.

Conclusions Grade inflation was reduced with faculty education and the 5-point centered evaluation scale.

Introduction

Many internal medicine subspecialty programs use end-of-rotation evaluations based on the recently modified American Board of Internal Medicine (ABIM) FasTrack 9-point Likert scale to assess trainee performance in the 6 Accreditation Council for Graduate Medical Education (ACGME) competencies: medical knowledge (MK), patient care (PC), interpersonal communication skills (ICS), professionalism (PROF), systems-based practice (SBP), and practice-based learning and improvement (PBLI). This ordinal item rating scale defines "superior" as 7 to 9, "satisfactory" as 4 to 6, and "unsatisfactory" as 1 to 3.[1] Validity may be reduced by grade inflation and poor interrater reliability.[2] Validity and reliability improve by employing an optimal number of response categories (4 to 7 for a Likert-type scale, with larger numbers adding little value) and "anchoring" descriptions for each response category.[3]

The ACGME Milestone Project requires that rotation evaluations meaningfully assess whether trainees are progressively improving and meeting competency milestones. Assessment is not peer comparison, but demonstrates individual objective milestone attainment.[4] In 2015, the ABIM introduced the ACGME Milestone reporting worksheet, which uses a 9-point Likert scale with anchoring descriptions of milestone progress as the annual trainee assessment.[5] This evaluation schema has not yet been validated.

In 2010, clinical rotation assessments for nephrology fellows at the Walter Reed National Military Medical Center demonstrated grade inflation. In the academic year (AY) 2010–2011, 70.5% of item assessments in the 6 competencies, using a 9-point Likert scale, were superior ratings (ie, in the 7 to 9 range). Fellows expected "superior" ratings.

To address this problem, we conducted faculty education regarding grade inflation. We also developed a 5-point rotation assessment anchored for each

DOI: http://dx.doi.org/10.4300/JGME-D-15-00218.1

Editor's Note: The online version of this article contains the evaluation form used in the study.

response category, which centers trainees performing at the level expected for their stage of training (ie, meeting milestones) at response category 3. This is 1 of the measurement tools used in our curricular milestone schema and informs clinical competency committee decisions regarding milestone achievement.[6,7]

What was known and gap
Grade inflation, with a majority of learners being given "superior" ratings, is common in graduate medical education.

What is new
A 5-point centered scale and faculty education reduced the percentage of superior ratings.

Limitations
Single site, single specialty study limits generalizability; dual intervention makes attribution of effect complex.

Bottom line
Faculty development and use of a 5-point centered scale reduced grade inflation.

Methods

Before AY 2012–2013, our clinical rotation evaluation was based on the ABIM FasTrack 9-point Likert scale, with anchor descriptions at the lowest and highest scale categories, and a scale item for each of the 6 competencies (14 additional items assessed nephrology-specific performance in physical examination, transplant management, renal replacement therapy, outpatient clinic, transitions of care, nephrology procedures, etc). The MK item is shown in FIGURE 1A. The ordinal rating scale for each item defined superior as 7 to 9 ("far exceeds reasonable expectations"), satisfactory as 4 to 6 ("always meets and occasionally exceeds reasonable expectations"), and unsatisfactory as 1 to 3 ("consistently falls short of reasonable expectations and does not show progress").[1] This rating scale will be referred to as the "9-point Likert" scale.

Faculty Education

In AY 2011–2012, clinical faculty received education from the program director regarding end-of-rotation grade inflation. Faculty raters were encouraged to give 4 to 5 ratings in each competency as baseline satisfactory performance early in training. As fellows advanced through training, this baseline would allow higher scores to be used to indicate progressive improvement and milestone achievement.

5-Point Centered Rating Form

In July 2012, we adopted a 5-point rating scale, centered at 3, defined as satisfactory performance for level of training. Ratings 4 and 5 indicated performance above level of training. Unsatisfactory ratings (1 and 2) and ratings of 5 required written explanations. The MK item is shown in FIGURE 1B, and the entire evaluation form is provided as online supplemental material. In addition to the 6 competency items, 5 items for assessment of transitions of care, outpatient clinic, transplantation, renal replacement therapy, and nephrology-related procedures were included. This rating scale will be referred to as the "5-point centered" scale.

At twice-yearly formative evaluations, fellows were assured that a 3 rating represented satisfactory performance for level of training (ie, milestones were

A. **Medical Knowledge** *(Question 2 of 20–Mandatory)*

Limited knowledge of basic and clinical sciences; minimal interest in learning; does not understand complex relations, mechanisms of disease.

Exceptional knowledge of basic and clinical sciences; highly resourceful development of knowledge; comprehensive understanding of complex relationships, mechanisms of disease.

No Interaction	Unsatisfactory			Satisfactory			Superior		
0	1	2	3	4	5	6	7	8	9

B. **Medical Knowledge** *(Question 1 of 13–Mandatory)*

Does the fellow display knowledge of basic and clinical sciences, interest in learning, and an understanding of complex relationships and mechanisms in nephrologic disease?

Not Observed	Not acceptable. Must address in comments.	Below average. Identify areas that require improvement in comments.	Satisfactory. At expected level for this degree of experience.	Excellent. Above the level normally seen for this degree of experience.	Superior performance at level rarely seen for this degree of experience. Must address in comments.
0	1	2	3	4	5

FIGURE 1
Example of Medical Knowledge Item for 9-Point Likert Scale (A) and 5-Point Centered Scale (B)

TABLE 1
Timeline of Grade Inflation Project

Training Year	No. of Evaluations[a]	No. of Faculty Evaluators	No. of First-Year Fellows	Event
2010–2011	35 All 9-point Likert	9	4	Recognition of grade inflation with 9-point Likert scale
2011–2012	30 All 9-point Likert	9	3	Counseling of attending staff development of 5-point centered scale
2012–2013	39 5-point centered (36) 9-point Likert (3)	12	3	Introduction of new 5-point centered evaluation scale Counseling of attending staff Counseling of fellows
2013–2014	27 All 5-point centered	9	3	Continuation of new 5-point evaluation scale Counseling of attending staff Counseling of fellows

[a] The number of evaluations represents the number of faculty evaluations of first-year fellow clinical rotations performed at Walter Reed National Military Medical Center, the primary training site. Six items were assessed per evaluation, 1 for each of the ACGME competencies.

being met), and that absence of superior ratings did not indicate poor performance. Faculty were informed that a 3 should be the most frequent rating given to a successful fellow, that 4 indicated that milestones were being met earlier than expected, and that a rating of 5 required explanation. Evaluations were not to be referenced to peer performance, but to individual progress in meeting milestones. Thus, a graduating fellow "ready for unsupervised practice" in a given competency would have the same rating (3) as a successful fellow in the first month of clinical training.

Data Collection

Evaluations were programmed and distributed using medical evaluation software (E*Value, Advanced Informatics, Minneapolis, MN). Scores were accessed using the Trainee Reports/Aggregate Performance search feature, filtered by start date, end date, type of rotation (activity), primary training site, faculty evaluator, evaluation type, and trainee cohort. All faculty evaluations of first-year fellows during inpatient and outpatient rotations at the primary training site were reviewed between AY 2010–2011 and AY 2013–2014. Item ratings in each competency based on academic year, fellow cohort, rotation block, rotation type, faculty, and type of evaluation (9-point Likert versus 5-point centered) were entered in a Microsoft Excel (Microsoft Corp, Redmond, WA) spreadsheet.

The study was determined exempt from Institutional Review Board review and approved by the Walter Reed National Military Medical Center Department of Research Protections. The manuscript was approved by the Walter Reed National Military Medical Center Department of Research Programs and Office of Public Affairs.

Data Analysis

For each evaluation, 6 items were evaluated, 1 for each competency. AY 2011–2012 was the reference year when the project began and the 9-point Likert evaluation was in use. Data are presented as absolute numbers, proportions, or percentages. Descriptive statistics were performed in Microsoft Excel. Comparisons were made as appropriate using Fisher exact test (QuickCalcs, GraphPad Software Inc, La Jolla, CA). $P = .05$ was considered to be significant.

Logistic regression was performed in Stata SE 12.1 (StataCorp LP, College Station, TX), with a binary outcome variable "no superior score" versus "$\geq$ 1 superior scores" in the 6 competency items for each evaluation. There were a total of 131 evaluations (observations). Nonsuperior scores were defined as less than 7 to 9 for the 9-point Likert scale and less than 4 to 5 for the 5-point centered scale. Independent variables (covariates) were attending faculty anonymous identifier, fellow anonymous identifier, AY quarter (with the first quarter as the reference), and type of evaluation (9-point Likert or 5-point centered). Of 131 evaluations, 32 (24.4%) did not have a superior score in any of the 6 competency items. Ninety-nine (75.6%) had 1 or more superior scores in the 6 competency items.

Fellow performance was independently assessed between AY 2010–2011 and 2011–2012 versus AY 2012–2013 and 2013–2014, by determining the percentage of first-year fellow chart audit deficiencies for each time period, as previously described.[8]

Results

TABLE 1 shows the number of first-year fellow evaluations by academic year at the primary training site, the number of faculty, the number of entering first-year fellows, and yearly project events. Seven

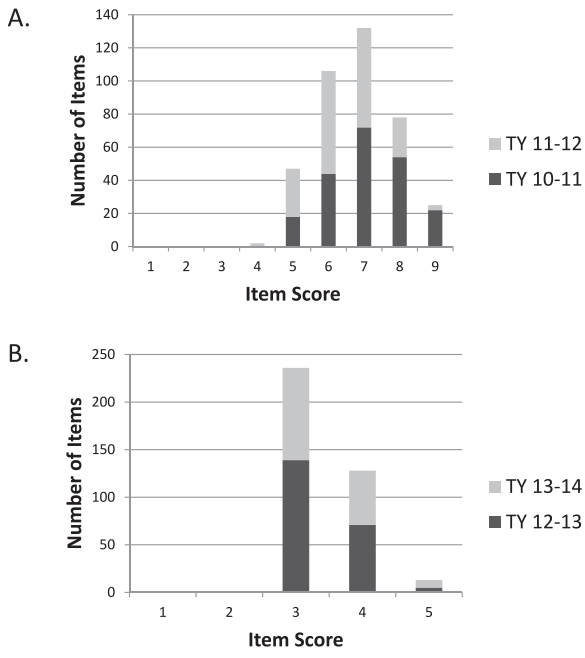

FIGURE 2
Distribution of 6 Core Competency Item Scores for 9-Point Likert Scale Evaluation Form (A) and 5-Point Centered Evaluation Form (B) by Academic Year

faculty did evaluations in all 4 training years. There was minimal overlap between the evaluation types. Before AY 2012–2013, all evaluations were 9-point Likert. Three 9-point Likert evaluations were done early in AY 2012–2013. Subsequently, all evaluations were 5-point centered. Rating distribution for each evaluation type is shown in FIGURE 2. The item rating distribution for the 9-point Likert scale in 2010–2011 and 2011–2012 was confined to a 6-point spread (4 to 9) centered at 7 (FIGURE 2A).

Faculty education alone was associated with fewer superior ratings from 2010–2011 to 2011–2012 (70.5% versus 48.3%, $P = .001$), declining further after 5-point centered scale introduction (2012–2013; 48.3% versus 35.6%; $P = .012$; FIGURE 3). There were 68 nine-point Likert evaluations (408 items total), and 63 five-point centered evaluations (378 items total). A total of 242 (59.3%) 9-point Likert scale evaluation items were rated superior versus 141 (37.3%) with the 5-point centered scale ($P = .001$). The proportion of evaluations without superior scores increased significantly in AY 2012–2013, with the introduction of the 5-point centered scale (FIGURE 3). Among 68 nine-point Likert evaluations, only 7 (10.3%) had no superior score in any of the 6 competency items, while 25 of 63 (39.7%) 5-point centered evaluations had no superior score ($P = .001$). Three of 7 faculty who did evaluations in all 4 study years were the authors. There was no difference in percentage of evaluations without superior scores between authors (n = 3, 7 of 47, 14.9%) versus non-authors (n = 4, 14 of 53, 26.4%; $P = .22$).

The percentage of superior ratings in MK, PC, PBLI, and PROF significantly declined in association with the 5-point centered scale (TABLE 2). This was most marked for MK, where superior ratings decreased from 42.6% to 14.5% ($P = .001$).

The percentage of first-year fellow chart audit deficiencies between 2010–2011 and 2011–2012 versus 2012–2013 and 2013–2014 were not different (15.1% versus 14.5%, $P = .62$), suggesting that the

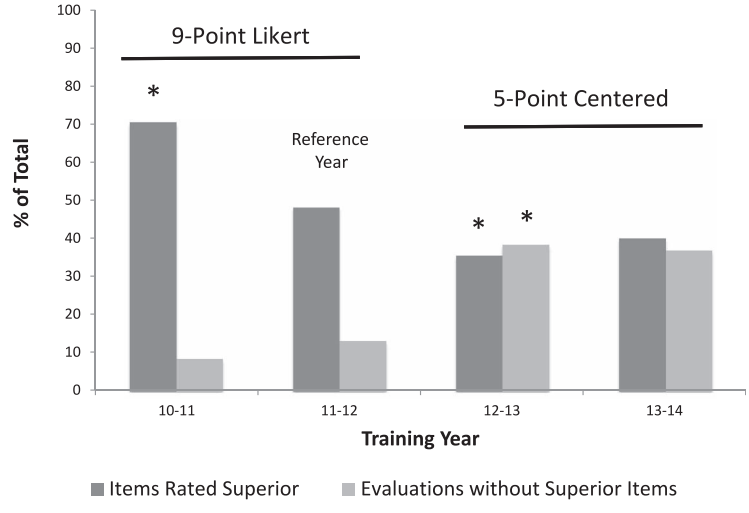

FIGURE 3
Superior Rating Evaluation Patterns by Academic Year (2011–2012 is Reference Year in Which Faculty Education Began)

Note: The 5-point centered scale was introduced at the beginning of academic year 2012–2013.

* $P = .05$ versus reference year. Items rated superior academic year (AY) 2011–2012 versus AY 2010–2011 ($P = .001$). Items rated superior AY 2011–2012 versus AY 2012–2013 ($P = .012$). Evaluations without superior items AY 2011–2012 versus AY 2012–2013 ($P = .029$).

TABLE 2
Proportion of Superior Ratings for Each Competency Item

Competency	MK	PC	PBLI	SBP	ICS	PROF
9-point Likert scale, N = 68	29 (42.6%)	44 (64.7%)	40 (58.8%)	33 (48.5%)	40 (58.8%)	56 (82.4%)
5-point centered scale, N = 63	9 (14.3%)	23 (36.5%)	21 (33.3%)	20 (31.7%)	32 (52.4%)	35 (55.5%)
	$P < .001$	$P = .002$	$P = .005$	$P = .07$	$P = .49$	$P = .001$

Note: Data are shown as percentage of items rated 7 to 9 (9-point Likert) or 4 to 5 (5-point centered). Comparisons performed using Fisher exact test. Two-tailed *P* values are shown, with $P = .05$ considered significant.
Abbreviations: MK, medical knowledge; PC, patient care; PBLI, practice-based learning and improvement; SBP, systems-based practice; ICS, interpersonal communication skills; PROF, professionalism.

decline in superior scores was unrelated to performance differences between the 2 cohorts.

On logistic regression, evaluations without superior scores in any competency were significantly associated with the 5-point centered versus the 9-point Likert scale (adjusted odds ratio [aOR] = 8.26; 95% CI 1.53–44.64; $P = .014$). Evaluations without superior scores were also associated with a faculty identifier (aOR = 1.18; 95% CI 1.03–1.35; $P = .013$), but not with a fellow identifier or academic year quarter.

Discussion

Faculty education and the introduction of a 5-point centered end-of-rotation evaluation were associated with significant declines in superior scores, specifically for MK, PC, PBLI, and PROF. It is not possible to differentiate the relative contributions of faculty education and the 5-point centered evaluation. Education alone was associated with 22% absolute reduction in superior item ratings, but a significant increase in evaluations without superior scores occurred only after 5-point centered evaluation introduction.

Evaluations without superior scores were significantly associated with the 5-point centered scale on logistic regression. However, superior scores did not decline significantly for SBP and ICS, and 37% of items received superior scores after full implementation of education and the 5-point centered evaluation. Regardless, these interventions were associated with significant reductions in grade inflation and could be applied to other evaluation scenarios, such as mini-clinical evaluation exercises.[9]

The 5-point centered anchoring statements were designed to reassure faculty and trainees that response category 3 describes completely satisfactory performance for stage of training, while not falsely suggesting attainment of "ready for unsupervised practice" or aspirational milestones.[4] Receiving this type of rating may make trainees more prepared to focus on deficiencies. The requirement that extreme superior ratings (response category 5) have written explanations served as a disincentive. On logistic regression, evaluations without superior scores were associated with faculty identifier, indicating differences in rating strictness or leniency.[10] Fellow identifier was not a significant predictor for evaluations without superior scores, suggesting that the decline in superior scores was not due to poor fellow performance. This is supported by unchanged first-year fellow outpatient chart audit deficiencies before and after 5-point evaluation introduction.[8]

Grade inflation is a tenacious problem for medical educators, due to leniency bias, halo effect, desire to reward well-liked trainees, and unpleasant message avoidance.[10–12] Fifty-five percent of internal medicine clerkship directors reported difficulty with grade inflation.[11] The standardized letter of recommendation for residency applicants to emergency medicine programs placed 40.1% of potential trainees in the top 10%.[13] Varney et al[2] devised a criteria-based, anchored evaluation system for their internal medicine residency program, resulting in a significant decline in superior ratings. Before intervention, their residents expected scores of 8 to 9 on the 9-point Likert scale.

The 9-point Likert scale is associated with grade inflation for residents and practicing physicians.[10] Our most frequent item score was 7, the mean rating given by program directors to graduating nephrology fellows who passed the ABIM nephrology examination.[14] The 9-point Likert scale may have too many categories.[3,10] Four to 7 categories are considered optimal, and extreme categories may complicate ratings. This is demonstrated by the frequency distribution of our 9-point Likert item scores, essentially confined to item scores of 5 to 9 (FIGURE 2A).

Limitations of our study include its retrospective study design, a single training program, a small number of trainees and faculty, and a relatively short study period, with all limiting the ability to generalize our findings. Given time, grade inflation may reoccur despite faculty education and 5-point centered scale implementation.

Future investigation should be directed to prospectively evaluating the individual effects of faculty education and the 5-point centered score on grade

inflation, and the introduction of an integrated set of evaluation tools to sufficiently assess clinical competency and milestone achievement, above and beyond traditional end-of-rotation appraisals.[7]

Conclusion

Faculty education and the introduction of a 5-point centered end-of-rotation evaluation were associated with significant declines in superior scores. End-of-rotation evaluations by faculty should not be used as the sole determinant of milestone achievement due to many inherent biases, including grade inflation, which are independent of instrument design and resistant to faculty training.[10] Faculty education may moderate these biases, but cannot completely remove them.

References

1. American Board of Internal Medicine. FasTrack Clinical Competence Evaluation System. http://www.abim.org/program-directors-administrators/fastrack.aspx. Accessed February 5, 2016.
2. Varney A, Todd C, Hingle S, Clark M. Description of a developmental criterion-referenced assessment for promoting competence in internal medicine residents. *J Grad Med Educ*. 2009;1(1):73–81.
3. Lozano LM, García-Cueto E, Muñiz J. Effect of the number of response categories on the reliability and validity of rating scales. *Methodology*. 2008;4(2):73–79.
4. Accreditation Council for Graduate Medical Education; American Board on Internal Medicine. The Internal Medicine Subspecialty Milestones Project. October 2014. https://www.acgme.org/acgmeweb/Portals/0/PDFs/Milestones/InternalMedicineSubspecialtyMilestones.pdf. Accessed February 4, 2016.
5. American Board of Internal Medicine. FasTrack. http://www.abim.org/program-directors-administrators/fastrack.aspx. Accessed February 4, 2016.
6. Yuan CM, Nee R, Abbott KC, Oliver JD III. Milestones for nephrology training programs: a modest proposal. *Am J Kidney Dis*. 2013;62(6):1034–1038.
7. Yuan CM, Prince LK, Oliver JD III, Abbott KC, Nee R. Implementation of nephrology subspecialty curricular milestones. *Am J Kidney Dis*. 2015;66(1):15–22.
8. Yuan CM, Prince LK, Zwettler AJ, Nee R, Oliver JD III, Abbott KC. Assessing achievement in nephrology training: using clinic chart audits to quantitatively screen competency. *Am J Kidney Dis*. 2014;64(5):737–743.
9. American Board of Internal Medicine. Mini-CEX. http://www.abim.org/program-directors-administrators/assessment-tools/mini-cex.aspx#competencies. Accessed February 4, 2016.
10. Williams RG, Klamen DA, McGaghie WC. Cognitive, social, and environmental sources of bias in clinical performance ratings. *Teach Learn Med*. 2003;15(4):270–292.
11. Dudas RA, Barone MA. Setting standards to determine core clerkship grades in pediatrics. *Acad Pediatr*. 2014;14(3):294–300.
12. Fazio SB, Papp KK, Torre DM, Defer TM. Grade inflation in the internal medicine clerkship: a national survey. *Teach Learn Med*. 2013:25(1):71–76.
13. Love JN, Deiorio NM, Ronan-Bentle S, Howell JM, Doty CI, Lane DR, et al. Characterization of the Council of Emergency Medicine Residency Directors' standardized letter of recommendation in 2011–2012. *Acad Emerg Med*. 2013;20(9):926–932.
14. Shea JA, Norcini JJ, Kimball HR. Relationships of ratings of clinical competence and ABIM scores to certification status. *Acad Med*. 1993;68(suppl 10):22–24.

All authors are with the Nephrology Service, Department of Medicine, Walter Reed National Military Medical Center. **Christina M. Yuan, MD,** is Associate Program Director; **Robert Nee, MD,** is Staff Nephrologist; **Kevin C. Abbott, MD, MPH,** is Program Director; and **James D. Oliver III, MD, PhD,** is Chief, Nephrology.

Funding: The authors report no external funding source for this study.

Conflict of interest: The authors declare they have no competing interests.

The views expressed in this report are those of the authors, and do not reflect the official policy of the Department of the Army, the Department of the Navy, the Department of Defense, or the US government.

Corresponding author: Christina M. Yuan, MD, Walter Reed National Military Medical Center, Nephrology SVC, Department of Medicine, 8901 Wisconsin Avenue, Bethesda, MD 20814, 301.295.4330, fax 301.295.6081, christina.m.yuan.civ@mail.mil

Received May 23, 2015; revision received September 13, 2015; accepted November 12, 2015.