MULTIPLE CHOICE VERSUS CONSTRUCTED RESPONSE:
A SMALL-SCALE CLASSROOM ASSESSMENT EXPERIMENT

K. Shane Goodwin—Department of Mathematics

Over the last ten years of teaching in higher education, I have struggled with the assessment dilemma between constructed-response (CR) and multiple-choice (MC) summative examination tools. Each has its pedagogical strengths and weaknesses. Through enlightening discussions with my international students, I often find that they are surprised by the prevalence of MC examinations here in the U.S. This seems to be especially true for students coming from European schools (although my sample size of opinion is fairly small). These students mention that they are much more accustomed to taking write-on exams at both the secondary and higher-education levels.

I have questioned at times whether there is a relationship between our culture's fascination with games of chance and the MC format.

I have questioned at times whether there is a relationship between our culture's fascination with games of chance and the MC format. At the end of a recent semester, an anxious student dropped by my office after taking one of my multiple-choice statistics exams. The exams had neither been scored nor handed back, but he knew there was one question in particular that had stumped him. Disappointed that he had not properly prepared himself in that specific area, he then asked if I could show him how to work the problem. His eyes lit up when he realized that, although he did not know how to do the problem, he received full credit for having guessed correctly. It is this element of chance that adds an interesting dimension to our testing methods. I thought at the time how curious it is that students can get full credit when they may be ignorant of the topic in question. That possibility, nonetheless, seems to provide hope to the under-prepared and can even give an advantage to the clever student who can "reverse-engineer" the problem by eliminating the distracters one by one. This seems to be especially true of a MC exam in mathematics.

To analyze the differences between MC and CR assessments more carefully, I conducted an experiment with my college algebra class (Math 110) during the First Summer term of 2003. What follows is a brief summary of my findings combined with a review of some of the research literature pertaining to these two forms of summative assessment.

The advantages of multiple choice testing are numerous. They include minimal grading, timely feedback to students, relative freedom from scoring biases, a wider sampling of course content, lower potential for measurement error, and high correlation to constructed-response test scores (Walstad, 1998). Some researchers argue, however, that despite the long list of advantages, MC testing does not adequately assess overall student performance and that some students are "predisposed to do

better on multiple choice or problems depending on the characteristics they possess" (Krieg & Uyar, 2001).

Gronlund (1976) suggests that since the MC format requires the selection of the correct answer, it isn't suited to measuring problem-solving skills in mathematics or science. Nor does it effectively measure the student's ability to organize and present ideas in a coherent and well-structured fashion. Because of these limitations many teachers and researchers in the field, of mathematics and statistics use a variety of assessment instruments: written and oral reports, projects, portfolios, essay questions, and so on (Garfield, 2000).

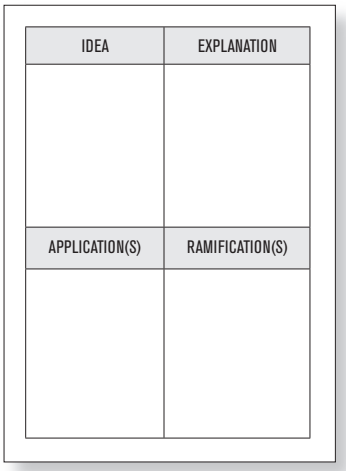
With this background in mind, I offered my summer college algebra students an alternative to the typical MC format that I have used in the past. By the middle of the term, my students had already taken two multiple-choice exams and were about to take a third exam covering polynomial, exponential, and logarithmic modeling. I developed a take-home/open-book summative examination that had a unique twist described on my exam cover sheet:

I offered my summer college algebra students an alternative to the typical MC format that I have used in the past.

> The following exam is meant to be a customizable assessment for you as the learner. You will be able to choose the modeling topics that interest you and allow you the best opportunity to demonstrate mastery. In other words, rather than finding out what you *do not know*, this particular exam will be assessing what you *do know*.

I offered the students a choice of four different mathematical topics for each of the first four open-ended questions. I then asked them to follow a particular summative format that included the following commentary and full-page quadrant structure:

> Each problem will be worth 12 points and will be graded using the rubric scale from the previous page. The three criteria are Organization, Mathematical Analysis, and Understanding of Mathematics.



- IDEA—refers to the main concept or big picture summary in your own words.
- EXPLANATION—refers to the comments and/or sketches necessary to concisely explain the idea.
- APPLICATION—refers to use of a real-world example of how the idea is put into action or how the idea can be utilized in a higher level math problem.
- RAMIFICATION—refers to the potential connectedness to other mathematical ideas, or the potential limitations that it might introduce, or finally the overall rationale for learning it.

Borrowing from rubric suggestions of Huba and Freed (2000), I struggled to produce the right wording and point values that would make sense to my college algebra students. The table below appeared on their cover sheet as well.

Table 1

Levels of Achievement				
Criteria	Exemplary 4 points	Proficient 3 points	Needs Improvement 2 points	Poor 1 point
Organization	Written work is well organized and easy to follow.	The organization is generally good, but some parts may not seem to fit well.	The organization of the work is unclear or cloudy.	The work is unorganized to the point of confusion.
Mathematical Analysis	Analysis is detailed, algebraically accurate, and complete.	The analysis is detailed, but some algebraic steps may be missing or not valid.	Some analysis is included and is correct, but many steps are missing.	Correct analysis is rarely used, and steps are not supported.
Understanding of Mathematics	The reader can clearly tell that the writer understands the underlying mathematics.	The student's work suggests some understanding, but the reader is left wishing for more to enrich the problem.	The work refers to correct mathematical principles, but relationships are weak or unclear.	No obvious references to mathematical principles exist suggesting a lack of understanding.

Beyond the open-ended questions, students had an opportunity to construct their own multiple-choice problems.

Beyond the open-ended questions, students had an opportunity to construct their own multiple-choice problems. This included providing a correct answer and its solution steps for each question along with three or four distracters. Finally, the examination concluded with a choice between two "mini-projects" that involved ill-defined problems from our current unit of study.

The students had one week to complete the exam and were encouraged to use textual sources but not to help one another. If anything, the exam became an individual project and was in no way a typical multiple-choice test. Upon turning in the exams, five students were randomly selected from the 43 in the class to form a focus group. We met for fifteen minutes before the end of class (after I had dismissed the rest of the students early) and I asked the focus group the following three questions:

What do you feel are the strengths of this type of assessment?

Gary: You learn a lot about a specific topic (overall techniques or concept). Good to see how multiple choice distracters are put together.

Mike: Not very positive. Didn't think the amount of time it took was worth it.

Jessica: You could find parts you were weak in and learn about them. Explore both your weaknesses and strengths. You could explain things in your own words, so you would understand it better.

Michelle: Liked how you could choose the concepts you felt you were stronger in. Learned a lot by going back into the book and digging in.

Sarah: All of the above. Learned more than in class because I had to reread concepts.

What do you feel are the weaknesses of this type of assessment?

Gary: Nice to be able to choose, but it gave too much choice and hard to pick which one I knew the best. Spent time changing mind. Hard to know what the teacher really wanted.

Mike: Took way too long. Would have learned more this way, but it wasn't worth the time invested. Would rather just go and take a quicker MC test.

Jessica: Not being able to figure out which concepts to focus on. I wanted to explore other concepts and not so much depth on just one concept.

Michelle: Wasn't sure exactly how you wanted the format. Example helped a bit but I focused only on the rational root theorem, and other concepts were different so didn't know how to apply the example.

Sarah: Difficult to read and then put in my own words. The example was a bit confusing.

What would you do differently if you were writing the exam?

Gary: Narrow the choices down. Less stressful if we had fewer to do.

Mike: Same thing. Fewer choices. One per question. Spent a lot of time deciding which topics.

Jessica: Have a few more examples to follow.

Michelle: Clearer examples.

Sarah: Same as above. Concentrate on more specific topics.

The overall message of the focus group was that the exam was a learning experience but that a few things should be modified to make the benefits worth the extra time demanded. Considering the comments of these students, I reasoned that many of them had not experienced a take-home summative exam where significant time is required to research the questions and provide substantive answers.

I was surprised that some in the focus group didn't appreciate the number of topic choices because of the time spent deciding on which to summarize. They seem to have an innate desire for economy when test taking. This is evident in the survey I conducted upon handing back the graded exams. When asked what was the number one advantage of taking a standard MC exam, more than half of my college algebra students responded that it is faster than a write-on exam. Likewise, when asked to give the primary disadvantage of taking a write-on exam, 25 of 43 students chose the response that it is more time consuming than a multiple choice exam. On the other hand, when asked to give the main advantage of the write-on exam, 24 students chose availability of partial credit.

They seem to have an innate desire for economy when test taking.

With regard to the overall learning experience, on a 1-to-5 Likert scale with 5 being the highest learning experience, my students averaged 3.8. Table 2 summarizes the data.

Please note that only 41 of the 43 students properly completed this portion of the survey.

Table 2

How would you rate this exam as a learning experience (low = 1 and 5 = high)?

Likert Rating	1	2	3	4	5
Students	0	3	9	22	7

As to whether the rubric was helpful to the students as they took the exam, my students averaged 3.4. Table 3 summarizes the data.

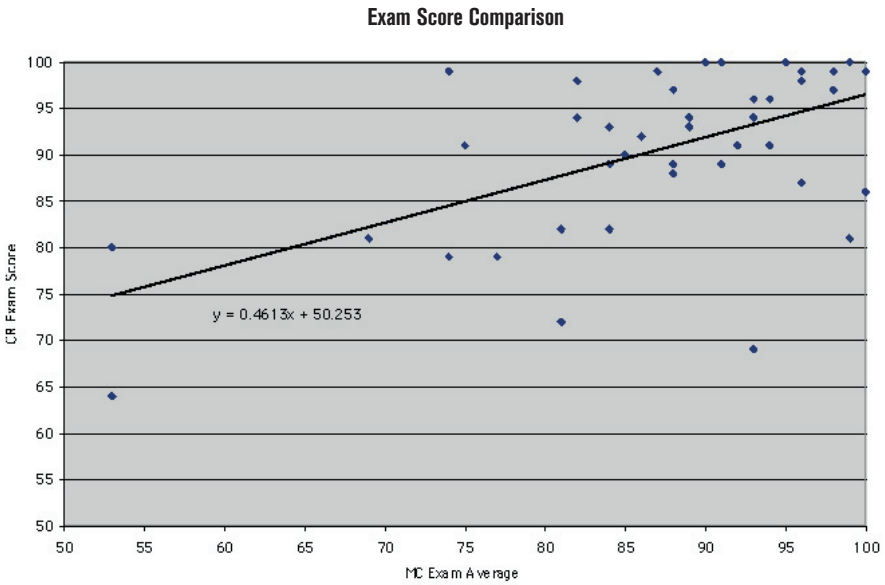
Table 3

How would you rate the helpfulness of the rubric (low = 1 and 5 = high)?

Likert Rating	1	2	3	4	5
Students	3	4	14	13	7

In addition to the class survey, I contrasted student test averages on the new CR test to see what differences might exist.

Figure 1



As one would expect, the better my students performed on the MC exams during the first part of the term, the better they performed on the experimental CR exam. Using the sample correlation coefficient of r = 0.56 and the sample size of 43, yields a p-value < 0.0001, suggesting there is a positive linear correlation between the MC and CR scores. To analyze this from another perspective, I performed a paired t-Test on the two samples of scores summarized in Table 4.

As one would expect, the better my students performed on the MC exams during the first part of the term, the better they performed on the experimental CR exam.

Table 4

t-Test: Paired Two Sample for Means

	Original Test Avg	Write-on Exam
Mean	87.41860465	90.58139535
Variance	121.7253599	81.63012182
Observations	43	43
Pearson Correlation	0.56334865	
Hypothesized Mean Difference	0	
df	42	
t Stat	-2.173593405	
P(T<=t) one-tail	0.017711568	
t Critical one-tail	1.681951289	
P(T<=t) two-tail	0.035423136	
t Critical two-tail	2.018082341	

At a level of significance of alpha = 0.05, there is a difference in the means of the two exam methods but not at alpha = 0.01. I am hesitant

to conclude that there is a significant difference in the means, primarily because I often strayed from the rubric while I was grading and as a result I became more and more lenient in awarding partial credit points. This is one of the drawbacks to constructed-response testing that Krantz (1999) asserts in his research and commentary. Unlike the objectivity in scoring a multiple-choice exam, there can be great subjectivity at times when professors grade the open-ended write-on style exams. Yet, Krantz also suggests that when class sizes are reasonable, perhaps a few constructed-response questions can be attached to an objective multiple-choice test to lend a personal touch. He contends that students can benefit from even brief comments written by the professor.

I am persuaded that I should aim for the middle of the road of assessment.

This brings me to the assessment debate between MC and CR examinations. I am persuaded that I should aim for the middle of the road of assessment. Because of the sheer number of students and the high overhead in hand grading that results, MC testing will likely remain an important tool in my "assessment toolkit." And yet, until I discover better what is going on in the mind of the learner by way of open-ended questions, I won't have a clear picture of the student's understanding nor how much he or she can articulate.

For this reason, I hope to adopt one of two assessment strategies. Either I will couple a student or group project with the traditional MC examinations or I will bring into each examination two or three constructed-response test items. Through either strategy, I hope to maximize academic feedback while minimizing grading time.

Works Cited

Garfield, Joan B. "Beyond Testing and Grading: New Ways to Use Assessment to Improve Student Learning," in Moore, Thomas L., ed., Teaching Statistics: *Resources for Undergraduate Instructors* (Washington, D.C.: Mathematical Association of America and American Statistical Association, 2000), 201-208.

Gronlund, Norman E. *Measurement & Evaluation in Teaching*, 3rd ed. (New York: Macmillan Publishing, 1976).

Huba, Mary E. & Freed, Jann E. *Learner-Centered Assessment on College Campuses: Shifting the Focus from Teaching to Learning*, 1st ed. (Boston: Allyn and Bacon, 2000).

Krantz, Steven G. *How to Teach Mathematics,* 2nd ed. (Rhode Island: American Mathematical Society, 1991).

Kried, Randall G. & Uyar, Bulent, "Student Performance in Business and Economics Statistics: Does Exam Structure Matter?" *Journal of Economics and Finance*, 25, (2001), 229-235.