Effect of Trainee Performance Data on Standard-Setting Judgments Using the Mastery Angoff Method

Stuart B. Prenner, MD
William C. McGaghie, PhD
Sarah Chuzi, MD
Eric Cantey, MD
Aashish Didwania, MD
Jeffrey H. Barsuk, MD, MS

ABSTRACT

Background Mastery learning in health professions education requires learners to learn and undergo assessment until they demonstrate a high level of competence. Setting defensible standards is key to accurately assessing educational outcomes in mastery learning. The Mastery Angoff method was proposed recently to set a minimum passing standard (MPS) for mastery learning curricula. However, it is unknown whether prior knowledge of trainee performance affects judges' decisions about setting an MPS using the Mastery Angoff method.

Objective We sought to determine the effect of introducing baseline data about trainee performance on faculty judges' decisions about the Mastery Angoff MPS for a written examination.

Methods We developed a mastery learning curriculum to train internal medicine residents and cardiology fellows about the correct interpretation of inpatient telemetry monitoring. All learners were required to meet or exceed an MPS on a 35-item written examination at the end of training. The MPS was set in 2017 by judges who used the item-based Mastery Angoff method without prior examinee performance information. The judges subsequently reevaluated the test items after receiving baseline data about trainee performance collected during pilot testing. Mastery Angoff MPSs with and without baseline performance data were compared.

Results Twelve judges participated in the standard-setting exercise. The initial MPS was similar to the repeat MPS set after judges received trainee performance data (86.2% versus 86.9%, $P = .23$).

Conclusions Prior knowledge about medical trainee performance data did not affect MPS as determined by the Mastery Angoff procedure.

Introduction

Advancement and promotion in health professions education schools have historically been based on completion of training programs of fixed duration, with learning outcomes measured using normal distributions to evaluate performance. Medical trainees in most settings graduate and begin independent practice after a fixed training period, frequently without rigorous assessment or benchmarked documentation of competence to safely practice medicine. This results in variation in clinical skills, which can lead to unsafe patient care practices.[1–6] In response, the Accreditation Council for Graduate Medical Education (ACGME) mandated the use of educational milestones, where resident physicians must reach a set proficiency level before graduating.[7]

Mastery learning provides a rigorous framework to standardize the process of documenting and reaching milestones, and it requires trainees to meet or exceed a minimum passing standard (MPS) before completing training. Learners initially unable to meet this standard participate in more training until they reach the MPS.[8] Setting defensible standards is critical for program accountability and for assuring learner readiness for independent practice.[9]

Several standard-setting methods have been described, including the Angoff, Hofstee, Contrasting Groups, and Mastery Angoff.[10,11] The Angoff and Hofstee methods have traditionally been used for setting standards in most health care–related mastery learning curricula.[12] However, these methods consider the borderline learner (Angoff) or a minimum and maximum failure rate (Hofstee), and are not appropriate for assessment when patient safety is a concern. Yudkowsky and colleagues[11] argued that, when determining an MPS in a mastery learning environment, judges should be asked to consider the performance of a trainee who is "well prepared to succeed at the next stage of instruction or practice." Setting an acceptable failure rate or an expected pass rate is inappropriate, because all learners are expected to pass with sufficient, high-quality training.

A new approach, the Mastery Angoff, has been proposed where judges rate each assessment item, while considering a trainee who is well prepared to

DOI: http://dx.doi.org/10.4300/JGME-D-17-00781.1

perform safely with minimal or no supervision. Prior research has shown that giving judges baseline performance data from former trainees affects judgments during standard-setting exercises using traditional Angoff and Hofstee procedures.[11,13] The effect of giving judges prior examinee performance data during Mastery Angoff standard setting is not known.

We sought to determine the effect of giving faculty judges examinee baseline data during standard setting using the Mastery Angoff method for an inpatient telemetry knowledge examination.

What was known and gap
Setting defensible standards in mastery learning is key to assessing educational outcomes in physicians.

What is new
Clinicians used the Mastery Angoff method to set a passing standard use for telemetry, and then received prior performance data. The researchers assessed whether the prior performance data changed the judges' minimum passing standard.

Limitations
Single institution, single specialty study limits generalizability.

Bottom line
Prior knowledge about medical trainee performance data did not affect minimum passing standards determined by the Mastery Angoff procedure.

Methods

We developed a mastery learning curriculum designed to teach Northwestern Memorial Hospital internal medicine (IM) residents and cardiology fellows how to interpret inpatient telemetry reports. Northwestern Memorial Hospital is an academic, tertiary care hospital in Chicago, Illinois, with 891 beds. Telemetry monitoring capability is available at 260 beds. We developed written pretraining and posttraining examinations to assess resident knowledge about telemetry use and interpretation. Faculty with telemetry expertise were recruited to participate in a standard-setting exercise to set an MPS for the posttraining telemetry examination using the Mastery Angoff procedure. Two standard-setting exercises were performed: (1) faculty made judgments using only their expectations and knowledge about IM residents and cardiology fellows, and (2) judges also were given telemetry performance data about postgraduate year 3 (PGY-3) residents and cardiology fellows from an examination pilot test. The PGY-3 residents and cardiology fellows did not have experience with a formal telemetry curriculum before the pilot test. We compared the results of the 2 standard-setting exercises and evaluated the number of judges who changed their decisions about telemetry items based on performance data.

Telemetry Curriculum

The mastery telemetry curriculum required trainees to take a written pretest, watch a video that demonstrated proper use and interpretation of telemetry monitoring, participate in deliberate practice with a telemetry monitoring device, and interpret telemetry output with feedback from faculty. Trainees then took a written posttest on which they needed to meet or exceed an MPS. Trainees who did not meet the MPS participated in further deliberate practice and data interpretation.

We wrote 71 multiple-choice questions about the proper interpretation of telemetry, in accordance with examination development guidelines.[14] Content included indications for use and discontinuation of telemetry, identification of artifacts, and interpretation of various types of clinical arrhythmias. Questions were reviewed for content and clarity by 6 attending cardiologists and were administered to 30 PGY-3 IM residents at the end of the 2017 academic year. The Kuder–Richardson Formula 20 coefficient of 0.75 denoted acceptable internal consistency. Using item performance from this pilot, a separate 36-item pretest and 35-item posttest were created that were equivalent in content and difficulty.[15,16] The 35-item posttest was subsequently administered to 14 cardiology fellows in various years at the end of an academic year.

Standard-Setting Exercise

The Mastery Angoff standard-setting method was used to establish the MPS for the posttest. Board-certified attending physicians from the Northwestern University Feinberg School of Medicine were asked to participate as judges, based on their experience supervising trainees and their expertise with the interpretation of telemetry monitoring. Judges were trained using the methods described by Norcini and Guille.[17,18] This involved (1) defining the qualities of the examination and examinees; (2) educating the judges about the pass/fail consequences of their decisions; (3) discussing the purpose of the telemetry evaluation and what constitutes adequate and inadequate skill; (4) defining the learner who is well prepared to succeed; and (5) group practice, feedback, and discussion.

Judges were asked to set standards for each posttest item in 2 iterations and were informed that the consequence of poor trainee performance would be more training, and that trainees could retake the posttest as needed. During the first iteration, judges used the Mastery Angoff method. In the second standard-setting iteration, which immediately

followed the first, judges received performance data in the form of the separate percentages of residents and fellows answering each item correctly from the pilot testing, as they reconsidered scores using the Mastery Angoff method. Judges were informed that performance data were obtained from PGY-3 IM residents and cardiology fellows who had not received formal telemetry instruction. Judges were asked to provide new judgments but could not revise their original judgments.

The Northwestern University Institutional Review Board approved this study, granting exempt status.

Analysis

We report the baseline performance scores of residents and fellows, and judge demographic data as means and SDs. The 2 iterations of MPS item data were compared using a paired *t* test. Statistical analyses were performance using STATA version 14 (StataCorp LLP, College Station, TX).

Results

Mean resident and fellow baseline performances on the posttest items were 73% (SD = 10) and 81% (SD = 8) correct, respectively. Twelve judges participated in the standard-setting exercises. Average time in current IM specialty or subspecialty was 9.8 years (SD = 9.1), with an average time supervising resident physicians of 11.3 years (SD = 9.4). A total of 9 of 12 judges had participated in prior standard-setting exercises; their average number of prior standard-setting experiences was 2.2 (SD = 2.0).

Judges' MPSs for the 2 iterations are shown in the TABLE. The first iteration (without test performance data) was set at 86.2%, which required 31 of 35 items to be correct on the posttest. The second iteration, informed by performance data, yielded an MPS of 86.9% and also required 31 of 35 test items to be correct ($P = .23$). Two of the 12 judges did not change any items from the first to the second iteration. Of the 10 judges who changed at least 1 item, the overall MPS decreased for 3 judges (88.7% versus 87.2%) and increased for 7 judges (85.1% versus 86.9%). All MPSs required 31 of 35 test items correct.

On average, 2 judges changed their assessment for any given question. However, among the 8 questions with greatest change between the 2 judgment iterations (up or down), an average of 5 judges changed their assessment. Of the 4 questions with the greatest increase in score after viewing baseline data, average trainee performance was 97.5%, whereas on the 4 questions with greatest decrease in score, average reported trainee performance was 49.5%.

Of the pilot group, 28 of 30 untrained PGY-3 IM residents and all 14 untrained cardiology fellows would not have met the MPS.

Discussion

This study shows that knowledge of trainee performance on a posttest assessment of telemetry skills resulted in no significant change to the MPS using the Mastery Angoff method.

Another study at our institution about the effect of giving judges trainee performance data concerning an MPS for central venous catheter insertion showed that, while baseline data led to an increase in an MPS set by the Angoff method, the changes were minimal and did not affect the number of trainees passing the exercise.[13] In contrast, in a subsequent study in which longitudinal data presented to judges showed that trainee pretest performance improved significantly over time, and many trainees exceeded the MPS without training, judges increased their subsequent MPS substantially.[19]

Standard-setting exercises of the US Medical Licensing Examination provided an additional opportunity to examine the impact of reality feedback on judge assessment. Judges participating in Angoff-style standard-setting exercises of Step 1 and Step 2 Clinical Knowledge examinations were provided with real trainee performance data as well as performance data that had been purposely manipulated. Judges were found to significantly modify their assessments based on performance data, regardless of whether performance data were real or manipulated. This suggests that judges seem to defer to the data in all cases, suggesting reliance on performance data may supersede content expertise.[20] One unanswered question from the study was the impact of providing performance data on the pass rate of the examination.

This study found that judgment based on expert knowledge of content and learners can be overwhelmed by performance data, particularly when the initial score was less than 100% for an item. The greatest change in item assessment occurred for questions where the initial item score was lower than 100%, and when reported performance data were either very high or low. However, when initial assessment approached 100%, even low trainee performance data did not alter the assessment significantly, reinforcing the concept that some items are in a "must know" category, and low performance data merely expose a knowledge gap. Ultimately, as in most mastery learning approaches, the overall MPS may be set high enough that such extreme performance data will not change the overall pass rate.

TABLE
Minimum Passing Standard Iterations 1 and 2 (Without and With Performance Data, Respectively) on a 35-Item Assessment, and Number of Items Changed by Each Judge

Judge	Minimum Passing Standard Iteration 1, % (No. Correct)	Minimum Passing Standard Iteration 2, % (No. Correct)	No. of Items Changed
1	83.9 (30)	86.1 (31)	7
2	90.6 (32)	90.6 (32)	0
3	82.1 (29)	82.1 (29)	0
4	91.9 (33)	92.9 (33)	5
5	86.2 (31)	89.5 (32)	10
6	77.7 (28)	82.1 (29)	4
7	86.1 (31)	86.4 (31)	10
8	82.9 (30)	83.4 (30)	0
9	91.7 (33)	89.3 (32)	12
10	86.7 (31)	87.7 (31)	8
11	82.9 (29)	81.4 (29)	10
12	91.4 (32)	90.9 (32)	19

The Mastery Angoff standard-setting method has been used more recently in health professions settings where patient safety is paramount.[11] In these situations, giving judges anchoring data may be irrelevant for several reasons. Anchoring data from traditionally trained participants' baseline tests do not necessarily inform performance prediction, and baseline tests typically have a low pass rate in mastery settings.[11] Learners may retrain several times between examinations and may retake a posttest several times before passing. When setting a mastery standard, item relevance is more important than item difficulty. Simply because a group of learners performed poorly on an item does not make that item less important. By contrast, judges may deem an item so essential to a clinical task that 100% of trained learners need to perform it correctly to pass. Such performance deficits expose a curriculum gap, rather than anchoring judge assessments. This was confirmed in the current study, as numerous items received judge scores of near 100%, even after judges learned actual trainee performance was much lower. Finally, anchoring data may be less useful when not connected to subsequent testing performance or actual performance in a clinical setting. Evidence-based approaches to mastery standard setting have shown that performance data are particularly valuable when the data link information about past examinees' success or failure to *subsequent* learning experiences or actual clinical performance.[21]

This study has several limitations. It was performed at 1 institution using 1 data set and a single panel of judges, and we did not collect data about subsequent clinical performance. We did not measure the stability of the 2 iterations with the Mastery Angoff over time, and we did not evaluate the credibility and reliability of the Mastery Angoff method in this study. Further study is needed to show that the Mastery Angoff technique yields valid and reliable data.

Conclusion

Judges participating in a mastery MPS-setting exercise were provided with baseline data from IM residents and cardiology fellows. Knowledge of actual trainee test performance led to minimal, nonsignificant changes in judges' scoring, and the overall number of test items needed to pass was unchanged. Therefore, we do not believe showing baseline data is important when using the Mastery Angoff technique.

References

1. McQuillan RF, Clark E, Zahirieh A, et al. Performance of temporary hemodialysis catheter insertion by nephrology fellows and attending nephrologists. *Clin J Am Soc Nephrol*. 2015;10(10):1767–1772.
2. Barsuk JH, Cohen ER, Feinglass J, et al. Residents' procedural experience does not ensure competence: a research synthesis. *J Grad Med Educ*. 2017;9(2):201–208.
3. Barsuk JH, Cohen ER, Caprio T, et al. Simulation-based education with mastery learning improves residents' lumbar puncture skills. *Neurology*. 2012;79(2):132–137.
4. Birkmeyer JD, Finks JF, O'Reilly A, et al. Surgical skill and complication rates after bariatric surgery. *N Engl J Med*. 2013;369(3):1434–1442.
5. Wayne DB, Barsuk JH, O'Leary KJ, et al. Mastery learning of thoracentesis skills by internal medicine residents using simulation technology and deliberate practice. *J Hosp Med*. 2008;3(1):48–54.

6. Barsuk JH, Cohen ER, Nguyen D, et al. Attending physician adherence to a 29-component central venous catheter bundle checklist during simulated procedures. *Crit Care Med*. 2016;44(10):1871–1881.
7. Nasca TJ, Philibert I, Brigham T, et al. The next GME accreditation system—rationale and benefits. *N Engl J Med*. 2012;366(11):1051–1056.
8. McGaghie WC. Mastery learning: it is time for medical education to join the 21st century. *Acad Med*. 2015;90(11):1438–1441.
9. Cusimano MD. Standard setting in medical education. *Acad Med*. 1996;71(suppl 10):112–120.
10. Downing SM, Tekian A, Yudkowsky R. Procedures for establishing defensible absolute passing scores on performance examinations in health professions education. *Teach Learn Med*. 2006;18(1):50–57.
11. Yudkowsky R, Park YS, Lineberry M, et al. Setting mastery learning standards. *Acad Med*. 2015;90(11):1495–1500.
12. McGaghie WC, Issenberg SB, Barsuk JH, et al. A critical review of simulation-based mastery learning with translational outcomes. *Med Educ*. 2014;48(4):375–385.
13. Wayne DB, Barsuk JH, Cohen E, et al. Do baseline data influence standard setting for a clinical skills examination? *Acad Med*. 2007;82(suppl 10):105–108.
14. Linn RL, Gronlund NE. *Measurement and Assessment in Teaching*. 8th ed. Upper Saddle River, NJ: Prentice Hall; 2000.
15. Issenberg SB, McGaghie WC, Brown DD, et al. Development of multimedia computer-based measures of clinical skills in bedside cardiology. In: Melnick DE, ed. *The Eighth International Ottawa Conference on Medical Education and Assessment Proceedings. Evolving Assessment: Protecting the Human Dimension*. Philadelphia, PA: National Board of Medical Examiners; 2000:821–829.
16. Downing SM. Reliability: on the reproducibility of assessment data. *Med Educ*. 2004;38(9):1006–1012.
17. Kane MT. Validating the performance standards associated with passing scores. *Rev Educ Res*. 1994;64(3):425–461.
18. Norcini J, Guille R. Combining tests and setting standards. In: Norman GR, van der Vleuten CPM, Newble DI, eds. *International Handbook of Research in Medical Education*. Dordrecht, the Netherlands: Kluwer Academic Publishers; 2002:811–834.
19. Cohen ER, Barsuk JH, McGaghie WC, et al. Raising the bar: reassessing standards for procedural competence. *Teach Learn Med*. 2013;25(1):6–9.
20. Clauser BE, Mee J, Baldwin SG, et al. Judges' use of examinee performance data in an Angoff standard-setting exercise for a medical licensing examination: an experimental study. *J Educ Meas*. 2009;46:309–407.
21. O'Malley K, Keng L, Miles J. From Z to A: using validity evidence to set performance standards. In: Cizek GJ, ed. *Setting Performance Standards: Foundations, Methods, and Innovations*. 2nd ed. New York, NY: Routledge; 2012:301–322.

Stuart B. Prenner, MD, is a Fellow, Advanced Heart Failure and Transplant Cardiology, Hospital of the University of Pennsylvania; **William C. McGaghie, PhD,** is Professor, Department of Medical Education, Northwestern University Feinberg School of Medicine; **Sarah Chuzi, MD,** is Chief Resident, Department of Medicine, Northwestern University Feinberg School of Medicine; **Eric Cantey, MD,** is Chief Resident, Department of Medicine, Northwestern University Feinberg School of Medicine; **Aashish Didwania, MD,** is Associate Professor and Internal Medicine Residency Program Director, Department of Medicine, Northwestern University Feinberg School of Medicine; and **Jeffrey H. Barsuk, MD, MS,** is Professor, Departments of Medicine and Medical Education, Northwestern University Feinberg School of Medicine.

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

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

Corresponding author: Stuart B. Prenner, MD, Perelman Center for Advanced Medicine, South Pavilion, 11th Floor, 3400 Civic Center Boulevard, Philadelphia, PA 19104, 267.593.0115, stuart.prenner@uphs.upenn.edu

Received October 23, 2017; revision received January 7, 2018; accepted January 17, 2018.