More Evidence on the Use of Constructed-Response Questions in Principles of Economics Classes

Stephen Hickson and Bob Reed

Abstract

This study provides evidence that constructed response (CR) questions contribute information about student knowledge and understanding that is not contained in multiple choice questions (MC). We use an extensive data set of individual assessment results from Introductory Macro- and Microeconomics classes at a large, public university. We find that (i) CR scores contain information not contained in MC questions, (ii) this information is correlated with a measure of student knowledge and understanding of course material, and (iii) CR questions are better able to 'explain' academic achievement in other courses than additional MC questions. There is some evidence to suggest that this greater explanatory power has to do with the ability of CR questions to measure higher-level learning as characterised by Bloom's taxonomy (Bloom, 1956). Both (i) the generalisability of our results to other principles of economics classes, and (ii) the practical significance (in terms of students' grades) of our findings, remain to be determined.

JEL classification: A22

> '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)

1. Introduction

University principles of economics courses often have enrolments 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.[1] On the other hand, CR questions are much more costly to grade. In addition, the marking of CR questions is less reliable due to the subjective nature of

[1] Multiple choice (MC) questions present students a set of answers and ask them to select the correct one(s). Constructed response (CR) questions require students to provide their own answers. These can range from fill-in-the-blank questions; to definitional or short-answer questions; to questions requiring mathematical solutions; to long essay questions.

the questions. The tradeoff is a very real one to university instructors facing declining budget environments where marking assistance on CR questions may be reduced or eliminated.

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'.[2] 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 affirmatively, 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, p. 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 Advanced Placement (AP) composite scores on MC scores. Kennedy and Walstad (1997) simulate grade distributions using different test formats. Becker and Johnston (1999) utilise two-stage least squares regression. 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 the others. Further, most of this research has focused on AP exams. These results may not be valid for principles of economics classes taught at universities.

Our study takes yet another approach to the MC-CR debate. We use extensive data from principles classes in macroeconomics and microeconomics from a large public university where assessments consist of both MC and CR questions. Our empirical methodology is targeted to an instructor trying to decide whether to use a composite MC-CR assessment, versus an assessment composed of all MC questions. It consists of three steps.

First, we investigate the degree to which 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'. If that were the case, there would be no reason to use the more costly CR questions, and our hypothetical instructor would be better off using an all-MC assessment. Our empirical analysis does not support this view. We find that the regression of CR scores on MC scores leaves a substantial residual.

The next step consists of determining whether the residual from the CR regressions represents noise, versus information relevant to student knowledge and understanding of course material. To address this question, we use MC data and the CR-residual from the term test to determine whether the CR data can help predict student performance on the final exam. If the residual variable were insignificant, that would suggest that the CR-residual was just noise. In contrast, we find that the CR-residual is large

[2] 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 microeconomics and macroeconomics exams from switching to an all-MC format. Unfortunately, in order to categorise 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.

in size and statistically significant. Since the residual represents the component of CR scores that cannot be explained by MC scores, and since it is significantly correlated with final exam performance, we conclude that CR questions contain information about student knowledge and understanding that is not contained in the original set of MC questions.

It is possible that the information provided by the CR-residual supplies the same information that could have been provided by additional MC questions. In other words, our results to this point are not able to help our hypothetical instructor decide whether to use CR questions or additional MC questions. To address this question, the third step constructs a pseudo-counterfactual experiment. We use MC and CR data from the midterm and the final exam to measure whether the CR component of a test provides more information than additional MC questions in explaining students' GPAs in other courses. In each of our 12 sub-samples, the CR component provides substantially more explanatory power than additional MC questions. This suggests that CR questions contain useful information beyond MC questions that may be helpful in assessing students' learning.

The second half of our study investigates 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. Finally, we recognise that our results reflect the nature and quality of our questions. Therefore, we describe the makeup of the respective MC and CR questions that were used in our research. We conclude with caveats regarding the interpretation and application of our research, and recommendations for future research.

2. Data

Our analysis uses data compiled over a six-year period (2002–07) 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 was 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 was maintained by the fact that the same two instructors taught the classes, and wrote and graded the assessments across the whole time period. The Kruder-Richardson-20 statistics for the MC sections are consistently around 0.7. This indicates a good level of reliability for testing instruments that are measuring multiple dimensions, constructs or areas of interest (Nunnally, 1978).

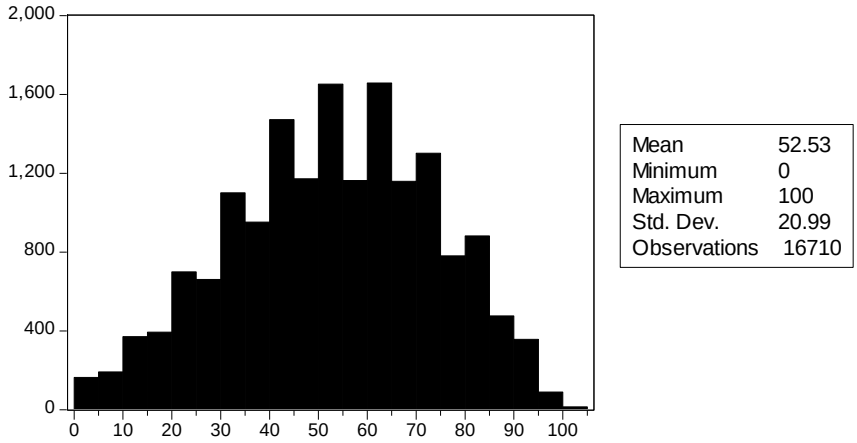
All together, the data set includes assessments from 10 separate offerings of Introductory Microeconomics and eight of Introductory Macroeconomics, for a total of 36 assessments (18 term tests plus 18 final exams). When we eliminate incomplete records and students for whom one of the assessments is missing, we are left with 16,710 observations.[3] By way of comparison, Walstad and

[3] 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.

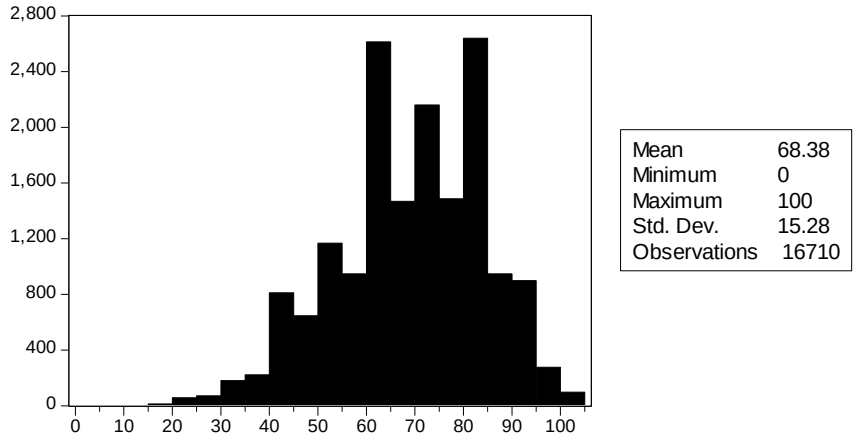
Becker (1994) have a total of 8842 observations and Becker and Johnston (1999) have 4178. Most studies have far fewer.[4]

Figure 1

PANEL A: Constructed-response scores



PANEL B: Multiple-choice Scores



There are two features which make our data set unique. First, we have repeated observations – i.e. both a midterm test and a final exam – for the same student for a given course. Second, we have data about the student's achievement in other courses. We exploit both of these features in our analysis.

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 characterised 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

[4] For example, Krieg and Uyar (2003) have only 223 observations.

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 estimate the key regressions using the more statistically appropriate fractional logit procedure. The results are virtually identical.[5]

Table 1 provides a statistical summary of the students represented in our study. Approximately 55% of the sample derive from Introductory Microeconomics 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 twice the allotted time as the term test, it is designed to require 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 both ECON 104 and 105.[6] 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.

Table 1: Statistical summary of data

Variable	Observations	Mean	Minimum	Maximum	Std. dev.
Micro	16,710	0.554	0	1	0.497
Term test	16,710	0.500	0	1	0.500
Constructed-response (Term Test)	8,355	50.0	0	100	20.4
Constructed-response (Final Exam)	8,355	55.0	0	100	21.3
Multiple-choice (Term Test)	8,355	66.8	0	100	15.7
Multiple-choice (Final Exam)	8,355	69.9	16.7	100	14.7
GPA	16,710	3.53	–1	9	2.49
Composite	16,710	63.1	10	100	15.5

While not reported in Table 1, approximately 56% 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% of the students are Asian. This high percentage is due to a surge in Asian enrollments that occurred in the early 2000s 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% of our sample. With respect to language, a little more than 60% of the sample declared English to be their first language. The great majority of the remainder identified with Chinese.

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

[6] 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.

3. Results

Step one

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 1, this would clearly indicate that CR scores added little information to that already provided by the student's MC performance. In this case, our hypothetical instructor would be better off discarding the CR component and using an all-MC assessment.

Table 2 summarises 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 percentage point on the MC component predicts an additional 0.7 to 1.1 percentage points on the CR component, depending on the sample.

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.

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%. Those for the term tests are even lower, in the low- to mid-30s.[7] (We discuss this difference between term tests and 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%.[8]

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, p. 194) report simple correlations of 0.69 and

[7] 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).

[8] 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.

0.64 for the Micro and Macro AP tests. Lumsden and Scott (1987, p. 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 these measure aspects of students' knowledge and understanding that are not measured by MC questions? Or are the two question types assessing the same thing(s) but with noise/measurement error?

Step two

If we had an alternative measure of student knowledge and understanding, we could take the residuals from the regressions in Table 2 and test if they were significant predictors of this alternative measure. If the residuals were unrelated to student knowledge and understanding, say were pure measurement error, then one would expect them to be uncorrelated 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 information about student knowledge and understanding that was not captured by MC responses.

Unfortunately, we do not have an alternative measure of student knowledge and understanding 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).[9] We separate the 2002–06 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.[10]

Our results are evidence that CR scores contain information not contained in the existing MC scores, and that this information is correlated with student academic performance. But is this 'information' really related to students' knowledge and understanding of course material? 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.

[9] 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.). Note that we would get the same coefficient and *t*-statistics for this variable if we substituted the actual CR (Term Test) variable for the associated residual (see Johnston and DiNardo, 1997, p. 82). We use the residual variable to emphasise that this variable contains information that is independent of the information contained in the MC (Term Test) variable.

[10] At the suggestion of a referee who was concerned that our results might be an artifact of a given year's type of assessment or demographic composition of test-takers, we re-estimated the regressions in Table 3, breaking out the observations by year and subject area (Macroeconomics, Microeconomics). The residual from the CR term test regression remained a significant determinant of final exam performance in every case (a total of 22 regressions). The results are available from the authors.

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 student knowledge and understanding, 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?

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	0.4424
Observations	3947	3947
Sample (2b): Micro (2007)		
Constant	–12.7999 (–4.93)	23.3048 (11.25)

Variable	Dep. Variable = Constructed-Response (Final Exam) (1)	Dep. Variable = Multiple-Choice (Final Exam) (2)
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.

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. 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 (cf. Table 1).

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).[11] 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

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

3). Given the greater time constraints, we hypothesise 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 (Application and Analysis) correct. 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. It will also cause the MC responses on the final exam to measure higher levels of knowledge and understanding 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.

The fact that (i) the CR-residual is a significant determinant of MC scores on the final exam, and (ii) the MC variable explains a smaller amount of variation in CR scores on term tests compared to final exams, is consistent with the hypothesis that the CR-residual measures higher-level learning according to Bloom's taxonomy (Bloom, 1956).

Step three

Summarising the above, our results suggest that CR scores contain information not contained in the responses to *existing* MC questions. However, we are still not in a position to help our hypothetical instructor decide whether to use CR questions or additional MC questions: perhaps the additional information provided by the CR questions is merely a substitute for information that could have been provided by including more MC questions. To address this concern, we would like to compare assessments using composite MC/CR questions with those using all-MC questions.

We could empirically address this if we were able to perform the following experiment. Suppose there were two groups of identical students. One group was given a composite test composed of MC and CR questions. Call these variables *MC1* and *CR*. The other group was given a test composed entirely of MC questions, where the first half of the questions was identical to what the first group received. Call these two sets of MC questions, *MC1* and *MC2*. Finally, suppose we had some objective measure of a student's knowledge and understanding of course material. Call this variable *Y*.

Now consider two regression models:

$$Y_i = \beta_0 + \beta_1 MC1_i + \beta_2 MC2_i + \varepsilon_i \quad (1)$$

$$Y_i = \alpha_0 + \alpha_1 MC1_i + \alpha_2 CR_i + v_i \quad (2)$$

If the CR questions contained the same 'information' as the additional MC questions, then the specification of Equation (2) should have approximately the same explanatory power as the specification of Equation (1). Alternatively, if the R^2 value for Equation (2) was smaller than that for Equation (1), that would suggest that the CR questions were less efficient at 'explaining' students' understanding than the additional MC questions. If the R^2 value for Equation (2) were greater, that would indicate that the CR questions contained information that had greater explanatory power than the additional MC questions.

We could get to this conclusion because we have a counterfactual to compare our composite test results with: One group takes a test composed entirely of MC questions. The other group takes a

composite test composed of both MC and CR questions. Unfortunately, our data does not contain a real counterfactual. Instead, we manipulate our data to create a pseudo-counterfactual.

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. As discussed above, we use this information to calculate a GPA value based on their performance in non-introductory economics classes. We use this GPA variable to proxy for *Y* in the experiment described by Equations (1) and (2) and the subsequent discussion. Our working assumption is that GPA in non-economics classes is positively correlated with students' knowledge and understanding of course material in their economics principles class. Our rationale is that students who have a good understanding of course material in one class are also likely to get high grades in their other classes (because better students are more likely to have good knowledge and understanding in all their classes).

For each student in a given principles of economics course, we also have their MC score on the (i) term test and (ii) final exam in that course; and their CR score on the (iii) term test and (iv) final exam. We divide our observations into the same six samples that we used in Table 3. With reference to Equations (1) and (2) above and the corresponding discussion, let *MC1* be the MC component on the term test, and let *MC2* and *CR* be the MC and CR scores from the final exam. Note that *MC2* and *CR* should be from the same assessment to make the comparison as clean as possible.

Using the same logic as above, if the R^2 values are higher from the equations with the *CR* component, that suggests that the CR responses contain more/better information than the *MC2* responses – and not just the same information. Accordingly, we compare the following regression models:

(i) $\text{GPA}_t = \beta_0 + \beta_1 \text{MC(Term)}_t + \beta_2 \text{MC(Final)}_t + \varepsilon_t$, and

(ii) $\text{GPA}_t = \alpha_0 + \alpha_1 \text{MC(Term)}_t + \alpha_2 \text{CR(Final)}_t + \eta_t$.

To recapitulate, the pair of models above proxies for the following thought experiment: Suppose an instructor had given an all-MC term test. Would he or she more effectively assess academic achievement if the final exam consisted of all MC questions, or a mix of CR and MC questions? Specification (i) represents the case where assessment is based solely on MC questions. Specification (ii) represents a composite CR/MC assessment. If MC and CR questions measure the same thing(s), a comparison of the R^2 values from estimating models (i) and (ii) across different samples should show no clear pattern. However, if CR questions measure information not captured by the additional MC questions – such as higher levels of the Bloom (1985) taxonomy – then the R^2 values from Specification (ii) regressions should be consistently higher.

As a further test, we also compare an alternative pair of regression models:

(iii) $\text{GPA}_t = \beta_0 + \beta_1 \text{MC(Final)}_t + \beta_2 \text{MC(Term)}_t + \varepsilon_t$, and

(iv) $\text{GPA}_t = \alpha_0 + \alpha_1 \text{MC(Final)}_t + \alpha_2 \text{CR(Term)}_t + \eta_t$.

Table 4 reports the results of this test. We divide the data into the same six samples used for Table 3. Consider the first two rows of Table 4. For the sample of all observations from 2002–06 (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.

Table 4: 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)	----

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. Samples (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)	----
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)	----

Taken together, the results from Tables 2 through 4 provide evidence that CR questions measure student knowledge and understanding that is not captured by MC questions. Our evidence is consistent with the hypothesis that the CR variable measures higher-level learning, as defined by Bloom's taxonomy (1956). While other studies, such as Kennedy and Walstad (1997) and Becker and Johnston (1999), provide evidence that CR and MC responses are 'different', our study is the first to link these differences to student academic performance in university principles of economics classes.

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.

4. Relating our findings to those of previous studies

Our finding that CR scores comprise information not contained in MC scores 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 Advanced Placement (AP) examination in Computer Science. TWW re-analyse 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.[12]

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 analysed 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 five free-response questions. The AP exam in Chemistry consists of 75 MC questions and four sections of free-response questions, some of which contain multiple problems. BRW and TWW break up the respective components into multiple 'parcels'. BRW re-organise the 50 MC questions into five sets ('parcels') of 10 questions each. TWW convert the original 75 MC questions into 15, 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. Unfortunately, our data contain fewer questions than BRW and TWW and are thus less amenable to 'parcelisation'. Instead, we apply principal component analysis (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.[13] 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.

[12] 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.

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

Table 5 reports the results of applying PCA to the same five samples we previously analysed 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 5 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 1. The second of these plots the eigenvalues in decreasing order and selects all factors immediately preceding an abrupt levelling 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 5. 'Proportion' translates these eigenvalues into shares of total variation in the correlation matrix. These range from 78–85% across the different samples.

Table 5: 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.

In summary, we find evidence that (i) 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.[14] In

[14] 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.

other words, when we use an empirical procedure similar to what BRW and TWW employ, we are led to the same conclusion that they reach.

Walstad and Becker (1994) is another study that has been very influential in the debate over CR versus MC questions. Their study analyses 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% 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'.

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 6 reports the results. Of interest here are the R^2 from the respective regressions. These range between 85 and 90%.[15] 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.[16]

Table 6: 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.

In summary, the strongest evidence that CR and MC questions measure the same thing comes from factor analysis and WB-style regressions. When we replicate these procedures using our data, we get results similar to the original authors. What can we learn from this? It means that one can get different conclusions from the same data, if one uses different methodologies. We argue that our methodology is more directly applicable for the instructor who is trying to decide whether to use a composite MC-CR assessment, versus an assessment composed of all MC questions.

On the other hand, our results are consistent with two studies that have been influential on the other side. Kennedy and Walstad (1997) 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

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

[16] 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.

grades is small but statistically significant. Further, alternative simulation assumptions produce larger effects.

Becker and Johnston (1999) examine results from the Victorian (Australia) Certificate of Education assessment of high school economics. The VCE assessment consists of both MC and CR components. Like previous studies before them, BJ find a high correlation between MC and CR scores. However, when they instrument the explanatory variable with school-wide performance on that component, they find the correlation becomes small in size and statistically insignificant. They therefore conclude that the MC and CR components measure different dimensions of knowledge.

The KW and BJ studies are complements to ours. Both find differences in what MC and CR responses measure. The unique contribution of our study is that we provide evidence that these differences are related to student academic achievement.

5. A closer look at the CR and MC questions analysed in this study

The debate over CR versus MC questions is to some extent an idiosyncratic one that is course- and instructor-dependent. 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 analysed 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. Despite the best of intentions, CR questions may only be testing recall and recognition. A key issue is whether the student could have memorised the answer in advance.

We next describe the nature of the MC and CR questions used in the assessments included in our data set.[17] 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.*

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).

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

A following 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 follows up 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 must 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 analyse 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.

6. Conclusion

This study provides evidence that constructed response (CR) questions contribute information about student knowledge and understanding that is not contained in multiple choice questions (MC). This finding may be useful to university instructors of principles of economics classes trying to decide whether to use constructed response (CR) questions on assessments, with their higher marking costs; or to employ all multiple choice (MC) questions.

To address this issue, our study empirically investigates the relationship between CR and MC questions using a 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% of the variation in CR scores. However, unlike other studies, we are able to provide evidence that the corresponding residuals are related to student knowledge and understanding. Specifically, we find that the component of CR scores that cannot be explained by MC responses is positively and significantly related to performance on a subsequent exam in the same course.

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 lead to different conclusions about the relationship between CR and MC questions when applied to our data. We argue that our methodology is more directly applicable for an instructor trying to decide between a composite MC/CR assessment, and an assessment composed entirely of MC questions.

We have two sets of cautions with regard to interpretation and application of our results. First, although this study employs a large number of observations, these all come from two courses at a single university. It is difficult to determine how generalisable these results may be. While we have attempted to give the reader an understanding of the type of MC and CR questions used in these courses, the most direct way to establish external validity is to replicate our methodology using data from other principles of economics classes.

Second, while this study presents evidence that CR questions contain information not contained in MC questions, it does not address the practical importance of this additional information. For example, if the use of CR questions resulted in only a small modification of students' grades, then our hypothetical instructor trying to decide between a MC/CR and an all-MC assessment might well choose the latter. The only study that has attempted to measure the effects of switching from a composite to an all-MC test is Kennedy and Walstad (1997). Their study focuses on AP test results, and they found little difference in outcomes between MC/CR and all-MC assessments. There are no studies that attempt to do the same for university classes. We hope this study will stimulate further work on this topic.

References

Becker, W. E. and Johnston, C. (1999). 'The Relationship between Multiple Choice and Essay Response Questions in Assessing Economics Understanding', *Economic Record,* vol. 75*,* pp. 348–57.

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

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

Buckles, S. and Siegfried, J. J. (2006). 'Using Multiple-Choice Questions to Evaluate In-Depth Learning of Economics', *Journal of Economic Education,* vol. 37, pp. 48–57.

Iz, H. B. and Fok, H. S. (2007). 'Use of Bloom's Taxonomic Complexity in Online Multiple Choice Tests in Geomatics Education', *Survey Review,* vol. 39(305), 226–37.

Johnston, J. and DiNardo, J. (1997). *Econometric Methods,*. New York: The McGraw-Hill Companies, Inc.

Kennedy, P. E. and Walstad, W. B. (1997). 'Combining Multiple-Choice and Constructed Response Test Scores: An Economists View', *Applied Measurement in Education,* vol. 10(4)*,* pp. 359–75.

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

Lumsden, K.G. and Scott, A. (1987). 'The Economics Student Reexamined: Male-Female Differences in Comprehension', *Journal of Economic Education*, vol. 18(4), 365–75.

Nunnally, J. (1978) *Psychometric Theory*, New York: McGraw-Hill.

Thissen, D., Wainer, H. and 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,* vol. 31, pp. 113–23.

Wainer, H. and Thissen, D. (1993). 'Combining multiple-choice and constructed response test scores: Towards a Marxist theory of test construction', *Applied Measurement in Education,* vol. 6, pp. 103–18.

Walstad, W. (2006). 'Testing for Depth of Understanding in Economics Using Essay Questions', *Journal of Economic Education,* vol. 37, pp. 38–47.

Walstad W. and Becker, W. E. (1994). 'Achievement Differences on Multiple-Choice and Essay Tests in Economics', *American Economic Review,* vol. 84, 193–96.

Zheng, A. Y., Lawthorn, J. K.,Lumley, T. and Freeman, S. (2008). 'Application of Bloom's Taxonomy Debunks the "MCAT Myth"', *Science,* vol. 319, pp. 414-15.

Author Biography

W. Robert Reed is Professor of Economics at the University of Canterbury. His previous appointments were at the University of Oklahoma and Texas A&M University. His main areas of research are public economics and applied econometrics, with a particular research interest in taxes and economic growth. He has published in the Journal of Political Economy, the Journal of Public Economics, the Journal of Labour Economics, and elsewhere. This is his first foray into economics education.

Stephen Hickson is a Teaching Fellow at the University of Canterbury (UC). He has been at UC full time since 2003. Prior to that appointment he worked full-time for Statistics New Zealand and part-time for UC. His main areas of research are in economics education particularly assessment. Stephen also teaches on the University of Canterbury MBA programme.

Contact details

W. Robert Reed and Stephen Hickson
University of Canterbury
Private Bag 4800
Christchurch 8140
Tel: +64 3 366 7001
Fax: +64 3 364 2635
Email: bob.reed@canterbury.ac.nz and stephen.hickson@canterbury.ac.nz