TO USE CONSTRUCTED-RESPONSE ASSESSMENTS, OR NOT TO USE CONSTRUCTED-RESPONSE ASSESSMENTS? THAT IS THE QUESTION

by

Stephen Hickson, W. Robert Reed, and Nicholas Sander*

Abstract

This study examines one of the hypothesized benefits of using constructed response (CR) questions. Namely, that CR questions test a higher-level of understanding than multiple choice (MC) questions. We argue that if this benefit is to justify the higher costs of CR questions, then one important result is that grade outcomes from CR questions should be substantially different than those from MC questions. We use a data set composed of thousands of observations on individual students in introductory economics classes at a large public university. We note that the instructors of these classes made conscientious efforts to write CR questions that assessed higher levels of learning (Bloom, 1956). Despite this, we find relatively little difference in grade outcomes. Our analysis suggests that switching from an all-CR assessment to an all-MC assessment would produce grade variations that are similar to the differences that are observed for students across different tests. However, we also note that there are potential issues in moving to an all-MC format relating to student study habits and their perception of fairness of the assessment. While other studies have focused on test scores, frequently AP test scores, our study is the first to focus attention on grades. We hope that our inability to identify substantial benefits to CR questions will stimulate further research to identify substantive benefits from using the more costly CR questions.

JEL Categories: A22

Keywords: Principles of Economics Assessment, Multiple Choice, Constructed Response, Free Response, Essay.

October 1, 2010

*Hickson, Reed, and Sander are, respectively, Teaching Fellow, Professor, and Honours student at the University of Canterbury. Sander was supported by a Summer Scholarship 2009-2010 jointly funded by the University of Canterbury and the Tertiary Education Commission of New Zealand (TEC). Reed is the contact author and his contact details are: Department of Economics and Finance, University of Canterbury, Private Bag 4800, Christchurch 8042, New Zealand; Email: bobreednz@yahoo.com; Phone: +64 3 364 2846.

> "A main message of this article is that the decision to abandon the constructed-response portion of tests in favour of an all-multiple-choice format should rest on the analysis and considered judgment of the costs and benefits of such a decision."
>
> - Kennedy and Walstad (1997)

I. INTRODUCTION

Most principles classes in economics employ assessments that use both multiple-choice (MC) and constructed-response (CR) questions. CR questions are costly to evaluate, and subject to greater subjectivity in marking. However, many instructors believe that (i) CR questions measure a different type of knowledge/learning than MC questions, and (ii) that this benefit is sufficiently large to outweigh their higher costs.

The research literature is divided with respect to the first of these points. Many studies conclude that MC and CR questions measure the same thing (Bennet, Rock, and Wang, 1991; Lukhele, Thissen, and Wainer, 1994; Thissen, Wainer, and Wang, 1994; Wainer and Thissen, 1993; and Walstad and Becker, 1994). Other studies find evidence that MC and CR measure different things (Krieg and Uyar, 2001; Lumsden and Scott, 1987; Becker and Johnston, 1999; and Hickson and Reed, 2009).

Even if CR questions measure a different type of knowledge/learning, there remains the question of whether this additional information makes much of a difference in terms of assessment outcomes or grades. If CR questions produce the same grades as MC questions – or near to the same grades – then there may be little benefit to using CR questions.

This study investigates the degree to which grades based solely on MC questions differ from grades based solely on CR questions. We show that this difference is due to two factors: a "systematic difference" and sampling error. Further, we show why it is impossible to identify the "systematic" component from observed MC-CR grade differences unless one is able/willing to make restrictive assumptions. Absent that, the best one can do is benchmark the size of the grade differences between MC and CR assessments against other kinds of observed grade differences.

Our data consist of thousands of observations on individual students who took principles of economics classes at a large public university. As we detail below, the instructors of these classes made a conscientious effort to write CR questions so that they would assess higher-level learning missed by MC questions (Bloom, 1956). This provided an exceptional opportunity to evaluate whether CR questions produce different assessment outcomes than MC questions.

Our analysis finds that differences in assessment outcomes between CR and MC components on the same test are comparable to differences in the same components across different tests. Specifically, the grade differences between CR and MC questions on a given assessment are similar in distribution to (i) differences in CR grades across different assessments, and (ii) differences in MC grades across different assessments. This suggests that, holding other things constant, switching from an all-CR assessment to an all-MC assessment would have produced grade variations similar to those already observed across different tests for the students in these courses.

While other studies have focused on test scores, frequently AP test scores, our study is the first to focus attention on grades. Further, the data for our study is collected directly from university introductory economics classes. This should make our findings of particular interest to instructors of large, university classes who are considering switching to all-MC assessments or indeed all-CR assessments.

The paper proceeds as follows. Section II describes our data. Section III provides the theoretical framework for understanding the empirical analysis. Section IV presents our results. Section V describes the type of MC and CR questions underlying our data, and discusses why we believe our results may be applicable to other settings at other universities. Section VI concludes.

II. DATA

Data. Our data consist of 7754 assessment records for students who took introductory microeconomic and/or introductory macroeconomics at the University of Canterbury (UC) from 2002-2007. For each student in each class we have a record of both their (i) Term Test and (ii) Final Exam results, including their performance on the MC and CR components of each assessment.[1] Term Tests for both micro- and macroeconomics classes consisted of a MC component consisting of 25 questions worth one point each, and a CR component consisting of 2 questions worth a total of 50 points. Final Exams consisted of 30 MC questions worth one point each, and 3 CR questions worth a total of 70 points. The same two instructors taught both courses over this time period.

We converted scores for each of the four components – MC-Term Test, MC-Final Exam, CR-Term Test, CR-Final Exam – to a 100-point scale. We then converted these numerical scores to grades using the UC grading scheme: A+ is assigned to scores 85 and over; A to scores between 80 and 85, A- to scores between 75 and 80, … , C- to scores between 45 and 50, D to scores between 40 and 45, and E to all scores less than 40. The respective grade distributions for each of the four assessment components are reported in TABLE 1, as are some other assessment characteristics.

The bottom two rows in the table report the mean and standard deviation of raw assessment scores (adjusted to a 100-point scale) for each of the assessment components. The first four columns report the distribution of grades that would be given if each of the

[1] We excluded students who did not complete both a term test and a final exam. Likewise, we excluded all students who were awarded an "aegrotat" as this flagged assessment results that were deemed unrepresentative due to circumstances which substantially impaired a student's performance.

components constituted a separate assessment. The last column reports the historical distribution of final grades given to the students in our sample.

The first thing we note is that students in our sample typically scored higher marks on the MC components of the assessments in our sample. The mean scores of the MC components were 66.5 and 70.3 on the Term Tests and Final Exams, respectively. The corresponding means for the CR components were 49.5 and 53.4. The higher mean scores for the MC components translate to higher grades. If the MC components were stand-alone assessments, students would receive more As and Bs, and fewer Cs, Ds, and Es. In other words, if the grading schedules were kept the same and everything else remained constant, students would see their grades substantially increase if assessments switched to all-MC questions. This highlights a major problem with measuring the impact of a greater reliance on MC questions.

If MC questions tend to produce, say, higher scores than CR questions, we suspect that instructors would either write harder MC questions, or scale the marking scheme to make it more difficult to achieve higher grades. We incorporate this response by assuming that any change in the question makeup of assessments will maintain the overall grade distribution at the historical pattern. This requires that we adopt a pre-determined grade distribution to standardize the grade outcomes from MC and CR questions. We proceed by using the historical grade distribution for the students in our sample.

The last column of TABLE 1 reports the grade distribution associated with all final grades for all student observations in our sample. We use this historical distribution to "norm" individual assessment components by imposing this grade distribution on each of the respective components. For example, for a given class with say, 400 students, we have the following four assessment observations for each student: a MC(Term Test), a MC(Final Exam), a CR(Term Test), and a CR(Final Exam). For each assessment component we assign

A+s to the top 3.4 percent of students, A's to the next 4.7 percent of students, and so on. In this way, the same grade distribution is imposed on each assessment component for each class in our sample. Thus, when we take the difference between the MC and CR components of an assessment, we hold constant the overall grade distribution.

TABLE 2 maps students' MC grades conditional on their CR grades on the same assessment. The top panel reports grade distributions for Term Tests. The bottom panel does the same for Final Exams. The table is interpreted thusly: 28.5 percent of students who received an A+ on the CR component of Term Tests also received an A+ on the MC component (once we rescaled CR and MC scores to have the same grade distributions). 17.2 percent received an A on the MC component. And so on.

Along any given row in the table, the sum of percentages excluding the diagonal element represents the percent of students who would experience a grade change if their assessment grade was based entirely on their MC component rather than their CR component. Thus, 72.5 percent of students (=100 – 28.5) who received an A+ on the CR component of the Term Test would receive a different grade (in this case, a lower grade) if their grade was based on the MC component. The numbers in the table can also be used to calculate the percentage of students who experienced a change of one full grade (e.g., B+ to C+, B- to A-) or more.

TABLE 2 is insightful for assessing how the rank order of students is changed when grades are based on MC questions rather than CR questions. Small changes in rank order are likely to show up as no change in grade. In contrast, rank orders must be substantially altered in order for a change in question format to produce large grade changes. The next section provides a theoretical framework for analyzing observed grade differences.

III. THEORY

Grade differences between MC and CR components on the same assessment. Let an individual observation of student *i*'s MC component on a given assessment *j* (*j* = Term Test, Final Exam) be given by

(1) $y_{ij,MC} = \mu_{ij,MC} + \varepsilon_{ij,MC}$,

where $y_{ij,MC}$ is the observed grade measured in points (cf. TABLE 1), $\mu_{ij,MC}$ is the mean grade the *ith* student would achieve on the *jth* MC assessment, and $\varepsilon_{ij,MC}$ is the "sampling error" associated with the fact that the same student would score different marks on different offerings of a given MC assessment. "Sampling error" may be due to student-specific factors, such as how hard the student studied for a particular assessment, how mentally alert and focused the student was on a given day, etc. Or it may be due to instructor-specific factors such as the general difficulty of the questions posed on the given assessment, or the specific areas that the questions tested. "Sampling error" also includes the impact of guessing.

Note that the systematic component, $\mu_{ij,MC}$, may incorporate factors that are not directly related to the understanding of course material. For example, some students may generally experience less test anxiety than others. This may result in systematically higher grades, even though the student may not have a superior understanding of the material.

Similarly, let

(2) $y_{ij,CR} = \mu_{ij,CR} + \varepsilon_{ij,CR}$

represent individual *i*'s CR grade, where $\mu_{ij,CR}$ and $\varepsilon_{ij,CR}$ are defined as above, except that $\varepsilon_{ij,CR}$ also includes "marking error" associated with subjective evaluations by potentially different markers subjectively evaluating the student's CR answers.

It follows that the observed difference between a student's MC and CR grades on a given assessment equals

(3) $$y_{ij,MC} - y_{ij,CR} = \left(\mu_{ij,MC} - \mu_{ij,CR}\right) + \left(\varepsilon_{ij,MC} - \varepsilon_{ij,CR}\right).$$

This difference is composed of two components: (i) a "systematic" difference, $\left(\mu_{ij,MC} - \mu_{ij,CR}\right)$, which represents the expected value of the difference in scores from a student repeatedly taking CR and MC assessments covering the same material; and (ii) "sampling error," $\left(\varepsilon_{ij,MC} - \varepsilon_{ij,CR}\right)$, associated with a variety of student- and instructor-specific factors.

If CR questions are able to assess higher-level learning in a way that MC questions are not, this should be reflected in the systematic component, $\left(\mu_{ij,MC} - \mu_{ij,CR}\right)$. If the systematic component shows little variation, then it would call into question the benefit of using CR questions. Unfortunately, this systematic component may incorporate the influence of other factors not directly related to understanding. For example, non-native English speakers may have a more difficult time expressing their thoughts on CR questions. Markers may not be able to distinguish (nor care to distinguish) that an unsuccessful answer is due to poor English facility rather than poor comprehension of course material.

Accordingly, we wish to identify the systematic component in observed MC-CR grade differences. FIGURE 1 illustrates the problem. The first two panels show the distributions of MC and CR grades for a given assessment, measured in points.[2] Both distributions are normed to have the same overall grade distribution. For each student and each assessment, we match their grades from the MC and CR components of that assessment and take the difference. Consider the following cases:

[2] For the relationship between grades and points, see TABLE 1.

Case One. Case One represents no sampling differences and no systematic differences, $\left(\varepsilon_{ij,MC} - \varepsilon_{ij,CR}\right) = \left(\mu_{ij,MC} - \mu_{ij,CR}\right) = 0$, for all *i,j*. In this case, all students would have exactly the same grades from the MC and CR components.[3] Panel (C) of FIGURE 1 plots the distribution of MC-CR grade differences for this case. Note that the values on the horizontal scale run from -10 to +10. “-10” represents the grade difference associated with a student receiving an A+ and an E on the CR and MC components of an assessment, respectively. “+10” represents the grade difference from receiving an E on the CR component and an A+ on the MC component of that assessment. In this case, all students receive the same grade on the MC and CR components, so that the distribution of $y_{ij,MC} - y_{ij,CR}$ is represented by a mass point at 0. The rank order of students is the same for both MC and CR.

Case Two. Case Two occurs when there is no sampling error but systematic differences exist between students’ performances on the MC and CR components of assessments: $\left(\varepsilon_{ij,MC} - \varepsilon_{ij,CR}\right) = 0,\ \left(\mu_{ij,MC} - \mu_{ij,CR}\right) \neq 0$ for all *i,j*. Case Two will produce distributions of grade changes like those in Panel (D) of FIGURE 1. In this case, the distribution of $y_{ij,MC} - y_{ij,CR}$ contains useful information because it measures the systematic effects of switching from an all-CR to an all-MC question format.

The near symmetry of the grade difference distribution in Panel (D) is no accident. Given the fact that the MC and CR grade distributions are normed to be the same, the total number of grade point increases must equal the total number of grade point decreases. For example, if one student experiences a grade increase from an B- to a B+, either (i) another student must experience a grade decrease from a B+ to a B-, or (ii) two students must experience a grade change where one decreases from B+ to B, and another decreases from B

[3] In terms of TABLE 2, all the diagonal terms would equal 100 and the off-diagonal terms would be zero.

to B-, or (iii) a similar chain of grade changes must take place to compensate the original change.

Case Three. Case Three represents the scenario where there are no systematic differences between students' performances on the MC and CR components of assessments, but sampling errors cause the observed grades to be different, $\left(\mu_{ij,MC} - \mu_{ij,CR}\right) = 0$, $\left(\varepsilon_{ij,MC} - \varepsilon_{ij,CR}\right) \neq 0$ for all *i,j*. This will also produce distributions of grade differences like those reported in Panel (D) of FIGURE 1. However, in this case, the difference distribution tells us nothing about the systematic effects of switching from CR to MC assessments. Unfortunately, there is no way to distinguish Case Two from Case Three as they are observationally identical.[4]

Case Four. In Case Four, both systematic differences and sampling errors are jointly present, $\left(\varepsilon_{ij,MC} - \varepsilon_{ij,CR}\right) \neq 0$, $\left(\mu_{ij,MC} - \mu_{ij,CR}\right) \neq 0$ for all *i,j*. This will produce distributions like the two previous cases. While this is the case that most certainly represents reality, it is observationally equivalent to the two previous cases. Without imposing an assumption on their relative sizes, it is impossible to identify the systematic component of the MC-CR grade differences.[5]

In fact, the problem of identifying systematic differences from observed differences is even more vexing than the previous discussion acknowledges. Note that the size of the systematic component, $\left(\mu_{ij,MC} - \mu_{ij,CR}\right)$, in the grade difference distribution of Panel (D) is not independent of the behavior of the error terms. As the variance of the sampling errors ε

[4] Note that the norming of the MC and CR distributions throws away information about the relative sizes of the sampling errors associated with the original test scores. This is a direct consequence of assuming that instructors would use similar/identical grade distributions irrespective of whether the assessments were MC or CR.

[5] Kennedy and Walstad (1997) get around this problem because they (i) use nominal scores, rather than a normed distribution; and (ii) assume the size of the variance of the error terms in their sample of AP scores based on external analyses, i.e., they do not estimate it from their sample.

increases, so must the variance of the corresponding grade point distributions *y*. These need to get re-normed in order to maintain the overall grade point distribution. This effectively reduces the variance of the systematic component μ, and thus, similarly, the variance of $(\mu_{ij,MC} - \mu_{ij,CR})$.

The preceding discussion illuminates the problems of identifying the effects of switching to an all-MC assessment format based on observed differences in MC and CR scores. The remainder of this section discusses how observed MC and CR differences can be analyzed to achieve a more modest goal that may still provide useful information regarding the benefits of CR questions.

<u>Grade differences between similar components across different assessments</u>. Let us now consider differences in the MC grades between the Term Test (*j*) and Final Exam (*k*) for a given student *i*. Using the same notation as above, the observed difference is given by

$$(4) \qquad y_{ij,MC} - y_{ik,MC} = \left(\mu_{ij,MC} - \mu_{ik,MC}\right) + \left(\varepsilon_{ij,MC} - \varepsilon_{ik,MC}\right).$$

A similar relationship holds for the differences in CR grades across different assessments *j* and *k*.

$$(5) \qquad y_{ij,CR} - y_{ik,CR} = \left(\mu_{ij,CR} - \mu_{ik,CR}\right) + \left(\varepsilon_{ij,CR} - \varepsilon_{ik,CR}\right).$$

Once again the observed difference is composed of two components: (i) a "systematic" difference, and (ii) "sampling error."

Note, however, that the two components differ in important ways from the two components in Equation (3). While MC and CR questions on the same test may assess different parts of course material, the overlap in course material is likely to be greater than the overlap on the Term Test and Final Exam. Further, the sampling error component now includes differences across different points in time that are not related to understanding. For example, a student may feel fine for the Term Test but feel sick for the Final Exam. Nevertheless, all of the problems associated with inferring the sizes of these two components from observed MC-CR differences hold *a fortiori* when analyzing MC-MC and CR-CR differences across assessments.

The value of cross-assessment differences in MC and CR grades. Given the above it is clear that one cannot use MC-MC or CR-CR differences across assessments to identify the systematic component of the distribution of MC-CR grade differences on the same assessment. Nevertheless, differences across assessments can still provide useful information. Among other things, they allow one to benchmark the MC-CR grade differences to determine whether they are larger, smaller, or approximately the same size as grade differences across tests.

Students are used to the existence of both a term test and a final exam changing the rank order of students and hence grades. In a sense, the existence of two assessment items introduces some “noise” (be it systematic or random) into the rank order. A shift to one form of assessment (e.g. all MC questions) also introduces “noise” in that the rank order is changed compared to the current mixed format. Suppose the change in the rank order of students that occurs from moving from an all-CR to an all-MC format is substantially larger than that which exists between the Term Test and Final Exam. That suggests that switching to an all-MC test could result in a world where the change in rank order and hence grades is much larger than that currently experienced by students. Alternatively, suppose the change is the same or smaller compared to grade differences between the Term Test and Final Exam. That would provide evidence that the magnitude of the change in rank order would not be dramatically different compared to what students’ already experience – at least in the sense that any grade differences that resulted would be of the same magnitude as those that currently occur across assessments.

IV. RESULTS

The first two rows of TABLE 3 summarize the grade differences between the MC and CR components on Term Tests and Final Exams, respectively. We categorize the differences in terms of whether there is (i) a one-letter grade negative difference or more between the CR and MC components, (ii) a difference of less than one-letter grade, and (iii) a one-letter grade positive difference or more. An example of a negative, one-letter grade difference would be a student who received a grade of A- on the CR component while achieving a B- on the MC component. I.e., their MC grade was a letter grade lower than their CR grade.

For the Term Test observations in our sample, roughly one-third of students had a MC-CR grade difference of a letter grade or more (32.9% = 16.2% + 16.7%). For the Final Exam observations, the corresponding number is a little less than a fourth (22.7% = 11.4% + 11.3%). As discussed above, it is unclear how to map this empirical result to the decision to move to all-MC assessments. If they are simply measuring sampling error, then they have nothing to say about the systematic changes in switching assessment formats. On the other hand, if all of the observed differences measure systematic differences, then they could indicate that a substantial proportion of students would be affected by a move to all-MC questions.

The third and fourth rows of TABLE 3 summarize the differences between MC grades on the Term Test and Final Exam, and CR grades on the Term Test and Final Exam, respectively. It is worth repeating from above that while these differences also have a systematic and a sampling error component, these differ in important ways from the MC-CR differences. While the cross-assessment differences of rows 3 and 4 do not neatly map onto the MC-CR differences of rows 1 and 2, they do provide useful information. Specifically, they provide a benchmark for gauging the size of the grade differences that might arise from switching to all-MC assessments.

TABLE 3 indicates that the associated distribution of MC-CR grade differences (i.e. the change in rank ordering) is approximately equal to the grade difference (rank ordering) that occurs across assessments. The first two rows of TABLE 3 indicate that between a fourth and a third of students would experience a change of a full letter grade or more if the assessment were based on the MC component rather than the CR component. This compares with approximately 33.3% of students who experience a change of a letter grade or more in the CR component from the Term Test to the Final Exam; and approximately 25.1% who see their grade change by a letter grade or more on the MC component between these two assessments. In other words, switching to all-MC assessments, holding others things constant, would not introduce any greater change in the rank order than students currently experience. The associated changes would be roughly the same magnitude as the changes students are already accustomed to experiencing across assessments. It is certainly correct that many individual students would receive a different grade with an all-MC assessment but they also receive a different grade with a term test and a final exam compared to, say, a final exam only. The extent of this difference is about the same in both cases.

Recall that the literature is somewhat divided on whether or not MC and CR measure different aspects. However, previous work by these authors (2009) suggests that MC and CR questions do measure different levels of learning. What this study indicates is that the measurement of the different levels of learning is highly enough correlated that changes in rank order are not, from the students' perspective, unreasonable.

Does this mean that CR questions can be abandoned with no impact? Not necessarily. One must remember that CR questions do serve to spread the students out along the rank order somewhat. The data in table 1 shows not only that the mean for MC questions is higher but also that students are far more bunched around this mean than for CR given the lower variance in MC. This feature of CR may well be important for points in the rank order where students tend to cluster and distinguishing between individual students is required.

Students also perceive a qualitative difference between the change in rank order introduced by having a test and an exam vs. the change in rank order that would be introduced in moving to an all-MC format. The authors surveyed the 2010 Principles of Macroeconomics class about their perceptions of assessment and the results are presented in Table 4. In this survey 93 percent of students prefer to have both a term test and an exam while 87 percent believe that both a test and an exam is “fairest”. Approximately 30 percent of students expressed a preference for either all-MC or all-CR in terms of the best chance of their highest rank in class. However, of that group 73 percent said that a mixed format (both MC and CR questions) is “fairest for students” despite they themselves having expressed a preference for one or the other. Only three percent of students believe that an all-MC test is “fairest”. What we conclude from this is that students would not see the change in rank order that would result from a move to an all-MC format as being “fair”. Hence they are unlikely to perceive the change in rank order of the move to an all-MC as being “fair” compared to that which occurs with a term test and a final exam despite both being of similar magnitude.

It is also possible that the actual presence of CR questions may drive different study habits. If CR questions do measure different levels of learning then one might expect students to study harder in order to do well on those questions. In our survey 35 percent of students declared that they would study harder for a test that contained all CR questions. We note that this is a stated rather than a revealed preference but it is consistent with what one could expect.

By separating the difference between MC and CR into systematic error, we are able to think about the importance of each of these in relation to the use of MC and CR questions. Consider again Equations (4) and (5).

(4) $$y_{ij,MC} - y_{ik,MC} = \left(\mu_{ij,MC} - \mu_{ik,MC}\right) + \left(\varepsilon_{ij,MC} - \varepsilon_{ik,MC}\right).$$

(5) $$y_{ij,CR} - y_{ik,CR} = \left(\mu_{ij,CR} - \mu_{ik,CR}\right) + \left(\varepsilon_{ij,CR} - \varepsilon_{ik,CR}\right).$$

Suppose that (i) the difference distributions, ($y_{ij,MC} - y_{ik,MC}$) and ($y_{ij,CR} - y_{ik,CR}$), had equal dispersion; i.e., var($y_{ij,MC} - y_{ik,MC}$) = var($y_{ij,CR} - y_{ik,CR}$). Suppose further that one was willing to assume that (ii) the sampling error inherent in MC assessment was less than that for CR assessment; so that, var($\varepsilon_{ij,MC} - \varepsilon_{ik,MC}$) < var($\varepsilon_{ij,CR} - \varepsilon_{ik,CR}$) . Assuming that (iii) the covariance between the systematic and sampling error components was zero, this would imply that var($\mu_{ij,MC} - \mu_{ik,MC}$) > var($\mu_{ij,CR} - \mu_{ik,CR}$). Finally, let us suppose that (iv) these systematic differences were primarily related to differences in learning outcomes.

If the data showed (i), and one was willing to make assumptions (ii)-(iv), then one could conclude that MC assessments had greater discriminatory power than CR assessments for measuring learning outcomes. This follows directly from the assumption that CR assessments have more sampling error. Holding constant the overall grade distribution, a larger sampling error component implies a smaller systematic component. If the variance of MC-MC differences is greater than CR-CR differences, as it is in our data (compare Rows 3 and 4), then *a fortiori* MC assessments would contain more information about learning outcomes than CR assessments.

Unfortunately, we cannot in good faith impose these assumptions in our study. In our assessments, marking of the CR questions was done according to a highly prescribed marking schedule, with one or at most two markers grading a given CR question. While there no doubt remained some sampling error due to subjective marking, we cannot claim that this dominated the sampling error associated with random guessing on MC questions.

If CR questions are used then it becomes important to minimise “sampling error” as per our framework above. This can be done at principles level by (i) minimsing the number of markers involved in marking questions; and (ii) using tight, point marking schedules with little room for judgement on the part of the marker.

V. A CLOSER LOOK AT THE MC AND CR QUESTIONS UNDERLYING THIS STUDY

The main argument of this study is that even when a conscientious effort is made to write CR questions that test different levels of knowledge than MC questions, CR and MC questions produce similar grade outcomes. In this section, we first review the literature on the ability of MC and CR questions to measure higher-order learning outcomes. We then describe the CR and MC questions used in the assessments analyzed by this study. This information is useful for determining the extent to which our results may be valid for other university, introductory economics courses.

Bloom (1956) defines the following six levels of learning (our expanded explanations are in parentheses);

1. Knowledge (knowing facts);
2. Comprehension (understanding the importance of known knowledge);
3. Application (putting knowledge and understanding to use);
4. Analysis (using knowledge to breaking down a problem into component parts);
5. Synthesis (combining different parts to form new knowledge and ideas); and
6. Evaluation (determining the worth or usefulness of knowledge, application, analysis or synthesis).

Textbook, MC test banks tend to consist of questions that disproportionately sample from the first two levels of learning. Buckles and Siegfried (2006) conclude that MC questions can be effectively used to assess up through the first four levels of Bloom's taxonomy. In contrast, they argue that while it is possible to use MC questions to assess synthesis and evaluation, these are more reliably measured through CR questions. According to Buckles and Siegfried (2006), the key ingredient for assessing these higher-level learning outcomes is the requirement that students work through a chain of reasoning using a number of logical steps. It is difficult to write a sequence of MC questions that get at this learning dimension, especially when the chain of reasoning can involve a complicated decision tree.

These conclusions find support elsewhere in the literature. As part of a wider study, Iz and Fok (2007) attempt to classify the set of 25 MC questions used in the test for the

Higher Diploma of Surveying. They classify 21 of the 25 as levels 1 to 4. The remaining four questions were simply lumped together as "they were few in numbers… and difficult to discriminate". Zheng et al (2008) assert that it is "…much more difficult to write multiple-choice questions at the application and analysis levels of Bloom's taxonomy than at the knowledge or comprehension levels." It is even more difficult to write synthesis and evaluation MC questions. Thus it is no surprise that standard textbook question banks are dominated by recognition-, recall-, and understanding-type questions.

Walstad (2006) concurs with Buckles and Siegfried to a large extent, but notes that many CR questions are not well-designed to assess higher-level learning. Unless they are carefully constructed, CR questions may only be testing recall and recognition. A key issue is whether the student could have memorized the answer in advance.

We next describe the nature of the MC and CR questions used in the assessments included in our data set.[6] MC and CR were deliberately constructed to assess different levels of knowledge. The first example is a MC question that was designed to test for Knowledge (Level 1 of Bloom's taxonomy).

Which of the following is NOT an impact of inflation?

1. *Wealth is transferred from savers to borrowers.*
2. *Important price signals become more difficult to read.*
3. *The currency loses value.*
4. *The value of money assets rises.*

The next example is another MC question, but this one was designed to test for Application and Analysis (Levels 3 and 4).

A recession in the rest of the world is likely to cause ___________ GDP growth and ____________ inflation in New Zealand.

1. *higher; higher.*
2. *higher; lower.*
3. *lower; higher.*
4. *lower; lower.*

[6] The questions are taken from the term-test and final exam for Introduction to Macroeconomics (ECON 105), Semester One, 2006.

Assessing higher levels of knowledge becomes much more difficult with MC questions. This is where CR questions provide an opportunity to assess levels of knowledge that cannot, or at least are not, being measured by MC questions.

The following example is taken from the same course as the questions above. It illustrates how a CR question can be written such that higher levels of learning are progressively tested as the student works their way through the question.

In 1989, the Government passed the Reserve Bank Act. How would you characterise the NZ economy since that time in terms of growth, inflation and unemployment?

This question tests Knowledge and Comprehension (Levels 1 and 2). It could be easily rewritten in a MC format. Marks were awarded for stating how economic growth, inflation and unemployment had performed over this period in general terms (knowledge). Marks were also awarded for answers that commented on the importance of these facts (e.g. recent slowing of growth at that time).

A follow-up CR question is:

The Reserve Bank Monetary Policy news release above [not shown here] was issued on 9 March 2006. In this release the Bank identifies a number of factors that are influencing both inflation and growth. Use an AD/AS model to explain how the Reserve Bank currently sees the following factors influencing inflation and growth (remembering that the AD/AS model is a static model so you will need to interpret the results).
(i) the slowing (or cooling) of the housing market.
(ii) labour costs.
(iii) business confidence.

This question tests Application, Analysis and some Synthesis (Levels 3, 4, and 5). Students are required to break down the economic factors identified in the Reserve Bank news release and to use the AD/AS model to analyse the question. The student needs to have a good working knowledge of the AD/AS model because the question does not explicitly identify how AD/AS are affected by the respective factors. Further, the student must bring these factors together to determine their overall impact on growth and inflation. The latter involves extending results from the static model (price and GDP level) to a dynamic world (inflation and growth).

The next CR question succeeds the previous one and moves to Synthesis and Evaluation (Levels 5 and 6):

> *If the three influences analysed above were the only factors impacting the NZ economy, what conclusions would you make about the outlook for inflation and growth?*

Students must combine all three answers into one overall judgement. From the answers to the previous question there is no ambiguity about the impact on economic growth but the impact on inflation of these three influences is ambiguous. Students need to recognise this and answer accordingly. The question and the resources provided with the question contain little guidance for the student. Further, students most provide a consistent answer based on their previous answer.

Typically, students who have learnt some facts will achieve a good score on the first CR question. Students who have learnt the mechanics of the AD/AS model will earn at least some of the marks for the second CR question. The most able students will earn marks for the last CR question.

These latter examples are designed to illustrate the difficulty with writing MC questions to assess the highest levels of learning. These levels of learning are best assessed when the student is asked to analyze a complex economic question that requires them to assemble a chain of logical arguments. Consider the problem of assessing such a problem with MC question(s). If a single MC question is used to assess a problem of great complexity, fairness would dictate that it be worth many more points than simple recognition, MC questions. But the all-or-nothing marking of MC questions makes this a risky measure. In contrast, if a sequence of MC questions are used to assess the different parts of the logical chain, it is difficult to not lead the student into the answer by virtue of asking the question(s). The combination of their free-response nature, along with partial-credit marking, endows the CR question format with the potential to better assess higher-level learning while maintaining fairness to students.

In conclusion, MC and CR questions are most likely to produce different outcomes when an intentional effort is made to use them to assess different levels of knowledge. Such was the case in the introductory economics classes from which our sample was drawn. If substantial grade differences were not observed here, then it is unlikely that they will be observed when such an effort is not made.

V. CONCLUSION

This study attempts to identify the benefits in terms of grade outcomes from using constructed-response (CR) questions. It analyses students' performances on multiple-choice (MC) and constructed-response (CR) questions. One possible benefit of using CR questions is that they allow an instructor to assess higher levels of learning than is possible with MC questions. If that is the case, then the grade outcomes using CR questions may well be different from those using MC questions.

We use a large dataset composed of individual assessment results from thousands of students in introductory economics classes at a large public university. The dataset allows us to analyze MC and CR differences both within the same assessment, and across different assessments (a Term Test and a Final Exam).

We first show that observed MC-CR differences on the same assessment consist of two components: (i) a "systematic" difference that is related to, among other things, students' relative advantages/disadvantages in answering CR questions; and (ii) a component based on sampling error. We demonstrate the impossibility of isolating the two components in the absence of imposing restrictive assumptions. However, we show that an analysis of MC-MC and CR-CR differences across assessments can provide information that is still useful to instructors considering switching to all-MC assessments.

In particular, we find that the differences between MC and CR components within an assessment are comparable in size to those that occur across assessments. This suggests that switching to all-MC assessments would not cause a reordering of students' grades that is significantly different in magnitude to that which occurs now with a term test and a final exam. Thus, if there is a benefit to using CR questions, it would not seem to be because CR questions produce substantially different grade outcomes than MC questions.

Our study is the only one that we are aware of that directly analyzes grades in university economics classes. Our results are broadly consistent with a number of other studies that report high correlations between students' performances on MC and CR questions (e.g., Walstad and Becker, 1994; Kennedy and Walstad, 1997). Previous research by two of the authors of this study concluded that MC and CR questions measure different things (Hickson and Reed, 2009). Using the same data employed in that study, we are now able to show that the differences are not large enough to result in substantially different grades. These things are not inconsistent. MC and CR questions can measure different things yet what they measure is so highly correlated that rank ordering is not substantially affected – at least is not affected any more than that which already exists with two assessments (term test and final exam).

Does this mean that there is no benefit to using CR questions? We emphasize that our results represent the experience of one university, and a particular set of classes within that university. Further research needs to be done at different locations to determine whether the results reported here are valid elsewhere.

Even if CR questions produce similar grade outcomes to MC questions, there may be other benefits to using CR questions. For example, the existence of CR questions that test higher levels of learning may prompt desirable behavioural changes in the study habits of students. It could cause students to study harder, and to study concepts more deeply. A short survey of students revealed at least a stated preference consistent with this possibility. Field experiments may be necessary to investigate whether these benefits exist and are of sufficient size to justify the higher cost of CR questions. Not surprisingly, students have a strong view of "fairness" when it comes to assessments. They clearly see that having two items of assessment (a term test and a final exam) is fairest and similarly tests with a mix of both MC and CR questions are perceived as the most fair. Even if CR and MC produce similar outcomes within a course, they may have different capacities to forecast success in subsequent courses. This is particularly important in introductory classes where some degree of screening is expected to take place.

In conclusion, a final contribution of this paper is that it highlights the need for research to identify the benefits associated with CR questions. The analysis used in our study should be easy to replicate from student records at other universities. We hope this research will stimulate further studies along these lines. Such research could be of great benefit to instructors of introductory, university economics classes deliberating on whether to continue using constructed-response assessments.

REFERENCES

Becker, W. E., & Johnston, C. (1999). The Relationship between Multiple Choice and Essay Response Questions in Assessing Economics Understanding. *Economic Record, 75,* 348-357.

Bennett, R., E., Rock, D., A., & Wang, M. (1991). Equivalence of Free-Response and Multiple-Choice Items. *Journal of Educational Measurement, 28(1),* 77-92.

Hickson, Stephen and Reed, W. Robert. "Do Constructed-Response and Multiple-Choice Questions Measure the Same Thing?" Working paper, University of Canterbury, May 2009.

Kennedy, P. E., & Walstad, W. B. (1997). Combining Multiple-Choice and Constructed Response Test Scores: An Economists View. *Applied Measurement in Education, 10(4),* 359-375.

Krieg, R., G., & Uyar, B. (2001). Student Performance in Business and Economic Statistics: Does Exam Structure Matter? *Journal of Economics and Finance, 25(2),* 229-241.

Lukhele, R., Thissen, D., & Wainer, H. (1994). "On the Relative Value of Multiple-Choice, Constructed Response, and Examinee-Selected Items on Two Achievement Tests." *Journal of Educational Measurement, 31(3), 234-250.*

Lumsden, K.G, & Scott, A (1987). The Economics Student Reexamined: Male-Female Differences in Comprehension. *Journal of Economic Education, 18(4),* 365-375.

Thissen, D., Wainer, H., & Wang, X. (1994). Are Tests Comprising Both Multiple-Choice and Free-Response Items Necessarily Less Unidimensional Than Multiple-Choice Tests? An Analysis of Two Tests. *Journal of Educational Measurement, 31,* 113-123.

Wainer, H. & Thissen, D. (1993). Combining multiple-choice and constructed response test scores: Towards a Marxist theory of test construction. *Applied Measurement in Education, 6,* 103-118.

Walstad, W. B., & Becker, W. E. (1994). Achievement Differences on Multiple-Choice and Essay Tests in Economics. *American Economic Review, 84,* 193-196.

TABLE 1
Grade Distributions

Grade (Points)[a]		*MC (Term Test) (1)*	*MC (Final Exam) (3)*	*CR (Term Test) (2)*	*CR (Final Exam) (4)*	*Historical Distribution of Final Grades (5)*
A+	***(9)***	11.1%	17.4%	4.0%	5.2%	3.4%
A	***(8)***	15.5%	16.0%	4.2%	5.5%	4.7%
A-	***(7)***	8.9%	8.6%	3.4%	5.6%	7.1%
B+	***(6)***	8.9%	16.3%	6.6%	8.8%	7.8%
B	***(5)***	9.3%	7.8%	5.7%	7.6%	10.6%
B-	***(4)***	17.2%	13.4%	9.5%	10.2%	10.9%
C+	***(3)***	6.4%	4.7%	6.9%	7.3%	12.6%
C	***(2)***	6.3%	7.3%	10.1%	9.9%	19.7%
C-	***(1)***	5.2%	2.9%	7.4%	6.8%	8.5%
D	***(0)***	6.8%	3.0%	10.1%	8.3%	5.5%
E	***(-1)***	4.3%	2.7%	32.2%	24.7%	9.3%
Mean		66.5	70.3	49.5	53.4	---
Std. Dev.		16.3	15.3	20.2	21.4	---

[a] "Points" identifies the point allocation for each grade as used to calculate Grade Point Average (GPA).

NOTE: The bottom two rows in the table report the mean and standard deviation of raw assessment marks (adjusted to a 100-point scale) by assessment category. Columns (1) through (4) report the grade distributions that would arise from applying the UC grading schedule to the raw assessment marks. Column (5) reports the historical distribution of Final Grades given in all classes in our sample. A total of 15,508 observations are represented in the table.

TABLE 2
Distribution of MC Grades Conditional on CR Grades

		A+	A	A-	B+	B	B-	C+	C	C-	D	E	Sum
CR Grade (Term Test)	A+	**28.5**	17.2	18.0	12.4	7.1	6.7	6.7	3.4	0.0	0.0	0.0	100
	A	14.7	**13.9**	22.1	13.9	8.7	10.4	9.5	6.3	0.0	0.3	0.3	100
	A-	6.2	13.1	**14.9**	15.5	13.5	14.5	10.7	7.5	2.7	0.7	0.7	100
	B+	6.1	7.3	12.1	**11.8**	15.8	14.1	12.0	14.1	4.0	1.3	1.3	100
	B	2.6	5.0	9.5	9.5	**15.1**	15.0	15.8	17.4	5.0	2.3	2.8	100
	B-	2.9	4.5	6.9	10.0	11.4	**11.3**	16.0	20.5	8.3	3.3	4.9	100
	C+	0.8	3.5	5.5	8.1	12.4	11.2	**14.3**	25.1	8.2	4.1	6.8	100
	C	0.6	2.0	4.0	5.8	10.0	10.9	12.8	**24.9**	11.4	7.0	10.6	100
	C-	0.6	0.8	1.5	2.3	9.6	9.1	12.3	26.1	**10.8**	11.4	15.5	100
	D	0.0	0.7	0.7	2.1	5.2	7.3	9.9	23.7	17.1	**14.8**	18.5	100
	E	0.0	0.4	0.3	1.3	3.1	5.0	9.6	21.4	15.3	11.2	**32.5**	100

		A+	A	A-	B+	B	B-	C+	C	C-	D	E	Sum
CR Grade (Final Exam)	A+	**35.6**	20.2	14.6	14.6	7.1	4.1	2.6	0.0	0.0	0.0	1.1	100
	A	12.8	**20.2**	21.8	13.6	13.9	10.1	3.3	3.3	0.5	0.0	0.5	100
	A-	8.7	15.1	**19.8**	15.5	16.9	8.9	8.5	5.6	0.5	0.2	0.2	100
	B+	4.8	10.8	16.3	**16.4**	17.1	13.0	11.1	8.1	1.2	0.3	0.8	100
	B	1.7	4.5	9.1	13.2	**19.5**	16.1	15.7	16.3	2.6	0.7	0.6	100
	B-	2.1	2.6	7.4	9.9	13.3	**15.6**	19.7	20.3	5.3	1.9	1.9	100
	C+	1.0	2.2	3.9	6.3	13.4	15.2	**16.5**	24.3	9.8	3.8	3.6	100
	C	0.1	0.5	2.5	3.5	7.4	11.1	15.9	**29.9**	13.5	8.1	7.3	100
	C-	0.5	0.2	1.1	2.0	3.0	6.5	11.5	30.5	**14.9**	12.9	16.9	100
	D	0.2	0.2	0.5	1.4	3.1	5.6	8.5	27.7	18.1	**14.1**	20.7	100
	E	0.0	0.1	0.3	0.6	0.8	2.6	4.3	16.1	14.3	13.2	**47.6**	100

NOTE: Numbers in tables are percentages and should be interpreted as follows: 28.5 percent of all Term Test takers who received an A+ on their CR component also earned an A+ on the MC component of the test (after norming the respective grade distributions to the historical averages in TABLE 1). 17.2 percent of all Term Test takers who received an A+ on their CR component earned an A on the MC component of that assessment. And so on. As the probabilities are conditional on grade received on the CR component, all row probabilities sum to 100%.

TABLE 3
Comparison of Grade Differences for Different Combinations of MC and CR Assessments

	Negative Difference of One Letter Grade or More	**Difference of Less Than One Letter Grade**	**Positive Difference of One Letter Grade or More**
MC-CR (Term)	16.2%	67.0%	16.7%
MC-CR (Final)	11.4%	77.3%	11.3%
MC(Term) - MC(Final)	16.8%	66.6%	16.5%
CR(Term) - CR(Final)	12.6%	74.9%	12.5%

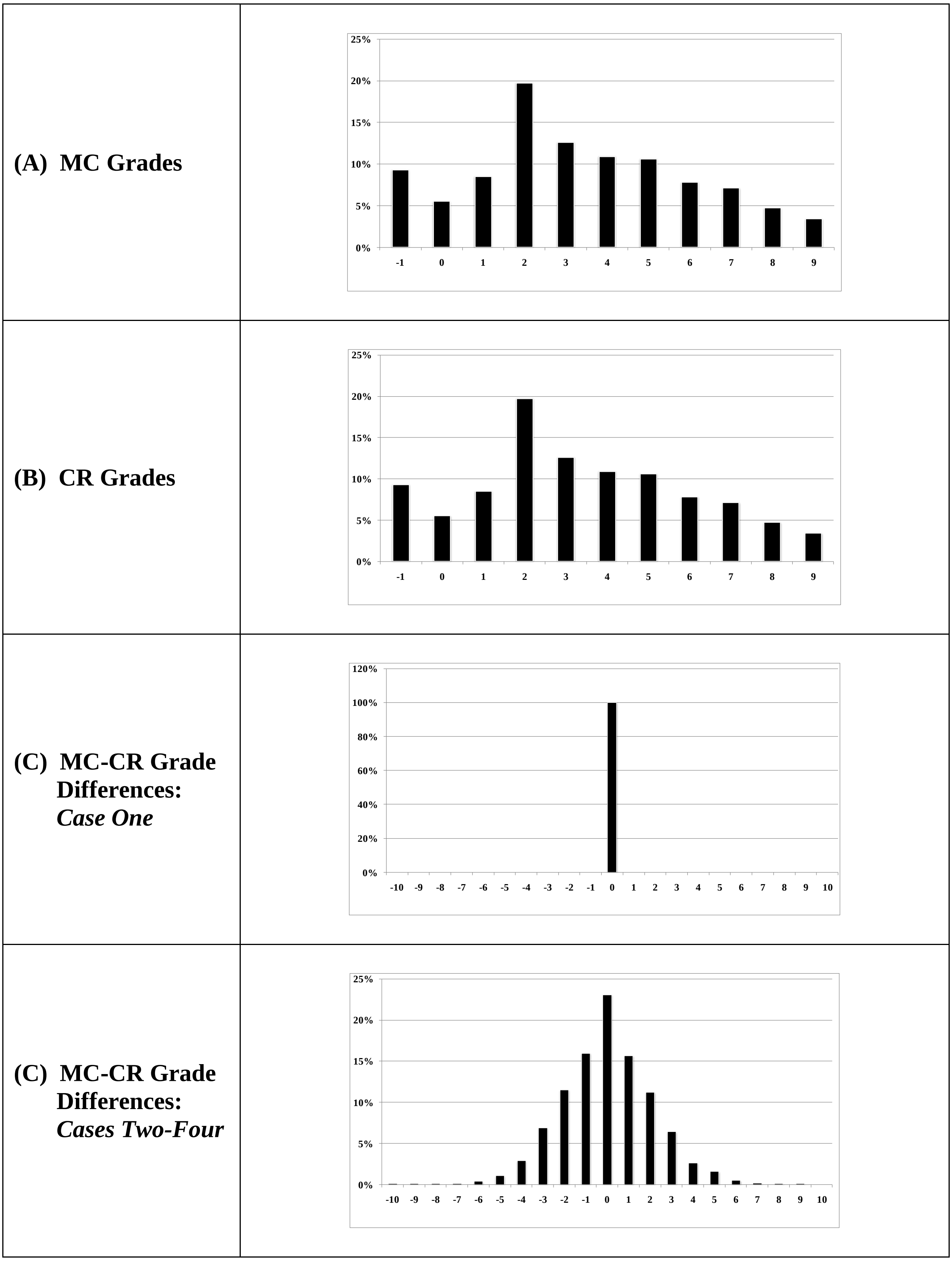
NOTE: The first two rows use the data from TABLE 2 to calculate differences between CR and MC grades on the same assessment. Examples of a “Negative Difference of One-Letter Grade” are CR Grade = A- and MC Grade = B-; CR Grade = B+ and MC Grade = C+; etc. A “Positive Difference of One-Letter Grade” is defined similarly. The last two rows report grade differences between (i) MC grades on the Term Test and Final Exam and (ii) CR grades on the Term Test and Final Exam, respectively.

TABLE 4
Student Perceptions of Different Assessment Options and Formats

	Question:	**Percent**
Q1	**Which of the following assessment options do you prefer?**	
	Having a term test and a final exam	93.1
	Having a final exam only	1.5
	I would be equally happy with either of the above two.	5.4
	TOTAL	100.0
Q2	**For a test or exam, which format do you think would give you the highest rank in class?** **(This question is not about which format gives you the highest score. Raw scores for multiple choice are usually higher due to guessing. This is about your rank in class).**	
	A test where all the questions are multiple choice.	14.5
	A test where all the questions require a written answer.	15.3
	A test with a mixture of multiple choice and written answer questions.	59.5
	The test format would not make any difference to my rank in class.	10.7
	TOTAL	100.0
Q3	**Which assessment option do you think is the fairest for students?**	
	Having both a term test and a final exam.	86.9
	Having a final exam only.	1.5
	Both of the above options are equally fair.	11.5
	TOTAL	100.0
Q4	**For a test or exam, which format do you think is fairest for students?**	
	A test where all the questions are multiple choice.	3.1
	A test where all the questions require a written answer.	6.9
	A test where there are both multiple choice and written answer questions.	79.2
	All of the above are equally fair.	10.8
	TOTAL	100.0
Q5	**Which one of these best describes you?**	
	I would study harder for a test if all the questions were multiple choice.	5.3
	I would study harder for a test if all the questions required a written answer.	35.1
	The test format makes no difference to how hard I would study.	59.5
	TOTAL	100.0
	For students who expressed a preference for MC or CR in Q2, what were their responses in Q4?	
	A test where all the questions are multiple choice.	5.1
	A test where all the questions require a written answer.	15.4
	A test where there are both multiple choice and written answer questions.	69.2
	All of the above are equally fair.	10.3
	TOTAL	100.0

Total number of responses: 131

FIGURE 1
Distributions of MC Grades, CR Grades, and MC-CR Grade Differences



(A) MC Grades

(B) CR Grades

(C) MC-CR Grade Differences: ***Case One***

(C) MC-CR Grade Differences: ***Cases Two-Four***

NOTE: Panels (A) and (B) report the distributions of MC and CR grades for a given assessment, where grades are measured in points as reported in TABLE 1. Each assessment is “normed” to have the same distribution as the historical distribution of Final Grades for all student observations in our sample (cf. Column 5 in TABLE 1). Panels (C) and (D) report the distributions of MC-CR grade point differences that would arise under different cases. For example, a student who earned an A- (7 points) on the MC component of an assessment, and a B- (4 points) on the CR component of that assessment would have a grade difference of -3 points. The four cases characterize: (i) no sampling error and no “systematic differences;” (ii) no sampling error and “systematic differences;” (iii) sampling error and no “systematic differences;” and (iv) both sampling error and “systematic differences.” The four cases are discussed in greater detail in the text.