DEPARTMENT OF ECONOMICS AND FINANCE

COLLEGE OF BUSINESS AND ECONOMICS

UNIVERSITY OF CANTERBURY

CHRISTCHURCH, NEW ZEALAND

DO CONSTRUCTED-RESPONSE AND MULTIPLE-CHOICE QUESTIONS MEASURE THE SAME THING?

Stephen Hickson and W. Robert Reed

WORKING PAPER

No. 08/2009

Department of Economics and Finance
College of Business and Economics
University of Canterbury
Private Bag 4800, Christchurch
New Zealand

WORKING PAPER No. 08/2009

DO CONSTRUCTED-RESPONSE AND MULTIPLE-CHOICE QUESTIONS MEASURE THE SAME THING?

Stephen Hickson[1], W. Robert Reed[2]

May 2009

Abstract: Our study empirically investigates the relationship between constructed-response (CR) and multiple-choice (MC) questions using a unique data set compiled from several years of university introductory economics classes. We conclude that CR and MC questions do not measure the same thing. Our main contribution is that we show that CR questions contain independent information that is related to student learning. Specifically, we find that the component of CR scores that cannot be explained by MC responses is positively and significantly related to (i) performance on a subsequent exam in the same economics course, and (ii) academic performance in other courses. Further, we present evidence that CR questions provide information that could not be obtained by expanding the set of MC questions. A final contribution of our study is that we demonstrate that empirical approaches that rely on factor analyses or Walstad-Becker (1994)-type regressions are unreliable in the following sense: It is possible for these empirical procedures to lead to the conclusion that CR and MC questions measure the same thing, even when the underlying data contain strong, contrary evidence.

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

JEL Classifications: A22

Acknowledgements: We want to acknowledge Peter Kennedy for copious comments on earlier versions of this paper. His comments have resulted in substantial improvements in our analysis.

[1] Department of Economics and Finance, University of Canterbury, Private Bag 4800, Christchurch 8042, New Zealand
[2]Department of Economics and Finance, University of Canterbury, Private Bag 4800, Christchurch 8042, New Zealand

*Corresponding Author: W. Robert Reed, Department of Economics and Finance, University of Canterbury, Private Bag 4800, Christchurch 8042, New Zealand; Email: bobreednz@yahoo.com; Phone: +64 3 364 2846.

WORKING PAPER No. 08/2009

DO CONSTRUCTED-RESPONSE AND MULTIPLE-CHOICE QUESTIONS MEASURE THE SAME THING?

> "In sum, the evidence presented offers little support for the stereotype of multiple-choice and free-response formats as measuring substantially different constructs."
>
> - Bennett, Rock, and Wang (1991)

> "Whatever is being measured by the constructed-response section is measured better by the multiple-choice section…We have never found any test that is composed of an objectively and subjectively scored section for which this is not true."
>
> - Wainer and Thissen (1993)

> "The findings from this analysis of AP exams in micro and macro principles of economics are consistent with previous studies that found no differences, or only slight differences, in what the two types of tests and questions [multiple-choice and essay] measure."
>
> - Walstad and Becker (1994)

I. INTRODUCTION

University principles of economics courses often have enrollments of several hundred students or more. Instructors of these courses face a potential tradeoff when designing tests: On the one hand, constructed-response (CR) questions are thought to assess important learning outcomes that are not well-addressed by multiple-choice (MC) questions. On the other hand, constructed-response questions are much more costly to grade. In addition, the marking of CR questions is less reliable due to the subjective nature of the questions.

Ideally, one would weigh the respective benefits and costs of CR and MC questions to decide the optimal mix of each to employ. However, this is a difficult task,

especially given the subjective nature of "benefits."[1] Perhaps because of this, much attention has focused on the question, "Do CR and MC questions measure the same thing?" If this question could be answered in the affirmative, it would mean there was no "tradeoff," and one could eliminate CR questions. In fact, a number of influential studies claim to demonstrate this result. The implications of this have been well-understood:

> The educational measurement literature suggests that multiple-choice questions measure essentially the same thing as do constructed-response questions. Given the higher reliability and lower cost of a multiple-choice test, a good case can be made for omitting constructed-response questions from a test containing both multiple-choice and constructed-response questions because they contribute little or no new information about student achievement (Kennedy and Walstad, 1997, page 359).

Previous research has taken different approaches to this question. Bennett, Rock, and Wang (1991) and Thissen, Wainer, and Wang (1994) employ factor analysis. Walstad and Becker (1994) regress AP composite scores on MC scores. Kennedy and Walstad (1997) simulate grade distributions using different test formats. Each of these has its own notion of what it means to "measure the same thing," and none attempts to reconcile their approach to those of others.

Our study proposes its own approach to this question. We investigate whether CR scores are "predictable" from MC scores. If a student's performance on the CR component of a test can be perfectly, or near-perfectly, predicted by their performance on the MC component, we could easily conclude that the two components "measure the same thing."

[1] The only study that we are aware of that attempts such an approach is Kennedy and Walstad (1997). They frame the decision to use CR questions as a tradeoff between reduced "misclassifications" and higher marking costs. "Misclassifications" are defined as estimated differences in the grade distribution (beyond natural sampling variation) that would arise on the AP micro- and macroeconomics exams from switching to an all-MC format. Unfortunately, in order to categorize these as "misclassifications," KW must assume that the mix of CR and MC questions on the AP tests is optimal. If the mix is not optimal, then it doesn't follow that the grade distribution under an all-MC format is worse than under the mixed format. This highlights the practical difficulties of implementing the "benefits versus costs" approach.

Not surprisingly, we find that the regression of CR scores on MC scores leaves a substantial residual. The first innovation of our study is that we are able to demonstrate that this residual is empirically linked to student achievement. Since the residual represents the component of CR scores that cannot be explained by MC scores, and since it is significantly correlated with learning outcomes, we infer that CR questions contain "new information about student achievement" and therefore do not measure the same thing as MC questions.

The preceding analysis demonstrates CR questions provide extra information not contained in the existing set of MC questions. But do they contain information beyond what could be provided by an all-MC assessment? It is the latter question that is salient for those contemplating a switch from a composite assessment to one consisting of all MC questions. The second innovation of our study is that we exploit the panel nature of our data to construct a quasi-counterfactual experiment. We show that combining one CR and one MC component always predicts student achievement better than combining two MC components.

The final section of our study explores explanations for why our research obtains results that are at variance with many previous studies. We are able to replicate the key findings of a number of these studies. This suggests that our different results are not driven by differences in the data, but by differences in empirical methodologies. We conclude with recommendations for future research.

II. DATA

Our analysis uses data compiled over a six-year period (2002-2007) from approximately 8400 students in two different courses at the University of Canterbury in New Zealand.

Introductory Microeconomics and Introductory Macroeconomics are semester-long courses typically taken by business students in their first year of study. Both courses administer a mid-semester "term test" and an end-of-semester final exam.

Both term tests and final exams consist of a CR and a MC component. While the weights given to these components are different for the term test and the final exam, and change somewhat over the years, the structure of these components has remained constant. For both courses, the term test is 90 minutes long and consists of 25 MC and two CR questions. The final exam is longer at 180 minutes, and consists of 30 MC and three CR questions. There has been little change in the coverage of the respective assessments over the years with one exception: In 2007, the final exam gave more coverage to material in the first half of the course. Inasmuch as possible, quality control across assessments is maintained by the fact that the same two instructors taught the classes, and wrote and graded the assessments across the whole time period.

All together, the data set includes assessments from ten separate offerings of Introductory Microeconomics and eight of Introductory Macroeconomics, for a total of 36 assessments (18 term tests plus 18 final exams). When we eliminate incomplete records and students for whom one of the assessments is missing, we are left with 16,710 observations.[2] By way of comparison, Walstad and Becker (1994) have a total of 8,842 observations. Most studies have far fewer.[3]

[2] The main reasons for deleting observations were the following: (i) A student received an aegrotat pass. Students apply for an aegrotat pass when they are unable to attend an assessment or their performance has been impaired due to illness or other unforeseen circumstances. (ii) A student had a missing term test or final exam score for some other reason. (iii) A student received a total score for the course equal to zero. These students did not attempt any assessment item.

[3] Krieg and Uyar (2003) have only 223 observations.

There are two features which make our data set unique. First, we have repeated observations on the same student for a given course. This allows us to test whether CR scores on the term test provide "new information" that can be used to predict student achievement on the final exam. Second, we have information about the student's achievement in other courses. This allows us to test whether CR scores in an economics course provide "new information" about student achievement outside the class.

The two key variables in our study are student scores on the CR and MC components of their term tests/final exams. These are calculated as percentages out of total possible scores. Panel A of FIGURE 1 reports a histogram and statistical summary for the full sample of CR scores. The average score is 52.53, and there is evidence of clumping as a result of the way in which the percentage scores are calculated. The lower panel of FIGURE 1 provides a similar report for the MC scores in our study. These are characterized by a higher mean (68.38) and smaller spread.

Also noteworthy in FIGURE 1 is that the distribution of test scores is constrained to lie between 0 and 100. Amongst other problems, this will cause the errors associated with a linear regression specification to be heteroscedastic. We address this problem in two ways. First, we use OLS but estimate the standard errors using the heteroscedastic-robust White procedure. OLS has the advantage of facilitating interpretation of the coefficient estimates. Accordingly, these are the results we report in our paper. However, we also estimated the key regressions using the more statistically appropriate fractional logit procedure. The results were virtually identical.[4]

TABLE 1 provides a statistical summary of the students represented in our study. Approximately 55 percent of the sample derived from Introductory Microeconomics

[4] The fractional logit results are available upon request from the authors.

classes. By construction, the data set consists of exactly half term test and half final exam results. TABLE 1 also breaks down the CR and MC scores by term test and final exam. Both components show higher scores on the final exam. This is consistent with the fact that the term test is more time-constrained than the final exam. While the final exam has the twice the allotted time as the term test, it is designed to have less than twice the work.

The variable *GPA* reports the student's grade point average for all courses outside of ECON 104 (Introductory Microeconomics) and ECON 105 (Introductory Macroeconomics) in the same year that the student was enrolled in the respective economics class. For example, if a student was enrolled in ECON 104 in Semester 1 of 2005, *GPA* reports their grade point average for all courses they took in calendar year 2005, excluding ECON 104 and 105. Grade points range from -1 (for a letter grade of E = fail) to 9 (for a letter grade of A+). The variable *COMPOSITE* is a weighted average of the CR and MC components, and is used later in the study when we estimate Walstad and Becker (1994)-type regressions.

While not reported in TABLE 1, approximately 56 percent of the sample is male. A little less than half of the students in our sample are New Zealand natives or of European extraction. Approximately 43 percent of the students are Asian. This high percentage is due to a surge in Asian enrollments that occurred in the early 2000's in New Zealand universities. This tapered off substantially in the latter years of the sample. Maori, Pacific Islanders, and Others (primarily Africans and Middle Easterners) account for less than 8 percent of our sample. With respect to language, most of the sample declared English as their "first language." Even so, a little less than 40 percent declared

that English was not their "first language," with the great majority of these identifying with Chinese.

III. RESULTS

The first step of our analysis consists of determining to what extent performance on the CR component of an assessment is "predictable" from the student's MC score on that assessment. If the corresponding regressions produce R^2 values close to one, this would clearly indicate that CR scores added little information to that already provided by the student's MC performance. We could then confidently conclude that CR and MC questions measured the same thing.

TABLE 2 summarizes the results of this analysis. We divided our data set into four, mutually exclusive sets of observations: (i) term tests and (ii) final exams from Introductory Microeconomics classes; and (iii) term tests and (iv) final exams from Introductory Macroeconomics classes. For each sample, we regressed students' CR scores on their MC scores for the same assessment. In addition, we aggregated all the observations into one sample. Not surprisingly, we find that MC scores are significant predictors of students' CR scores. An extra point on the MC component predicts an additional 0.7 to 1.1 points on the CR component, depending on the sample.

On the other hand, we also find that the R^2 values are never close to 1. The R^2 values for the final exam regressions are close to 50 percent. Those for the term tests are even lower, in the low- to mid-30's.[5] (We discuss this difference between term tests and

[5] Conventional wisdom is that CR questions are "noisier" assessments. This view is supported by the fact that CR scores have greater dispersion (cf. FIGURE 1 and TABLE 1).

final exams below.) For the full sample, the R^2 of the regression of CR scores on MC scores is a little less than 40 percent.[6]

To facilitate comparison with other studies, the last line of the table reports the simple correlation between CR and MC scores. Walstad and Becker (1994, page 194) report simple correlations of 0.69 and 0.64 for the Micro and Macro AP tests. Lumsden and Scott (1984, page 367) report correlations of 0.18 and 0.26 for introductory Micro and Macro courses, respectively. In contrast, they cite a number of other studies where the correlations range higher, though still lower than reported here. Thus, our finding that CR scores are far from being perfectly, or even near perfectly, predictable from MC scores appears to be the norm.

Unfortunately, while an R^2 close to 1 provides strong evidence that CR and MC questions measure the same thing, it is unclear what an R^2 far from 1 implies. Is the unexplained component in CR scores due to the fact that CR questions measure something different than MC questions? Or are the two question-types assessing the same thing(s) but with measurement error?

If we had an alternative measure of student learning, we could take the residuals from the regressions in TABLE 2 and test if they were independent predictors of academic achievement. If the residuals were unrelated to student learning, say were pure measurement error, then one would expect them to be unrelated to this alternative measure. Alternatively, if we could show that these residuals were positively related to this alternative measure, this would provide evidence that the residuals contained

[6] We also investigated the effect of including higher-order, polynomial terms for the MC variable. This added little to the overall explanatory power of the equations.

independent "information about student achievement" that was not captured by MC responses.

Unfortunately, we do not have an alternative measure of student learning for the same assessment. We do, however, have a close substitute. Because we have repeated observations for each student, we can test whether residuals from the term test regressions are related to achievement on the final exam. If the residuals represent pure measurement error, one would not expect to find any relationship with students' final exam performance.

Column (1) of TABLE 3 reports the results of a regression where students' CR scores from the final exam were regressed on (i) their MC scores from the term test, and (ii) the unexplained component of their CR score from the term test (i.e., the residual from the regression specification that was reported in TABLE 2).[7] We separate the 2002-2006 and 2007 final exams because the 2007 final exams included a larger share of material from the first half of the course. We also separate the Introductory Microeconomics and Introductory Macroeconomics final exams. In each of the six samples, the *Residual* variable has very large *t*-values. In addition, the respective coefficients are all positively-signed.

It is interesting to contrast these results with a prediction from a well-cited study by Lukhele, Thissen, and Wainer (1994). They fit item response models to AP tests in Chemistry and History and conclude that MC questions are more reliable than CR questions. In their words, "This means that, if we wish to predict a particular student's score on a future test made of constructed response items, we could do so more

[7] The residual variables come from term test CR regressions using the same observations as the TABLE 3 samples (e.g., "All Observations (2002-2006)," "All Observations (2007)," etc.)

accurately from a multiple-choice than from a constructed response test that took the same amount of examinee time" (page 246). In fact, LTW are unable to confirm this statement because their data are cross-sectional. The panel nature of our data allows us to test, and reject, LTW's prediction.

Our results are consistent with the hypothesis that CR scores contain unique information about student learning. But is this unique "information" related to academic achievement? For example, suppose students with bad handwriting receive lower marks on CR questions, ceteris paribus. Then a lower score on the term test CR section could be predictive of a lower score on the final exam CR section because it was predictive of bad handwriting – i.e., something unrelated to learning outcomes.

To check this possibility, we also regressed students' final exam MC scores on the same two variables used to predict their final exam CR scores. The qualitative results remain unchanged. For each sample, the *Residual* variable is positively correlated and highly, statistically significant. In other words, the unexplained component of term test CR scores predicts student achievement on both the (i) CR and (ii) MC components of the final exam.

While this latter finding is strong evidence that the CR residuals contain information about learning outcomes, it raises another concern: If CR and MC questions measure something different, why should the term test CR residual have predictive power for the final exam MC score?

Our explanation recalls a number of previously noted characteristics about our data, and combines this with the educational psychology literature on learning goals. First, both CR and MC scores are lower for the term test than the final exam (cf. TABLE

1). Second, the R^2 values from the term test regressions in TABLE 2 are lower than the corresponding final exam regressions. Third, the term test is more time-constrained than the final exam (as evidenced by lower mean CR and MC scores).

Bloom's (1956) taxonomy predicts that MC questions are more likely to test the lower levels of educational objectives (i.e., Knowledge, Comprehension, Application, and, perhaps, Analysis). While CR questions test these as well, they are uniquely suited for assessing the more advanced learning goals (Synthesis and Evaluation).[8] Accordingly, one would expect CR to contain some unique information compared to MC, but also some overlap.

We now attempt to explain both the poorer predictability of MC scores on term tests (cf. TABLE 2), and the fact that term test CR scores are significant predictors of final exam MC scores (cf. Column 2, TABLE 3). Given the greater time-constraints, we hypothesize that students will devote relatively less time to the MC component on the term test; since MC questions can be answered very quickly, if necessary. However, the cost of this test-taking strategy is that students are less likely to get the more difficult MC questions correct (Application and Analysis). It is these more difficult MC questions that will test higher levels of learning.

As a consequence, the amount of informational "overlap" between the MC and the CR questions – as measured by the levels of educational objectives that are assessed – is likely to be lower for the term test than for the final exam. This will cause MC scores to be a worse predictor of CR scores on term tests compared to final exams.

[8] The six levels of Bloom's taxonomy are sometimes recast as follows (from lowest to highest): (i) Remembering, (ii) Understanding, (iii) Applying, (iv) Analyzing, (v) Evaluating, and (vi) Creating.

It will also cause the MC responses on the final exam to measure higher levels than the MC responses on the term test. Because the CR responses also assess these higher levels, the CR *Residual* will be able to predict final exam MC scores even after controlling for term test MC scores.

Summarizing the above, our results suggest that CR scores contain unique information not contained in the responses to existing MC questions. But there is now another concern: If CR questions have overlap with MC questions, then perhaps all of their "extra information" would be subsumed by including additional MC questions. Ideally, we would like to compare assessments using composite CR/MC questions with those using all-MC questions. While we cannot do this directly, we can construct a quasi-counterfactual to evaluate this concern.

A unique feature of our data is that we have information on students' grades in every course they have taken at the University of Canterbury. We use this information to calculate a GPA value based on their performance in non-introductory economics classes. For example, suppose a student took Introductory Microeconomics (ECON 104) in the first semester of 2005. We calculate their GPA over all other courses during the 2005 academic year, excluding their performance in ECON 104. If they subsequently took Introductory Macroeconomics (ECON 105) in the second semester of 2005, we also exclude their performance in that class.[9]

We begin by exploring to what extent the CR *Residual* variables are able to predict outside *GPA*, holding constant student MC scores. TABLE 4 reports the results.

[9] We chose to exclude both introductory economics classes because of similarities in the way the two classes were assessed. Since the two lecturers work closely together, it is possible that their assessment styles were similar. Correlation in performance across the two classes might represent students' ability to perform well on a particular style of assessment, and not an independent observation about student learning outcomes.

The four measures of student achievement for a given economics course are: the student's MC score on the (i) term test and (ii) final exam in that course; and the residuals from the (iii) term test and (iv) final exam CR regressions, also from that course. These latter two variables are generated from TABLE 2-type regressions and represent the component of the student's CR score that cannot be explained by their MC performance on the same assessment.

We divide our observations into the same six samples that we used in TABLE 3. For each sample, we investigate whether the individual *Residual* variables are positively and significantly related to their outside *GPA* values. We also perform an *F*-test of the joint significance of the two *Residual* variables. Once again, the results in every case are consistent with the hypothesis that CR scores measure independent information not captured by existing MC scores. An extra "unexplained" point on the CR component of an assessment is associated with an increase in their outside *GPA* of anywhere from 0.0064 (cf. Column 3, Sample 3b) to 0.0662 points (cf. Column 4, Sample 3b). Interestingly, final exam performance seems to be a better predictor of outside *GPA* for both MC and CR scores. The individual *Residual* variables are each statistically significant at generally high *t*-values. Furthermore, the joint *F*-tests all have *p*-values that indicate significance at the 0.01% level.

Thus, the CR *Residual* variables are significant even after we control for both the term test and final exam MC scores. While this is additional evidence that the CR *Residual* variables contain information on learning outcomes, we are still dogged by the concern that the additional information provided by the CR variables is merely a substitute for information that could be provided by including more MC questions.

If MC and CR questions measure the same thing(s), then both should be equally good at predicting students' *GPAs*. Accordingly, we compare the following regression models:

(i) $GPA_t = \beta_0 + \beta_1 MC(Term)_t + \beta_2 MC(Final)_t + \varepsilon_t$, and

(ii) $GPA_t = \alpha_0 + \alpha_1 MC(Term)_t + \alpha_2 CR(Final)_t + \eta_t$.

The pair of models conceptualizes the following thought experiment: Suppose an instructor had given an all-MC term test. Would he/she more effectively assess academic achievement if the final exam consisted of MC questions or CR questions? Specification (i) represents the case where assessment is based solely on MC questions. Specification (ii) represents a composite CR/MC assessment. A comparison of the R^2 values from estimating models (i) and (ii) across different samples should show no clear pattern – if MC and CR questions measure the same thing(s). However, if CR questions measure unique information, such as higher levels of the Bloom (1985) taxonomy, and if competency at these higher levels is positively correlated with student achievement, then the R^2 value from Specification (ii) should be consistently higher. As a further test, we also compare an alternative pair of regression models:

(iii) $GPA_t = \beta_0 + \beta_1 MC(Final)_t + \beta_2 MC(Term)_t + \varepsilon_t$, and

(iv) $GPA_t = \alpha_0 + \alpha_1 MC(Final)_t + \alpha_2 CR(Term)_t + \eta_t$.

TABLE 5 reports the results of this test. We divide the data into the same six samples used for TABLES 3 and 4. Consider the first two rows of TABLE 5. For the sample of all observations from 2002-2006 (Sample 1a), the regression of *GPA* on the two MC components produces an R^2 value of 0.424. In contrast, the "composite" regression of one MC and one CR component has an associated R^2 value of 0.526. The

composite "assessment" does a better job of predicting student achievement. Rows (3) and (4) perform a similar comparison, this time starting with the *MC(Final)* score and adding either the *MC(Term)* or *CR(Term)* score. Once again, the composite "assessment" does a better job of predicting student achievement. In fact, for every sample and every pair of regression models, a combination of CR and MC scores does a better job of predicting students' *GPAs* than relying solely on MC scores.

Taken together, the results from TABLES 2 through 5 provide strong evidence that the CR and MC questions in our data do not measure the same things. While other studies, such as Kennedy and Walstad (1997), find evidence that CR and MC responses are "different," our study is the first to link these differences to learning outcomes.

IV. RELATING OUR FINDINGS TO THOSE OF PREVIOUS STUDIES

Our finding that CR and MC scores do not measure the same thing is at variance with a number of influential studies. In this section, we want to explore whether this is due to differences in our data, or differences in empirical procedures.

Bennett, Rock, and Wang (1991) and Thissen, Wainer, and Wang (1994) are widely-cited studies from the educational measurement literature. BRW base their analysis from a sample of responses from the College Board's Advancement Placement (AP) examination in Computer Science. TWW re-analyze BRW's data, and add a similar sample from the AP exam in Chemistry. Both employ common factor analysis to study the relationship between "free response" and MC questions. Both find that a single factor

explains most of the variation in the respective questions. They therefore conclude that these two question-types measure the same thing.[10]

While BRW and TWW employ factor analyses, they use somewhat different techniques. BRW use a model in which free response and MC questions are each loaded on a single factor. These two (correlated) factors are then analyzed to determine whether they contain unique information. In contrast, TWW employ a more general procedure to decompose the variation in the two types of questions into multiple factors.

The AP exam in Computer Science consists of 50 MC questions, and 5 CR questions. The AP exam in Chemistry consists of 75 MC questions and 4 sections of CR questions, some of which contain multiple problems. BRW and TWW break up the respective components into multiple "parcels." BRW re-organize the 50 MC questions into five sets "("parcels") of ten questions each. TWW convert the original 75 MC questions into fifteen, five-question parcels. These parcels become, in a sense, separate variables which are then decomposed into factors.

We attempt to replicate BRW's and TWW's factor analysis results. If we cannot replicate their results, this would suggest that our data are substantially different from theirs. In contrast, if are able to replicate their results, this would indicate that our different conclusions derive from different empirical procedures. Given the preceding evidence on CR and MC questions, it would suggest that the factor analysis approach is unreliable for determining whether CR and MC questions measure the same thing.

Unfortunately, our data contain fewer questions than BRW and TWW and are thus less amenable to "parcelization." Instead, we apply principal component analysis

[10] While both studies find more than one significant factor, they both conclude that a single factor is able to explain most of the variation in the two types of questions.

(PCA) to students' scores on the CR and MC components. PCA is related to factor analysis in that its "principal components" are akin to the factors identified by factor analysis. It has the advantage in that it produces a unique decomposition of the correlation matrix.[11] In contrast, factor analysis typically involves a subjective procedure ("rotation") that allows one to generate alternative sets of factors from the same data. A particularly attractive feature of PCA for our purposes is that it yields a straightforward measure of the amount of variation "explained" by each of the principal components.

TABLE 6 reports the results of applying PCA to the same five samples we previously analyzed in TABLE 2. As there are only two variables (*Multiple-Choice* and *Constructed-Response*), there are a total of two principal components. By construction, these two principal components explain all of the "variation" in the correlation matrix.

The first item of interest in TABLE 6 is the column of "eigenvalues." These provide a measure of importance for each of the principal components. In factor analysis, two common approaches for choosing the number of factors are Kaiser's eigenvalue rule and Cattell's scree test. The first of these selects factors having eigenvalues greater than one. The second of these plots the eigenvalues in decreasing order and selects all factors immediately preceding an abrupt leveling off of the values. Both approaches lead to the conclusion that there is one main factor underlying students' CR and MC responses in each of the samples. This finding is reinforced by the second column in TABLE 6. "Proportion" translates these eigenvalues into shares of total variation in the correlation matrix. These range from 78-85 percent across the different samples.

[11] Non-unique solutions can arise when two or more eigenvalues are exactly equal, but this is rarely encountered in practice.

In summary, we find evidence (i) that a single factor underlies students' CR and MC responses in our data, and (ii) this single factor is able to explain most of the variation in the respective scores.[12] In other words, when we use an empirical procedure similar to what BRW and TWW employ, we are led to the same conclusion. This raises serious doubts about the appropriateness of factor analysis for addressing the question, "Do CR and MC questions measure the same thing?" Our analysis demonstrates that it is possible for this empirical procedure to produce a positive answer to this question, even when the underlying data contain strong, contrary evidence.

Walstad and Becker (1994) is another study that has been very influential in the debate over CR versus MC questions. Their study analyzes AP Microeconomics and Macroeconomics exams. Each of these has CR and MC components from which an overall composite score is formed, with the components receiving weights of two-thirds and one-third, respectively. WB use these data to regress the composite scores on the MC scores. They find that the MC scores explain between 90 and 95 percent of the variation in composite scores. WB conclude that there are "no differences, or only slight differences, in what the two types of tests and questions [multiple-choice and constructed-response] measure."

Conveniently, WB report simple correlations between the CR and MC components of the AP exams. These fall in the same range as the correlations we report for our data in TABLE 2. Thus, it should not be surprising that we are able to produce WB-type regressions that are very similar to theirs.

[12] BRW conclude that one factor explains most of the variation by virtue of a battery of goodness-of-fit measures, finding that the second factor adds little in the way of goodness-of-fit. TWW reach this conclusion by noting that the factor loadings on the second factor are relatively small.

We construct composite scores from the MC and CR components using the same weights as the AP exams. We then estimate WB-type regressions using the same five samples we used for our original analyses. TABLE 7 reports the results. Of interest here are the R^2 from the respective regressions. These range between 85 and 90 percent.[13] Using the same specification, WB obtained an R^2 of 94% for the Microeconomics exams, and an R^2 of 90% for the Macroeconomics exams. Our macro results are about the same as WB's, while our micro results are somewhat lower.

In conclusion, the strongest evidence that CR and MC questions measure the same thing comes from factor analysis and WB-style regressions. The preceding analysis argues that both these approaches are unreliable in the following sense: It is possible for these empirical procedures to produce an affirmative conclusion, even when the underlying data contain strong, contrary evidence.

In placing these studies in perspective, it is useful to recall the "policy question" that motivates them. If it could be shown that CR and MC questions measure the same thing, then instructors could get the same information about learning outcomes using an all-MC format, at lower total cost. Our analysis suggests that CR and MC questions do not measure the same thing. Yet, it could still be the case that an all-MC format is preferable if the extra information provided by CR questions was not sufficient to justify their higher costs.

This highlights two separate, but related research questions: (i) Do CR and MC questions measure the same thing?, and (ii) Are the benefits of CR questions sufficient to compensate their costs? Perhaps WB-style regressions are more appropriate for addressing this second question. If composite scores are near-perfectly predictable from

[13] These results are very similar to those obtained by Krieg and Uyar (2001).

MC scores, this may suggest that the benefits of CR questions are relatively small. However, even this conclusion does not necessarily follow. The slippage occurs in mapping R^2 values to benefits.

As Kennedy and Walstad (1997) point out, it is grades, not R^2 values, which matter to instructors and students. KW use simulation exercises to estimate the effect of moving to an all-MC format for the AP test. They report that the number of students who would receive different AP grades is small but statistically significant. However, alternative simulation assumptions produce larger effects.

Like KW, we conclude that CR and MC questions do not measure the same thing. KW's approach has an advantage over ours in that they relate differences in CR and MC scores to an outcome that can be mapped into a benefit versus cost framework. The unique contribution of our study is that we provide evidence that these differences are related to student achievement.

V. CONCLUSION

Our study empirically investigates the relationship between constructed-response (CR) and multiple-choice (MC) questions using a unique data set compiled from several years of university introductory economics classes. Similar to other studies, we find that MC questions are able to explain, at best, about 50 percent of the variation in CR scores. However, unlike other studies, we are able to show that the corresponding residuals are related to student learning. Specifically, we find that the component of CR scores that cannot be explained by MC responses is positively and significantly related to (i) performance on a subsequent exam in the same course, and (ii) academic performance in other courses.

However, the key issue for instructors considering a switch to an all-MC format is whether CR questions provide information that could not be obtained by expanding the set of MC questions. We exploit the panel nature of our data to construct a quasi-counterfactual experiment. We show that combining one CR and one MC component always predicts student achievement better than combining two MC components.

A final contribution of our study is that we demonstrate that empirical approaches that rely on factor analysis or Walstad-Becker (1994)-type regressions are unreliable in the following sense: It is possible for these empirical procedures to lead to the conclusion that CR and MC questions measure the same thing, even when the underlying data contain strong, contrary evidence.

Further progress on the CR versus MC debate will likely come from more careful analyses of the benefits and costs of these two kinds of questions. Kennedy and Walstad (1997) show one way forward: Their paper models how one can compare grade distributions using alternative test formats. Another possible approach is to compare CR and MC scores on how well they predict future academic success. We hope this study stimulates future research efforts in this direction.

REFERENCES

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.

Bloom, B. S. (1956). *Taxonomy of educational objectives: The classification of educational goals, handbook 1: Cognitive Domain.* New York, McKay.

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.

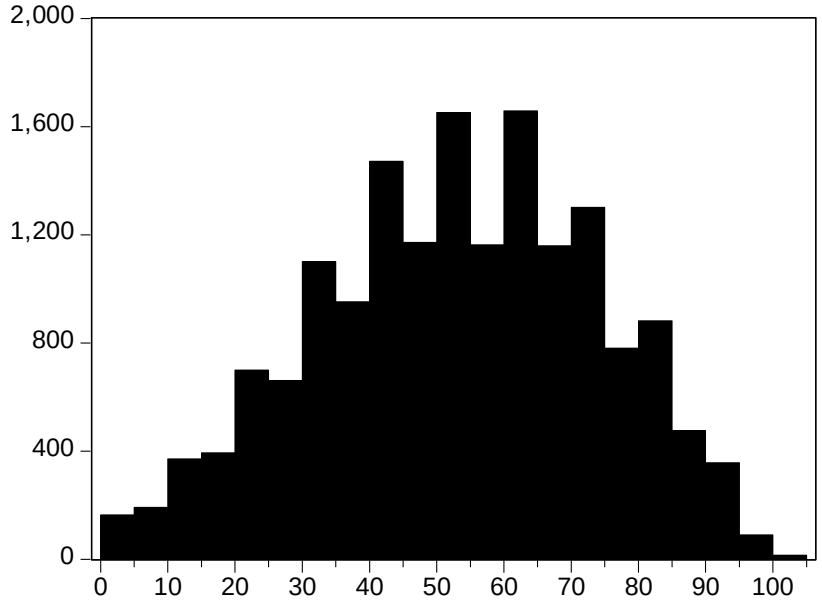
Thissen, D., Wainer, H., & Wang, X. (1994). Are Tests Comprising Both Multiple-Choice and Free-Response Items Necessarily Less Unidimensional Than Multiple-Choice Tests? An Analysis of Two Tests. *Journal of Educational Measurement, 31,* 113-123.

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.

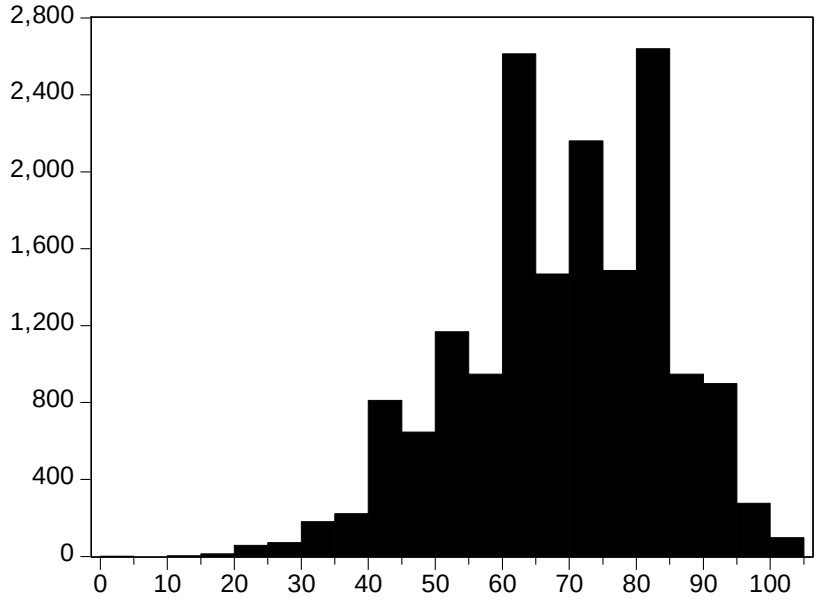
FIGURE 1
Statistical Summary of Multiple-Choice and Constructed-Response Scores

PANEL A: Constructed-Response Scores



Mean	52.53
Minimum	0
Maximum	100
Std. Dev.	20.99
Observations	16710

PANEL B: Multiple-Choice Scores



Mean	68.38
Minimum	0
Maximum	100
Std. Dev.	15.28
Observations	16710

TABLE 1
Statistical Summary of Data

VARIABLE	OBSERVATIONS	MEAN	MINIMUM	MAXIMUM	STD. DEV.
MICRO	16710	0.554	0	1	0.497
TERM_TEST	16710	0.500	0	1	0.500
CONSTRUCTED-RESPONSE (Term Test)	8355	50.0	0	100	20.4
CONSTRUCTED-RESPONSE (Final Exam)	8355	55.0	0	100	21.3
MULTIPLE-CHOICE (Term Test)	8355	66.8	0	100	15.7
MULTIPLE-CHOICE (Final Exam)	8355	69.9	16.7	100	14.7
GPA	16710	3.53	-1	9	2.49
COMPOSITE	16710	63.1	10	100	15.5

TABLE 2
Predicting Constructed-Response Scores Using Multiple-Choice Scores

	SAMPLE				
	Micro/Term Tests (1)	***Micro/Final Exams (2)***	***Macro/Term Tests (3)***	***Macro/Final Exams (4)***	***All Observations (5)***
Constant	-7.4980 (-6.72)	-12.1581 (-11.69)	6.1509 (5.79)	-21.2494 (-18.03)	-6.0626 (-10.69)
Multiple-Choice	0.8097 (50.96)	0.9832 (67.81)	0.7143 (43.55)	1.0608 (67.28)	0.8568 (106.63)
Observations	4628	4628	3727	3727	16710
R^2	0.347	0.470	0.318	0.508	0.389
Simple Correlation	0.589	0.686	0.564	0.713	0.624

NOTE: Values in parentheses are *t*-statistics calculated using heteroscedastic-robust (White) standard errors.

TABLE 3
Predicting Final Exam Performance From Term Test Scores

VARIABLE	*Dep. Variable = Constructed-Response (Final Exam) (1)*	*Dep. Variable = Multiple-Choice (Final Exam) (2)*
Sample (1a): ALL OBSERVATIONS (2002-2006)		
Constant	7.5982 (9.55)	37.3361 (60.72)
Multiple-Choice (Term Test)	0.7152 (63.24)	0.4933 (57.12)
Residual from Term Test Constructed-Response Regression	0.5292 (49.49)	0.3092 (38.97)
R^2	0.468	0.410
Observations	7270	7270
Sample (1b): ALL OBSERVATIONS (2007)		
Constant	-12.2469 (-5.97)	25.8495 (14.34)
Multiple-Choice (Term Test)	0.9591 (33.80)	0.6170 (25.09)
Residual from Term Test Constructed-Response Regression	0.6331 (22.03)	0.2198 (11.80)
R^2	0.579	0.415
Observations	1085	1085
Sample (2a): MICRO (2002-2006)		
Constant	-0.6955 (-0.58)	27.7901 (30.76)
Multiple-Choice (Term Test)	0.7954 (48.93)	0.5879 (48.29)
Residual from Constructed-Response Regression	0.4710 (31.79)	0.2740 (25.20)
R^2	0.459	0.424
Observations	3947	3947

VARIABLE	*Dep. Variable = Constructed-Response (Final Exam) (1)*	*Dep. Variable = Multiple-Choice (Final Exam) (2)*
Sample (2b): MICRO (2007)		
Constant	-12.7999 (-4.93)	23.3048 (11.25)
Multiple-Choice (Term Test)	0.9946 (26.90)	0.6108 (21.07)
Residual from Term Test Constructed-Response Regression	0.6112 (17.21)	0.2547 (11.73)
R^2	0.578	0.454
Observations	681	681
Sample (3a): MACRO (2002-2006)		
Constant	9.6417 (8.77)	40.0442 (46.99)
Multiple-Choice (Term Test)	0.7335 (44.66)	0.5055 (39.99)
Residual from Term Test Constructed-Response Regression	0.5757 (33.66)	0.2808 (22.50)
R^2	0.486	0.404
Observations	3323	3323
Sample (3b): MACRO (2007)		
Constant	-13.8856 (-4.07)	34.2929 (12.53)
Multiple-Choice (Term Test)	0.9375 (20.68)	0.5685 (15.96)
Residual from Term Test Constructed-Response Regression	0.6663 (13.09)	0.3167 (9.80)
R^2	0.581	0.479
Observations	404	404

NOTE: Values in parentheses are *t*-statistics calculated using heteroscedastic-robust (White) standard errors.

TABLE 4
Predicting Student GPAs Using Term Test and Final Exam Scores

	ALL OBSERVATIONS		***MICRO***		***MACRO***	
	2002-2006 (1a)	***2007 (1b)***	***2002-2006 (2a)***	***2007 (2b)***	***2002-2006 (3a)***	***2007 (3b)***
Constant	-3.8018 (-37.80)	-4.7479 (-19.01)	-4.2663 (-31.83)	-4.8550 (-16.72)	-3.666 (-24.65)	-5.4062 (-10.82)
Multiple-Choice (Term Test)	0.0317 (19.92)	0.0301 (5.99)	0.0356 (14.86)	0.0301 (4.74)	0.0363 (15.39)	0.0327 (4.10)
Multiple-Choice (Final Exam)	0.0742 (41.73)	0.0925 (19.63)	0.0752 (30.93)	0.0973 (15.50)	0.0705 (26.46)	0.0943 (11.31)
Residual from Term Test CR Regression	0.0269 (19.92)	0.0188 (5.38)	0.0209 (11.49)	0.0215 (5.15)	0.0321 (15.64)	0.0064 (2.88)
Residual from Final Exam CR Regression	0.0488 (34.76)	0.0498 (13.68)	0.0508 (26.92)	0.0440 (9.14)	0.0464 (22.19)	0.0662 (10.84)
Observations	7270	1085	3947	681	3323	404
R^2	0.569	0.630	0.567	0.629	0.577	0.642
Hypothesis Test (Residuals = 0)	F = 1218.45 (p = 0.000)	F = 195.92 (p = 0.000)	F = 589.44 (p = 0.000)	F = 92.87 (p = 0.000)	F = 616.87 (p = 0.000)	F = 97.96 (p = 0.000)

NOTE: Unless otherwise marked, values in parentheses are *t*-statistics calculated using heteroscedastic-robust (White) standard errors. Column numbers (e.g., 1a) identify the respective sample and are identical to the samples in TABLE 3.

TABLE 5
Predicting Student GPAs: Would an All-Multiple Choice Assessment Be Better?

	ESTIMATED COEFFICIENTS			
	Multiple-Choice (Term Test)	***Multiple-Choice (Final Exam)***	***Constructed-Response (Term Test)***	***Constructed-Response (Final Exam)***
I. SAMPLE (1a)				
A. MC(Term) + [MC(Final) OR CR(Final)]:				
(1) R^2 = 0.424	0.0392 (23.25	0.0811 (46.10)	----	----
(2) R^2 = 0.526	0.0277 (18.44)	----	----	0.0719 (65.76)
B. MC(Final) + [MC(Term) OR CR(Term)]:				
(3) R^2 = 0.424	0.0392 (23.25	0.0811 (46.10)	----	----
(4) R^2 = 0.485	----	0.0634 (35.53)	0.0491 (37.60)	----
II. SAMPLE (1b)				
A. MC(Term) + [MC(Final) OR CR(Final)]:				
(5) R^2 = 0.490	0.0497 (9.75)	0.0864 (18.01)	----	----
(6) R^2 = 0.593	(0.0328 (7.15)	----	----	0.0732 (25.73)
B. MC(Final) + [MC(Term) OR CR(Term)]:				
(7) R^2 = 0.490	0.0497 (9.75)	0.0864 (18.01)	----	----
(8) R^2 = 0.554	----	0.0764 (17.96)	0.0506 (16.13)	----

	ESTIMATED COEFFICIENTS			
	Multiple-Choice (Term Test)	***Multiple-Choice (Final Exam)***	***Constructed-Response (Term Test)***	***Constructed-Response (Final Exam)***
III. SAMPLE (2a)				
A. MC(Term) + [MC(Final) OR CR(Final)]:				
(9) $R^2 = 0.434$	0.0472 (18.75)	0.0753 (30.18)	----	----
(10) $R^2 = 0.534$	0.0364 (16.42)	----	----	0.0693 (44.23)
B. MC(Final) + [MC(Term) OR CR(Term)]:				
(11) $R^2 = 0.434$	0.0472 (18.75)	0.0753 (30.18)	----	----
(12) $R^2 = 0.468$	----	0.0671 (24.54)	0.0444 (24.54)	----
IV. SAMPLE (2b)				
A. MC(Term) + [MC(Final) OR CR(Final)]:				
(13) $R^2 = 0.515$	0.0406 (6.39)	0.0989 (16.13)	----	----
(14) $R^2 = 0.587$	0.0291 (4.89)	----	----	0.0723 (19.40)
B. MC(Final) + [MC(Term) OR CR(Term)]:				
(15) $R^2 = 0.515$	0.0406 (6.39)	0.0989 (16.13)	----	----
(16) $R^2 = 0.572$	----	0.0834 (15.06)	0.0450 (11.45)	----

	ESTIMATED COEFFICIENTS			
	Multiple-Choice (Term Test)	***Multiple-Choice (Final Exam)***	***Constructed-Response (Term Test)***	***Constructed-Response (Final Exam)***
V. SAMPLE (3a)				
A. MC(Term) + [MC(Final) OR CR(Final)]:				
(17) $R^2 = 0.421$	0.0410 (16.08)	0.0792 (29.40)	----	----
(18) $R^2 = 0.533$	0.0296 (13.43)	----	----	0.0702 (45.38)
B. MC(Final) + [MC(Term) OR CR(Term)]:				
(19) $R^2 = 0.421$	0.0410 (16.08)	0.0792 (29.40)	----	----
(20) $R^2 = 0.508$	----	0.0593 (22.79)	0.0550 (29.58)	----
VI. SAMPLE (3b)				
A. MC(Term) + [MC(Final) OR CR(Final)]:				
(21) $R^2 = 0.473$	0.0563 (6.83)	0.0855 (9.84)	----	----
(22) $R^2 = 0.630$	0.0289 (3.95)	----	----	0.0811 (18.22)
B. MC(Final) + [MC(Term) OR CR(Term)]:				
(23) $R^2 = 0.473$	0.0563 (6.83)	0.0855 (9.84)	----	----
(24) $R^2 = 0.522$	----	0.0672 (7.42)	0.0600 (9.67)	----

NOTE: Values in parentheses are *t*-statistics calculated using heteroscedastic-robust (White) standard errors. Sample numbers (e.g., 1a) identify the respective sample and are identical to the samples in TABLES 3 and 4.

TABLE 6
Summary of Principal Component Analyses

Sample (1): All Observations

Principal Component	*Eigenvalue*	*Proportion*
1	1.6236	0.812
2	0.3764	0.188

Sample (2): Micro/Term Tests

Principal Component	*Eigenvalue*	*Proportion*
1	1.5846	0.792
2	0.4154	0.208

Sample (3): Micro/Final Exams

Principal Component	*Eigenvalue*	*Proportion*
1	1.6855	0.843
2	0.3145	0.157

Sample (4): Macro/Term Tests

Principal Component	*Eigenvalue*	*Proportion*
1	1.5636	0.782
2	0.4364	0.218

Sample (5): Macro/Final Exams

Principal Component	*Eigenvalue*	*Proportion*
1	1.7129	0.856
2	0.2871	0.144

NOTE: Samples are identical to the samples in TABLE 2.

TABLE 7
Summary of Regressions Based on Walstad and Becker's (1994) Specification

	SAMPLE				
	Micro/Term Tests (1)	***Micro/Final Exams (2)***	***Macro/Term Tests (3)***	***Macro/Final Exams (4)***	***All Observations (5)***
Constant	-2.4999 (-6.72)	-4.0527 (-11.69)	2.0503 (5.79)	-7.0831 (-18.03)	-2.0209 (-10.69)
Multiple-Choice	0.9366 (176.85)	0.9944 (205.76)	0.9048 (165.51)	1.0203 (194.11)	0.9522 (355.55)
Observations	4628	4628	3727	3727	16710
R^2	0.862	0.891	0.871	0.896	0.876

NOTE: The dependent variable is a composite assessment score created by weighting the multiple-choice and constructed-response components by 2/3 and 1/2, respectively. These are the weights used by the Advanced Placement Economics test that was analysed by Walstad and Becker (1994). Samples are identical to the samples in TABLE 2.