Article

Setting Mastery Learning Standards

Rachel Yudkowsky MD, MHPE, Yoon Soo Park, PhD, Matthew Lineberry, PhD,

Aaron Knox, MD, and E. Matthew Ritter, MD

Academic Medicine. 2015. 90(11): 1495-1500.

R. Yudkowsky is associate professor, Department of Medical Education, and director, the Dr. Allan L. and Mary L. Graham Clinical Performance Center, University of Illinois at Chicago College of Medicine, Chicago, Illinois.

Y. S. Park is assistant professor, Department of Medical Education, University of Illinois at Chicago College of Medicine, Chicago, Illinois.

M. Lineberry is assistant professor, Department of Medical Education, University of Illinois at Chicago College of Medicine, Chicago, Illinois.

A. Knox is resident, plastic and reconstructive surgery, University of British Columbia Faculty of Medicine, Vancouver, British Columbia, Canada.

E. M. Ritter is associate professor, vice chairman for education, and program director for the general surgery residency, Norman M. Rich Department of Surgery, Uniformed Services University of the Health Sciences F. Edward Hébert School of Medicine/Walter Reed National Military Medical Center, Bethesda, Maryland.

Correspondence should be addressed to Rachel Yudkowsky, Department of Medical Education, 986 CMET, University of Illinois at Chicago College of Medicine, 808 S. Wood Street MC 591, Chicago IL 60612. Phone: 312-996-3598; e-mail: rachely@uic.edu.

Abstract

Mastery learning is an instructional approach in which educational progress is based on demonstrated performance, not curricular time. Learners practice and retest repeatedly until they reach a designated mastery level; the final level of achievement is the same for all, although time to mastery may vary. Given the unique properties of mastery learning assessments, a thoughtful approach to establishing the performance levels and metrics that determine when a learner has demonstrated mastery is essential.

Standard-setting procedures require modification when used for mastery learning settings in health care, particularly regarding the use of evidence-based performance data, the determination of appropriate benchmark or comparison groups, and consideration of patient safety consequences. Information about learner outcomes and past performance data of learners successful at the subsequent level of training can be more helpful than traditional information about test performance of past examinees. The marginally competent "borderline student" or "borderline group" referenced in traditional item-based and examinee-based procedures will generally need to be redefined in mastery settings. Patient safety considerations support conjunctive standards for key knowledge and skill subdomains and for items that have an impact on clinical outcomes. Finally, traditional psychometric indices used to evaluate the quality of standards do not necessarily reflect critical measurement properties of mastery assessments. Mastery learning and testing are essential to the achievement and assessment of entrustable professional activities and residency milestones. With careful attention, sound mastery standard-setting procedures

can provide an essential step towards improving the effectiveness of health professions education, patient safety, and patient care.

Mastery learning is an instructional approach in which educational progress is based on demonstrated performance rather than curricular time.[1] Learners are provided with terminal objectives and performance metrics, opportunities for study and practice, and repeated formative testing with feedback about their progress towards performance goals. Learners cannot advance to the next curricular module, stage of training, or level of practice until the predetermined performance levels are achieved.[1,2] A key characteristic of mastery testing is the ability to retest on multiple occasions to reach a designated "mastery" level; the final level of achievement is the same for all learners, although some learners may require more time and more test attempts than others. Mastery learning and testing can be important elements of competency-based curricula,[2] and are integral to the achievement and assessment of core entrustable professional activities (EPAs)[3] in undergraduate medical education and of milestones[4] during residency.

A thoughtful and rigorous approach to standard-setting, to establish the performance metrics that determine when a learner has demonstrated mastery, is essential to mastery learning. While traditional standards target minimal competence, the goal of mastery learning is to ensure that *all* learners are *well prepared to succeed* in subsequent stages of training. If standards are set too low, students will not be well prepared to succeed; if standards are set too high, students will expend unnecessary effort that could be better spent on other educational goals.

Standards, also called cut scores, pass/fail scores, or minimum passing levels, can be *normative*, as in requiring a score above 1.5 standard deviations below the mean examinee score, or *criterion-based* (also called *absolute*)—for example obtaining a score of 80% correct.[5,6] Normative standards, in which a learner's pass/fail status depends on the performance of other members of the group, have no place in competency-based curricula or mastery settings. Criterion-based standards, on the other hand, are especially appropriate for competency-based curricula in health professions education, providing public accountability towards licensure and certification.[7] Defensible standards are those determined through a systematic approach to capturing the opinions of trained content experts who are familiar with the learners and the inferences to be made about the learners, the test and the scoring method, the standard-setting procedure, and consequences resulting from the selected standard.[5,6]

Competency-based curricula frequently use traditional standard-setting procedures such as Angoff,[8] Hofstee,[9] borderline, or contrasting-groups.[6] While criterion-based methods are appropriate for mastery settings, the central inference of mastery standards— that they predict success in subsequent training or practice—demands an evidence-based approach.[10] Evidence can include the use of predictive past performance data, information about the consequences of different standards for future performance, the use of targeted reference groups, and consideration of patient safety in clinical settings. Additionally, repeated testing and uniformly high terminal achievement levels can have unique effects on the psychometrics of standards, making it challenging to evaluate their quality.

The purpose of this article is to identify elements of traditional standard-setting procedures that require modification in health care mastery learning settings, focusing on the use of evidence-based information to support mastery decisions. While many of our examples address standard-setting for performance tests of clinical skills, the principles apply equally to written tests administered within a mastery learning approach.

Standard-Setting Procedures

Standard-setting procedures[5,6,11] can be categorized as item-based, examinee-based, or test-based (see below); all elicit the opinions of subject matter experts, usually with some degree of iterative discussion. While the process of gathering expert judgments remains unchanged in mastery settings, the information on which judgments are based should be focused on *predicting future performance*, a type of evidence only rarely used in traditional standard-setting exercises.

Item-based standard setting procedures: predictive performance data

The item-based Angoff method,[6,8] frequently used for written tests and performance checklists, asks judges to predict the performance of the "borderline student," a student who is *just at the edge of minimal competence*. Judges indicate the probability that the borderline student would accomplish each item of a test or checklist correctly. In mastery settings, rather than predicting the behavior of a *minimally* competent student who is *just at the edge* of acceptable performance, judges will be modeling the performance of a student who is *well prepared to succeed* at the next stage of instruction or practice.

Data about past examinees' performance often are used to help judges calibrate item-based judgments.[12,13] Judges frequently refer to percent-correct statistics from past administrations of each test or checklist item to help estimate the probability that a minimally competent examinee would accomplish a particular item. In traditional curricula these statistics are based on a single test administration at the end of the learning unit, which most learners are expected to pass on the first attempt. In a mastery environment, on the other hand, the first test may have a very low pass rate. Learners may retake the exam a variable number of times—some will choose to retest early and often, others will wait until they have mastered most of the material. Eventually—after two, three, five, ten retests—they will reach the mastery level and move on. Which test results should be used to inform the judges?

When setting standards in the context of a mastery learning approach, item *difficulty* is less important than item *relevance* or *importance*. If a given item is important for learners to master prior to progressing to the next stage of learning or clinical practice, knowing that in the past only 50% of learners accomplished that item does not make the item any less important. Such a finding should be interpreted as a gap in curriculum and instruction that needs to be closed, not as cause to lower the mastery standard.

An evidence-based approached to mastery standards implies that performance data are most valuable when the data include information about past examinees' success or failure in *subsequent* learning experiences.[14] Suppose a cohort of residents completed a basic laparoscopic skills assessment on a simulator, and then completed a number of basic laparoscopic procedures on patients. Analyses showing how scores on the simulation-

based assessment predict examinees' performance on actual patients could be very useful to judges—for example, showing that examinees with four or more instrument collisions on the simulation-based assessment have a significantly elevated risk of unsafe behaviors during patient care would suggest that fewer than four instrument collisions be one of the criteria for advancement. Similarly, for preclinical written exams, predictive performance data might include the test performance of the subset of students who were subsequently successful in the preclinical curriculum overall.

As another example of the evidence base that can inform standard-setting judges about the impact of different standards on future performance, see Supplemental Digital Figure 1 at [LWW: INSERT LINK]. This figure shows a data display for a hypothetical simulation-based lumbar puncture (LP) training program. The figure's top panel shows past performance data typically provided during standard setting exercises; the bottom panel provides an example of predictive performance data. Suppose all learners had to score at least 80% on an LP checklist at the end of training, and that learners were then certified to perform LPs into the indefinite future. By showing how participants' immediate post-training scores relate to the same learners' retest scores six months later (shown in the bottom panel), this hypothetical example suggests that the mastery standard of 80% may have been too lenient, as a number of participants who scored below 95% on the post-training assessment earned extremely low scores at 6-month follow-up. Data of this type, clearly displayed, help standard-setting judges estimate the levels of performance needed in early mastery learning modules to ensure safe and effective subsequent learning or patient care activities.

Examinee-based procedures: Identifying appropriate benchmark groups

Examinee-based procedures or methods such as the borderline-group method or the contrasting-groups method[6,11] require judges or external criteria to categorize examinees into groups at contrasting levels of performance—for example, proficient versus non-proficient, or pass/marginal/fail. Group membership is defined by data other than scores on the test in question—for example, by data from direct observation of performance or the use of other relevant criteria. The standard for a particular exam is obtained by determining the test score that best discriminates between the two groups (contrasting-groups method) or the median score of the marginal group (borderline-group method). Examinee-based methods are often used to set mastery standards for instrument-based metrics: measures obtained by a simulator, computer, or other measurement device during dynamic, real-time assessments of performance.[15-17]

Traditional examinee-based methods generally need to be modified to support the "well prepared to succeed" inferences of a mastery setting. The marginally acceptable performance of peers identified by the traditional borderline-group method is not an appropriate final goal for mastery learners; on the other hand, benchmarking the performance of experts may result in standards that are inappropriately high and result in effort expended to little purpose. The "proficient group" approach[18,19] uses the performance scores observed from a developmentally-appropriate benchmark group to guide standard-setting. The proficient group performs a task such as knot tying in an instrumented environment (e.g., a virtual-reality simulator). Their performance data can then be used to guide standard-setting; for instance, judges may deem it appropriate to set

the average "time to secure knot" of the proficient group (second-year residents) as the mastery standard for first-year residents.

A highly proficient or even expert benchmark group may be appropriate for learners transitioning to independent practice. However, experts may perform the task using procedural variants that would be inappropriate and unsafe for early trainees with limited clinical judgment and skills. Experts also may demonstrate behaviors that are not essential for safe practice at earlier stages of training, such as very rapid task performance. Measures of experience alone, such as years of practice, do not well predict acceptable performance.[20] Suitably proficient individuals are best identified based on a combination of clinical experience and scores on an objective measure of performance. The proficient group method has been applied repeatedly in procedural simulation in the simulation lab, operating room, and procedural suite.[21-25]

Comparison groups for contrasting-groups methods used in mastery settings must be chosen with care. Several studies [15,17,26] have compared medical students' performance of basic surgical skills on a simulator to that of practicing surgeons, and derived a cut score that maximally discriminated between the two groups. However, in mastery learning we rarely need assessments that can tell novices from experts; instead, we need assessments that discriminate between novices who are sufficiently competent to move on versus novices who are not, or that distinguish trainees who are not quite ready for unsupervised practice from those who can graduate and practice safely. This requires careful choice of comparison and benchmark groups depending on the stage of training and the specific inferences desired.

Performance data of expert or proficient groups should not form the basis for a mechanistic generation of a standard (e.g., arbitrarily choosing "expert score minus 1.5 standard deviations," or "the point of intersection between experts' and novices' score distributions"). Rather, such data should serve as a point of departure for thoughtful deliberations among standard-setting judges about the importance of each metric for clinical and educational outcomes and the level of performance expected at different transition points. These deliberations are key to setting defensible, effective, and achievable mastery learning standards.[11`]

Test-based procedures

The test-based Hofstee method [6,9] (also called the whole-test method or compromise method) uses a combination of normative and criterion-based standards to ensure that the number of failed learners will be acceptable and the standards therefore implementable. Judges are asked to bracket the cut score by specifying the minimum and maximum acceptable passing scores and the minimum and maximum acceptable failure rates; the final cut score is based on the actual performance of the examinees. Data provided to judges include normative information about the distribution of scores and fail rates at different cut scores. The Hofstee method is arguably inappropriate for setting standards in a mastery context, in which practically all learners are expected to eventually achieve the specified standard and advance to the next phase of training. A curriculum in which required standards are lowered in order to meet constraints of acceptable fail rates would, by definition, be antithetical to a mastery learning approach. While one could pre-set minimum and maximum acceptable fail rates at 0% and 100% for mastery settings,

eliminating fail-rate judgments would remove the essential characteristic of the Hofstee procedure.

Mastery Standards to Support Patient Safety

A key goal of milestones[4] and EPAs[3] is to ensure that learners are well prepared to transition safely and successfully to the next level of clinical training or practice. Mastery learning, often simulation-based, can support this goal by ensuring that students and residents acquire a suitable level of proficiency in skills such as performing invasive procedures on live patients.[27]

Standards must be appropriate to the specific transition under consideration. For example, when setting standards for performing a lumbar puncture on a part-task trainer, we may be interested in whether the trainee is ready to perform the task on a live patient under close supervision later that week, or, more commonly, whether he or she is prepared to perform the procedure well into the future. The usual task in traditional standard-setting exercises is to specify *how much* of the content learners must master to proceed to the next learning experience, for example the number of multiple-choice or procedure checklist items accomplished. However, in consideration of patient safety consequences, judges may wish to specify process variables that indicate *how well* learners must master that content—for example, how quickly knowledge can be retrieved, the time frame in which a procedure must be performed, or evidence of overlearning and automaticity that help predict long-term retention.[21,22,28-30] Here again predictive information such as that shown in Supplemental Digital Figure 1's lower panel can be critical for judges. While skills decay is not unique to mastery learning, it is

especially salient for activities such as procedural skills that are often taught and tested using a mastery approach.

When setting mastery standards for clinical skills, judges should take note of the clinical relevance and patient safety implications of each item. Traditional standard-setting procedures are compensatory across items: as long as examinees achieve the cut score it does not matter which individual items are missed and which are accomplished. In clinical settings, however, the omission or incorrect performance of individual items may have a significant impact on patient safety and outcomes. One approach to setting mastery standards for basic procedural skills is to have judges rate each item as to its impact on dimensions such as patient safety, patient comfort, or procedure outcome, relying on evidence-based data when available; an item whose performance or non-performance has an impact on one of these dimensions can be considered "critical."[27] A similar approach can be taken for standardized patient and mannequin scenario checklists that include many actions that contribute to good outcomes but only a few truly critical actions. Standards can be set separately for critical and non-critical checklist items, such that performance of non-critical items does not compensate for omission or incorrect performance of critical items.[27] Setting this type of conjunctive standard for critical items is also important when assessing maintenance of skills from initial testing to a delayed retest, to avoid having retention of non-critical items mask the decay of critical skills. While conjunctive standards increase the risk of incorrectly classifying a capable student as failing, we may choose to tolerate the higher error rate and require another round of testing in order to avoid the false-positive of passing a student who is clinically unsafe.

Assessment of clinical skills in a simulated environment almost always involves some degree of construct underrepresentation[31] that, combined with the stress and distractions inherent in clinical environments, often leads to a decrement in performance in live-patient settings.[32, 33] Learners who aim for and reach only the traditional standard of "minimal competence" in a simulated environment are at risk of falling below minimal competence on the task as a whole when they attempt to perform it in the real world. High standards for those aspects of knowledge and skill that are measurable, trainable, and essential to successful outcomes will maximize learners' ability to perform adequately in distracting and complex real-life settings.[27]

Evaluating the Quality and Impact of Standards

Evaluating the quality of mastery standards can be challenging. Once a mastery learning system is implemented, it is difficult to obtain comparative data showing that learners who achieve the cut score are successful in the next stage of training and practice while learners who do not reach the passing score are likely to struggle or to be unsafe. When learners who pass the standard are successful, it is difficult to know whether a lower standard might have been sufficient to obtain the desired effect, since allowing learners who did not achieve the standard to progress may not be feasible or, in patient care settings, ethical. Comparison data collected *before* implementation of the mastery learning system may be the best source of evidence that the cut score is appropriately placed.

Reliability metrics for mastery tests are complex, especially when standards are conjunctive, and may require psychometric expertise. Each round of practice and

retesting increases the learners' probability of mastery and decreases the variance of test scores (see Figure 1), resulting in a higher reliability and a decreased standard error of measurement; thus the precision of mastery determination may increase with each retesting. On the other hand, the decreased variance across learners—which may approach zero with repeated testing, since all are achieving the mastery standard—means that traditional reliability metrics will be difficult to interpret and may not be relevant in a mastery setting. See Lineberry et al[34] in this issue of *Academic Medicine* for an in-depth discussion of validity evidence considerations for mastery tests and for reliability issues in particular.

Summing Up

With the advent of EPAs and milestones, medical education continues to move toward a true competency-based educational system in which students and residents are offered repeated opportunities to practice and achieve the skills critical to their future practice. Effective mastery learning within a competency-based curriculum requires a thoughtful, systematic, and evidence-based approach to setting mastery standards. Traditional item-based and examinee-based standard-setting procedures often need to be modified for mastery testing, with particular attention to the use of predictive performance and clinical outcome data and the selection of appropriate benchmark groups. List 1 provides a summary of key considerations when setting standards in a health care mastery setting; Table 1 shows how these might apply to different types of tests.

Evidence relating specific performance metrics and standards to clinical or learning outcomes is rare and sorely needed; psychometric guidelines and best practices under mastery learning conditions are under-researched and a fruitful area for future development. We hope that this article will serve to stimulate additional conversation and research regarding the implications for standard-setting of a mastery learning approach. With careful attention to these issues, mastery standards can provide an essential step towards improving the effectiveness of health professions education, patient safety, and patient care.

Acknowledgments: The authors thank Dr. Barry Issenberg and Dr. William McGaghie for their critical review of an earlier version of the manuscript, and Dr. Jeffrey Barsuk and Elaine Cohen for the use of performance data on which to loosely base our hypothetical example in the bottom panel of Supplemental Digital Figure 1.

Funding/Support : None.

Other disclosures: None.

Ethical approval: Not applicable.

References

1. JH Block, ed. Mastery Learning: Theory and Practice. New York, NY: Holt, Rinehart and Winston; 1971.
2. McGaghie WC, Miller GE, Sajid A, Telder TV. Competency-Based Curriculum Development in Medical Education: An Introduction. Geneva, Switzerland: World Health Organization; 1978. http://whqlibdoc.who.int/php/WHO_PHP_68.pdf. Accessed June 30, 2015.
3. Association of American Medical Colleges. Core Entrustable Professional Activities for Entering Residency: Curriculum Developers' Guide. https://members.aamc.org/eweb/upload/Core%20EPA%20Curriculum%20Dev%20Guide.pdf. Accessed June 30, 2015.
4. Accreditation Council for Graduate Medical Education Milestones. https://www.acgme.org/acgmeweb/tabid/430/ProgramandInstitutionalAccreditation/NextAccreditationSystem/Milestones.aspx. Accessed June 30, 2015.
5. Cizek GJ. Setting Performance Standards: Concepts, Methods, and Perspectives. Mahwah, NJ: Lawrence Erlbaum Associates; 2001.
6. Downing S, Tekian A, Yudkowsky R. Procedures for establishing defensible absolute passing scores on performance examinations in health professions education. Teach Learn Med. 2006;18:50-57.
7. Lammers RL, Davenport M, Korley F, et al. Teaching and assessing procedural skills using simulation: Metrics and methodology. Acad Emerg Med. 2008;15:1079-1087.

8. Angoff WH. Scales, norms, and equivalent scores. In RL Thorndike, ed. Educational Measurement. Washington, DC: American Council on Education; 1971:508-600.
9. Hofstee WKB. The case for compromise in educational selection and grading. In SB Anderson, JS Helmick, eds. On Educational Testing. San Francisco, CA: Jossey-Bass; 1983:107-127.
10. American Educational Research Association, American Psychological Association, National Council on Measurement in Education: Standards for Educational and Psychological Testing; Standard 5.23. Washington, DC: American Educational Research Association, 2014:108-109.
11. Livingston SA, Zieky, MJ. Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests. Princeton, NJ: Educational Testing Service; 1982.
12. Cizek G. Standard-setting guidelines. Educ Measurement: Issues and Practice. 1996;15:12-21.
13. Mee J, Clauser BE, Margolis MJ. The impact of process instructions on judges' use of examinee performance data in Angoff standard setting exercises. Educ Measurement: Issues and Practice. 2013;32:27-35.
14. O'Malley K, Keng L, Miles J. Using validity evidence to set performance standards. In GJ Cizek, ed. Setting Performance Standards. New York: Routledge; 2012:301-322.

15. Konge L, Clementsen P, Larsen KR, Arendrup H, Buchwald C, Ringsted C. Establishing pass/fail criteria for bronchoscopy performance. Respiration 2012; 83:140–146.
16. Madsen ME, Konge L, Nørgaard LN, et al. Assessment of performance measures and learning curves for use of a virtual-reality ultrasound simulator in transvaginal ultrasound examination. Ultrasound Obstet Gynecol. 2014:44:693-699.
17. Konge L, Annema J, Clementsen P, Minddal V, Vilmann P, Ringsted C: Using virtual-reality simulation to assess performance in endobronchial ultrasound. Respiration 2013;86:59–65.
18. Gallagher AG, Ritter EM, Champion H, et al. Virtual reality simulation for the operating room: Proficiency-based training as a paradigm shift in surgical skills training. Ann Surg. 2005;241:364-372.
19. Gallagher AG. Metric-based simulation training to proficiency in medical education: What it is and how to do it. Ulster Med J. 2012;81:107-113.
20. Choudhry NK, Fletcher RH, Soumerai SB. Systematic review: The relationship between clinical experience and quality of health care. Ann Int Med. 2005; 142:260-273
21. Rosenthal ME, Ritter EM, Goova MT, et al. Proficiency-based fundamentals of laparoscopic surgery skills training results in durable performance improvement and a uniform certification pass rate. Surg Endosc. 2010;24:2453-2457.

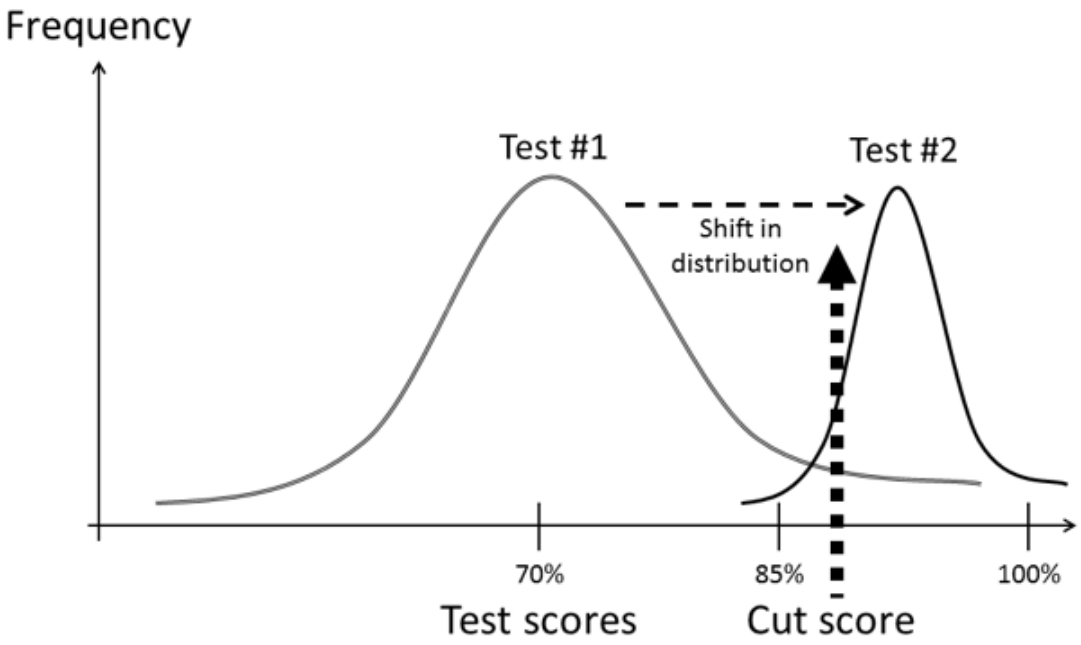
22. Stefanidis D, Korndorffer JR, Jr., Sierra R, Touchard C, Dunne JB, Scott DJ. Skill retention following proficiency-based laparoscopic simulator training. Surgery. 2005;138:165-170.
23. Seymour NE, Gallagher AG, Roman SA, et al. Virtual reality training improves operating room performance: Results of a randomized, double-blinded study. Ann Surg. 2002;236:458-463.
24. Scott DJ, Ritter EM, Tesfay ST, Pimentel EA, Nagji A, Fried GM. Certification pass rate of 100% for fundamentals of laparoscopic surgery skills after proficiency-based training. Surg Endosc. 2008;22:1887-1893.
25. Ahlberg G, Enochsson L, Gallagher AG, et al. Proficiency-based virtual reality training significantly reduces the error rate for residents during their first 10 laparoscopic cholecystectomies. Am J Surg. 2007; 193:797-804.
26. Fraser SA, Klassen DR, Feldman LS, Ghitulescu GA, Stanbridge D, Fried GM. Evaluating laparoscopic skills: Setting the pass/fail score for the MISTELS system. Surg Endosc. 2003;17:964–967.
27. Yudkowsky R, Tumuluru S, Casey P, Herlich N, Ledonne C. A patient safety approach to setting pass/fail standards for basic procedural skills checklists. Simulation in Healthcare. 2014;9:277-282.
28. Stefanidis D, Korndorffer JR, Jr., Markley S, Sierra R, Scott DJ. Proficiency maintenance: Impact of ongoing simulator training on laparoscopic skill retention. J Am Coll Surg. 2006;202:599-603.
29. Stefanidis D, Korndorffer JR, Jr., Markley S, Sierra R, Heniford BT, Scott DJ. Closing the gap in operative performance between novices and experts: Does

harder mean better for laparoscopic simulator training? J Am Coll Surg. 2007; 205:307-313.

30. Stefanidis D, Scerbo MW, Montero PN, Acker CE, Smith WD. Simulator training to automaticity leads to improved skill transfer compared with traditional proficiency-based training: a randomized controlled trial. Ann Surg. 2012;255:30–37.
31. Downing SM, Haladyna TM. Validity threats: Overcoming interference with proposed interpretations of assessment data. Med Educ. 2004;38:327–333.
32. Prabhu A, Smith W, Yurko Y, Acker C, Stefanidis D. Increased stress levels may explain the incomplete transfer of simulator-acquired skill to the operating room. Surgery. 2010;147:640–645.
33. Scerbo MW, Schmidt EA, Bliss JP. Comparison of a virtual reality simulator and simulated limbs for phlebotomy training. J Infus Nurs. 2006;29:214–224.
34. Lineberry M, Park YS, Cook D, Yudkowsky R: Making the case for mastery learning assessments: Key issues in validation and justification, Acad Med. 2015;90:XXX-XXX.

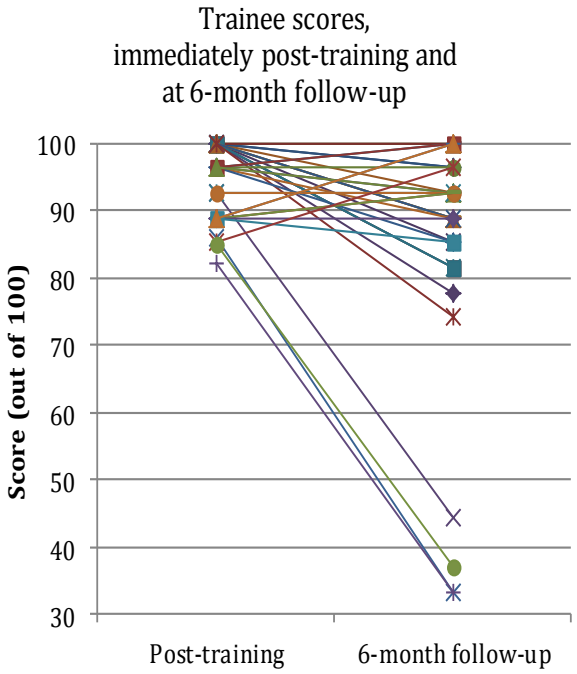
[FIGURE LEGEND]

Figure 1 Group performance distributions shift in a mastery learning setting as each round of practice and retesting increases the learners' probability of mastery.



Supplemental Digital Figure 1

Item	Residents who correctly performed the item, No. (%)	Residents who omitted or incorrectly performed the item, No. (%)
Checked that all necessary equipment is available and ready to use	181 (76)	48 (24)
Opened lumbar puncture kit carefully to maintain sterility	170 (71)	69 (29)
Prepped insertion point with iodine swabs in concentric circles	147 (61)	92 (39)
[Other checklist items, not shown here]	—	—



Correlation between post-training and 6-month follow-up: ***r* = 0.44**

If post-training mastery standard is set at...	Then at 6-month follow up... Minimum score	Mean score
100	74.07	88.10
95	74.07	89.95
90	44.44	88.74
85	33.33	86.20
80	33.33	86.20

Supplemental Digital Figure 1 Performance data examples for standard-setting judges in non-mastery and mastery contexts. The *top panel* presents a hypothetical past-performance data display for a traditional item-based standard setting exercise for lumbar puncture, showing the numbers and percentages of residents who did or did not accomplish each item; data based on performance of 239 residents. The *bottom panel* presents a hypothetical predictive performance data display for a mastery item-based standard-setting exercise for lumbar puncture; data based on performance of 34 residents.

Table 1

Setting Standards for Different Types of Exams in Mastery Learning Settings[a]

Type of exam data	Examples of standard-setting considerations and supporting information in a mastery learning setting [b]
Written exams such as multiple- choice questions Standardized patient checklists or rating scales	*If using a modified Angoff Method:* • As supporting information, use benchmark performance data of students who were successful at later stages of curriculum. • Redefine borderline student from "minimally competent" to "well prepared for next stage." • Consider identifying critical items when patient safety issues are present.
Procedural skills checklists or rating scales Mannequin scenario checklists or rating scales	*If passing the test will put live patients at risk:* • As supporting information, identify subset of items critical to patient safety or procedure outcome (or other salient dimensions). • Note that item difficulty is less salient than item relevance and patient safety implications. • Set standards separately and conjunctively for critical and non-critical items.
Simulator-based performance metrics[c]	*If using borderline group method:* • As supporting information, identify appropriate benchmark group: solidly competent or proficient, rather than marginally or minimally competent. *If using contrasting-groups method:* • As supporting information, identify appropriate "expert" or "passing" group: persons successful at the next stage of training or practice. Avoid contrasting novices with experts.

[a] A mastery learning setting is one in which learners take a variable amount of time to reach a uniformly high achievement standard. Learners may retest multiple times until the standard is achieved.

[b] Standard-setting methods shown are only examples; other standard-setting methods could be selected for the same type of exam data.

[c] Select relevant metrics with care; set mastery standards only for measures that have an impact on live performance.

List 1

Considerations for Setting Mastery Standards

- The inferences and decisions that will be based on the mastery cut score must be clear. What is the "next step" of training or practice? When will it occur? What is the level of supervision at the next step?
- Essential content and, when appropriate, process variables such as speed or automaticity of response needed for a safe and successful transition to the next step, should be identified.
- Absolute or criterion-based standard-setting methods should be used rather than than normative methods.
- Conjunctive rather than compensatory standards are appropriate for key knowledge and skill subdomains and for items that have an impact on patient safety.
- Information about the performance of past examinees, especially first-time test takers, is less helpful than performance of learners at the immediate next level of training or practice.
- Information about expert performance should be used with caution and as part of a thoughtful and deliberative standard-setting process.
- Information relating performance on the test to successful practice at the next stage of training is key to setting evidence-based mastery standards and should be a priority for mastery standards research.
- Traditional psychometric indices used to evaluate the quality of cut scores do not necessarily reflect measurement properties of mastery assessments and should be used with caution.