

# The Two-Stage Examination: A Method to Assess Individual Competence and Collaborative Problem Solving in Medical Students

Janet E. Lindsley, PhD, David A. Morton, PhD, Karly Pippitt, MD, Sara Lamb, MD, and Jorie M. Colbert-Getz, PhD

## Abstract

### Problem

Effectively solving problems as a team under stressful conditions is central to medical practice; however, because summative examinations in medical education must test individual competence, they are typically solitary assessments.

### Approach

Using two-stage examinations, in which students first answer questions individually (Stage 1) and then discuss them in teams prior to resubmitting their answers (Stage 2), is one method for rectifying this discordance. On the basis of principles of social constructivism, the

authors hypothesized that two-stage examinations would lead to better retention of, specifically, items answered incorrectly at Stage 1.

In fall 2014, they divided 104 first-year medical students into two groups of 52 students. Groups alternated each week between taking one- and two-stage examinations such that each student completed 6 one-stage and 6 two-stage examinations. The authors reassessed 61 concepts on a final examination and, using the Wilcoxon signed ranked tests, compared performance for all concepts and for just those students initially missed, between Stages 1 and 2.

### Outcomes

Final examination performance on all previously assessed concepts was not significantly different between the one- and two-stage conditions ( $P = .77$ ); however, performance on only concepts that students initially answered incorrectly on a prior examination improved by 12% for the two-stage condition relative to the one-stage condition ( $P = .02$ ,  $r = 0.17$ ).

### Next Steps

Team assessment may be most useful for assessing concepts students find difficult, as opposed to all content. More research is needed to determine whether these results apply to all medical school topics and student cohorts.

### Problem

Course examinations offer educators opportunities, as well as challenges in assessing, teaching, and providing feedback to learners. Medical schools must be able to demonstrate that each student is individually competent; therefore, typical examinations entail individuals working in isolation. By contrast, real-world problem solving involves collaboration; effective teamwork under stressful conditions is a skill that physicians must master.

### Approach

Student engagement with course material is most intense during examinations, making the assessment period the ideal time for providing immediate and individualized feedback to correct misunderstandings. One strategy that attempts to accomplish all of these goals (ensure individual competence, promote collaborative problem solving, and provide effective individual feedback) is the two-stage examination.

### Description, benefits, previous research, and theory

During two-stage examinations, students first complete the assessment and submit answers individually. Then, working in teams, they answer the same assessment questions again. During the second, teamwork-based stage, students typically engage in a lively discussion and receive immediate feedback from their peers regarding any errors in their problem solving. This two-stage method is commonly used in team-based learning (TBL) to assess readiness prior to a lesson. It is less commonly used in medical schools to assess learning or retention of previously taught content.

At the college level and among other health care professions, faculty have used two-stage examinations.<sup>1-7</sup> Faculty are often convinced of the utility of this examination format after witnessing the intensely productive discussions and full engagement of students that occur during the second stage.<sup>7</sup> Additionally, student satisfaction, according to the results of course surveys, is typically very high.<sup>3,7,8</sup> However, the few investigators who have attempted to determine whether retention of learning is better following two-stage examinations have reported mixed results.<sup>4,6,8,9</sup> One limitation of the prior research is that the results were not parsed out as concepts students initially missed versus concepts students already knew at the first stage.

According to social constructivism, learning is fostered by peer collaboration, which promotes the elaboration of knowledge structures and fosters individual awareness of personal learning processes.<sup>10</sup> By assessing students' individual understanding before moving them into teams, students are well prepared to participate in rich discussion. Examinations motivate students to study and to maximally

Please see the end of this article for information about the authors.

Correspondence should be addressed to Janet E. Lindsley, University of Utah, Department of Biochemistry, 15N. Medical Dr. East RM 4100, Salt Lake City, UT 84112-5650; e-mail: Janet.Lindsley@hsc.utah.edu.

This is an open-access article distributed under the terms of the Creative Commons Attribution-Non Commercial-No Derivatives License 4.0 (CCBY-NC-ND), where it is permissible to download and share the work provided it is properly cited. The work cannot be changed in any way or used commercially.

*Acad Med.* 2016;91:1384-1387.

First published online April 5, 2016

doi: 10.1097/ACM.0000000000001185

engage with course content. During the first encounter with the assessment, students have independently thought about the questions and committed to answers. During the team discussion, students articulate reasons for choosing their answers, which requires that they understand what the question is asking, that they can communicate their logic, and that they ultimately defend or abandon their original answers.<sup>7</sup> Other team members can either agree with an individual student's rationale or explain why they disagree. This discussion provides timely, individualized feedback that addresses each student's misunderstandings. Because all of the students share a mutual goal (correctly solving the problem), they see the immediate benefit of effective teamwork in the two-stage examination.

If students already have sound knowledge structures, then further elaboration from peers may not be necessary for retaining knowledge; thus, we hypothesized that examinations completed in two stages (individual + team) compared with one stage (individual only) would lead to better retention of only concepts (examination items) students missed in the first stage.

### Study design and participants

We used a randomized crossover design (see Figure 1) to determine whether two-stage examinations improved retention of learning in a 17-week foundational sciences

course. The course included 14 multiple-choice examinations. Examinations 1–13 (“preliminary examinations”) covered the prior week’s material, and Examination 14 was a comprehensive final.

In the fall of 2014, we divided 104 first-year medical students from the University of Utah School of Medicine into two even groups (Group A and Group B) based on the alphabetical order of their last names. We further randomly divided each group of 52 students into 13 teams of four for the Stage 2 discussions. Both Group A and Group B took Examination 1 as a two-stage examination to experience the process. For Examinations 2–13, the groups alternated each week between a one-stage (control) and a two-stage (intervention) condition such that, following the introductory week 1 examination, each student completed 6 one-stage and 6 two-stage examinations. We did not use performance data from Examination 1 for any analyses in the study. This study was deemed exempt by the institutional review board of the University of Utah School of Medicine. Because the study was exempt, informed consent was unnecessary. Although students were aware of our research question, they were unaware of our specific hypothesis.

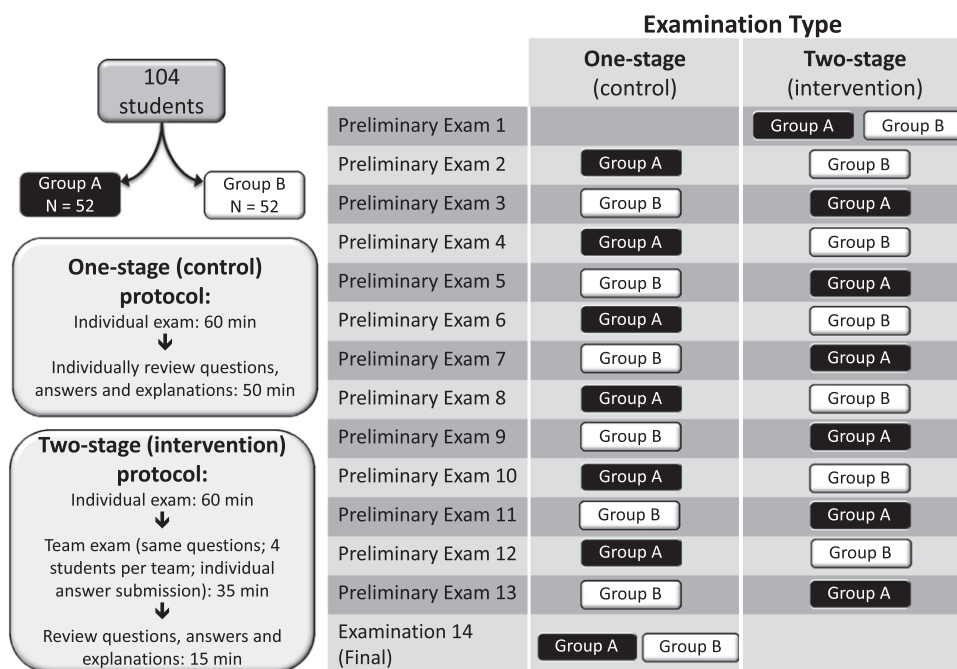
### Examinations

Examinations 1–13 consisted of 25 to 33 multiple-choice questions, and the final

(Examination 14) contained 150 multiple-choice questions. Course content integrated material from multiple disciplines (gross anatomy, biochemistry, embryology, medical ethics, genetics, histology, physiology, and pharmacodynamics/kinetics) each week so that each examination covered a variety of topics. We administered all examinations on iPads using ExamSoft (Dallas, Texas). All questions were tagged for level of cognition—recall, a single step of application/data interpretation, or a research/clinical scenario requiring two or more steps of interpretation/application.

Sixty-one questions on the final examination assessed concepts identical to those on preliminary Examinations 2–13 (the remaining 89 questions assessed related but slightly different concepts, concepts previously assessed on Examination 1, or concepts taught in the final week of the course). Seven of these 61 questions (11%) were recycled from preliminary examinations, while the remaining 54 (89%) were new. Of these 61 final examination items, 5 (8%) required recall, 25 (41%) required one-step application, and 31 (51%) required two or more steps of interpretation/application.

Students had access to laboratory reference values and a metabolic map but no other materials while completing all stages of all these proctored examinations. Group A and Group B completed



**Figure 1** Overview of crossover study design and protocol for a study of the effect of two-stage examinations on knowledge previously taught, University of Utah School of Medicine, 2014.

**Table 1**  
**Demographics for 104 First-Year Medical Students Participating in a Study of Two-Stage Examinations at the University of Utah, 2014**

Student characteristic	Group A, no. (% <sup>a</sup> of 52)	Group B, no. (% of 52)	P value
<b>Gender</b>			.33
Female	23 (44)	28 (54)	
Male	29 (56)	24 (46)	
<b>Race/ethnicity</b>			.14
Asian	6 (12)	7 (13)	
Caucasian	40 (77)	37 (71)	
Underrepresented minority <sup>b</sup>	4 (8)	2 (4)	
Unknown	2 (4)	6 (12)	

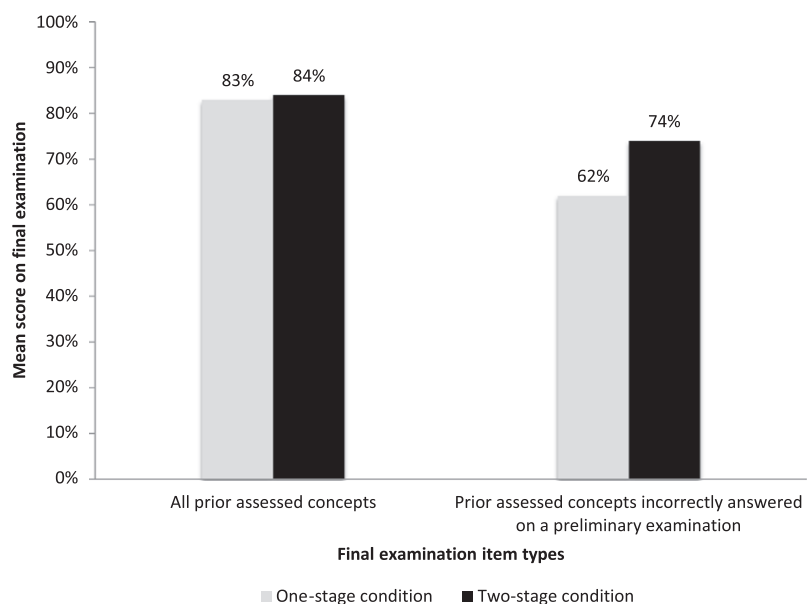
<sup>a</sup>Percentages may not total 100 due to rounding.

<sup>b</sup>Includes all students who self-identified as Indian or Pakistani, Filipino, or Hispanic.

their examinations in separate rooms. For both one-stage and two-stage examinations, students individually selected and submitted answers; therefore, unlike in traditional TBL, we did not require team members to achieve consensus. Scores for the 13 preliminary examinations were worth a total of 36% of the overall course grade for the first individual attempt, and 4% for the second “post-team-discussion” attempt. The final examination was worth 30% of the overall course grade. (The other 30% of the grade came from homework assignments [5%] and anatomy and laboratory quizzes and exams [25%]).

**Data analysis**

We analyzed all data using SPSS version 21 (IBM Corp., Armonk, New York). We calculated frequencies and percentages for demographic variables. We calculated mean Stage 1 performance and mean Stage 2 performance from Examinations 2–13 for each student. To ensure that there were no differences between Group A and Group B, we used Mann–Whitney *U* tests to compare gender and race distributions and average first (individual) attempt scores from Examinations 2–13. One of us (J.E.L.) identified the central concept required to answer each preliminary and final examination question.



**Figure 2** Retention of learning for 104 first-year medical students as measured by concepts assessed both on a final examination and one of six prior two-stage examinations (preliminary examinations completed as an individual and then in a team) or one of six prior one-stage examinations (preliminary examinations completed as an individual only), University of Utah School of Medicine, 2014.

For the final examination, we computed level of cognition frequencies and percentages for each question. We calculated four retention means for each student: final examination performance for all concepts previously assessed on an examination for (1) one-stage conditions and (2) two-stage conditions, and final examination performance just for concepts students initially missed on a preliminary examination for (3) one-stage conditions and (4) two-stage conditions. We compared retention means between one-stage and two-stage conditions using the Wilcoxon signed rank tests. We set alpha at .05 for all statistical tests.

**Outcomes**

**Demographics and results from Examinations 2–13**

Table 1 provides the demographic distributions of the 104 first-year students. Distributions of gender and race were not significantly different between Groups A and B (*P* = .33 and .14, respectively).

The mean individual (first attempt, two-stage; only attempt, one-stage) performance on Examinations 2–13 for Groups A and B were, respectively, 84% (standard deviation [SD] 5%) and 86% (SD 6%). The 2% difference in performance between Groups A and B was not significant (*P* = .06).

For two-stage examination conditions, all students across both groups averaged 85% (SD = 6%) for their first attempt (individual) and 96% (SD = 3%) for their second, postdiscussion attempt on Examinations 2–13.

**Results from the final examination**

When looking at just the 61 items on the final exam that assessed a concept identical to one on a preliminary exam, Group A, on average, missed 3.85 (SD = 2.48) of the 27 items that had been in the two-stage conditions (within the stage-one, individual component) and 3.35 (SD = 1.74) of the 34 items that had been covered in the one-stage conditions on these preliminary exams. Group B missed 3.41 (SD = 2.49) of the 34 items that had been covered in the two-stage conditions (within the Stage 1, individual component) and 2.70 (2.25) of the 27 items that had been covered in the one-stage conditions in Examinations 2–13.

Figure 2 provides final examination performance based on one-stage and

two-stage conditions. Final examination performance for all 61 concepts was not significantly different for the two-stage (mean [M] = 84%, SD = 9%) and one-stage (M = 83%, SD = 9%) conditions ( $Z = 0.29, P = .77$ ). Final performance on only the concepts that students initially answered incorrectly on a prior examination improved 12% (confidence interval 2%–22%) on the final examination for the two-stage (M = 74%, SD = 26%) relative to the one-stage (M = 62%, SD = 35%) condition ( $Z = 2.29, P = .02$ ) with a small effect size ( $r = 0.17$ ).

### Discussion and comparison to TBL

This is the first study to investigate the impact of two-stage examinations on, specifically, medical students' retention of previously learned content. A strength of this study is that we analyzed retention for concepts that students initially missed on an examination based on the hypothesis that the two-stage examination would impact learning of those concepts the most. Prior studies of two-stage examinations used in college-level and other health professions' courses have shown mixed results for retention of material.<sup>4,6,8,9</sup> In the current study, overall retention of learning from a two-stage examination was not different than retention from a one-stage examination. However, for specifically the items that students initially missed on a preliminary examination, we noted a retention benefit for the two-stage examination compared with the one-stage examination.

TBL uses a similar two-stage assessment at the start of the in-class session to ensure that students arrive having studied the assigned content and are prepared for the complex application exercises.<sup>5</sup> While the two-stage examinations described in the current study share several beneficial characteristics with the individual and group readiness assessment tests (iRATs and gRATs) used in TBL, we note one important distinction. In TBL, the iRATs and gRATs are designed to assess *readiness for learning*. In the current study we used two-stage examinations to assess learning of previously taught material, so expecting a higher level of learning (application, analysis, and synthesis of challenging concepts) compared with the expectations of iRATs and gRATs is reasonable. Indeed, very few of the final examination items required simple recall of information. That the items required

application also aligns with our findings that the two-stage examinations have the largest impact on concepts that students initially misunderstood.

### Next Steps

We conducted this study at one institution that uses an organ-systems-based integrated curriculum, and our participants were all students in their first semester of medical school; thus, additional research is needed to confirm the results of our study for other settings and populations. Specifically, research that shows whether our results are stable in more experienced students as their motivation and study habits change would be valuable.

Additionally, the benefit of the two-stage examination compared with the one-stage examination for prior missed concepts produced a small effect size and wide confidence interval. That the mean difference in scores on previously missed concepts between one-stage and two-stage conditions was fairly large may suggest that the types of students who benefit the most from two-stage examinations may vary. However, other factors could also contribute to the wide confidence interval, so future research should identify when, for whom, and for what material two-stage assessment is most beneficial.

The results of this study suggest that difficult concepts are best for team assessment, but more research is needed to determine whether two-stage examinations are feasible in other settings, such as during the clerkship years. Finally, because two-stage examinations require teamwork skills, examining whether those skills positively correlate with the most learning from the team will be of value.

Too often, assessments are used just to determine individual competence. Balancing the need to protect assessment integrity and the need to make assessment a learning experience for students is difficult. Using two-stage examinations for assessment of prior learned content is one innovative method that allows faculty and administrators to determine individual competence while still giving students a chance to solve problems in a team.

*Acknowledgments:* The authors would like to thank Michael Lauder, MA, Amy Peterson, and Justen Harris for the extra work involved in administering the one-stage vs. two-stage

examinations to the two cohorts of students. They are also grateful to Dr. Steven Baumann for his input on the study design.

*Funding/Support:* None reported.

*Other disclosures:* None reported.

*Ethical approval:* This study was approved by the institutional review board at the University of Utah School of Medicine (00074585).

**J.E. Lindsley** is associate professor of biochemistry and assistant dean of curriculum, University of Utah School of Medicine, Salt Lake City, Utah.

**D.A. Morton** is associate professor of neurobiology and anatomy, University of Utah School of Medicine, Salt Lake City, Utah.

**K. Pippitt** is assistant professor and director, Medical Student Education, Department of Family and Preventive Medicine, University of Utah School of Medicine, Salt Lake City, Utah.

**S. Lamb** is associate professor, associate dean of curriculum, and program director, Internal Medicine/Pediatrics Residency Program, University of Utah School of Medicine, Salt Lake City, Utah.

**J.M. Colbert-Getz** is assistant professor, Internal Medicine Administration, and director, Medical Education Research, University of Utah School of Medicine, Salt Lake City, Utah.

### References

- 1 Giuliodori MJ, Lujan HL, DiCarlo SE. Collaborative group testing benefits high- and low-performing students. *Adv Physiol Educ.* 2008;32:274–278.
- 2 Sinner GJ, Briggs JC, Stevenson FT, Nazian SJ. Group testing in medical education: An assessment of group dynamics, student acceptance, and effect on student performance. *Med Sci Educ.* 2013;23:346–354.
- 3 Meseke CA, Nafziger R, Meseke JK. Student attitudes, satisfaction, and learning in a collaborative testing environment. *J Chiropr Educ.* 2010;24:19–29.
- 4 Meseke JK, Nafziger R, Meseke CA. Facilitating the learning process: A pilot study of collaborative testing vs individualistic testing in the chiropractic college setting. *J Manipulative Physiol Ther.* 2008;31:308–312.
- 5 Parmelee D, Michaelsen LK, Cook S, Hudes PD. Team-based learning: A practical guide: AMEE guide no. 65. *Med Teach.* 2012;34:e275–e287.
- 6 Rao SP, Collins HL, DiCarlo SE. Collaborative testing enhances student learning. *Adv Physiol Educ.* 2002;26:37–41.
- 7 Wieman CE, Rieger GW, Heiner CE. Physics exams that promote collaborative learning. *Physics Teacher.* 2014;52:51–53.
- 8 Meseke CA, Bovée ML, Gran DF. Impact of collaborative testing on student performance and satisfaction in a chiropractic science course. *J Manipulative Physiol Ther.* 2009;32:309–314.
- 9 Cortright RN, Collins HL, Rodenbaugh DW, DiCarlo SE. Student retention of course content is improved by collaborative-group testing. *Adv Physiol Educ.* 2003;27:102–108.
- 10 Vygotskii LS, Cole M. *Mind in Society: The Development of Higher Psychological Processes.* Cambridge, Mass: Harvard University Press; 1978.