University of Nebraska - Lincoln
DigitalCommons@University of Nebraska - Lincoln

CBA Faculty Publications | Business, College of

May 1994

Achievement Differences on Multiple-Choice and Essay Tests in Economics

WILLIAM WALSTAD
University of Nebraska-Lincoln, wwalstad1@unl.edu

William Becker
University of Nebraska - Lincoln

Follow this and additional works at: http://digitalcommons.unl.edu/cbafacpub

 Part of the Business Commons

WALSTAD, WILLIAM and Becker, William, "Achievement Differences on Multiple-Choice and Essay Tests in Economics" (1994). *CBA Faculty Publications*. 34.
http://digitalcommons.unl.edu/cbafacpub/34

This Article is brought to you for free and open access by the Business, College of at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in CBA Faculty Publications by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.

American Economic Review 84:2 (May 1994), pp. 193-196. Copyright © 1994 American Economic Association. Used by permission.

RESEARCH ON ECONOMICS EDUCATION[†]

Achievement Differences on Multiple-Choice and Essay Tests in Economics

By WILLIAM B. WALSTAD AND WILLIAM E. BECKER*

Multiple-choice and essay tests are the typical test formats used to measure student understanding of economics in college courses. Each type has its features. A multiple-choice (or fixed-response) format allows for a wider sampling of the content because more questions can be given in a testing period. Multiple-choice tests also offer greater efficiency and reliability in scoring than an essay. The major disadvantage of a multiple-choice item is that the fixed responses tend to emphasize recall and encourage guessing. In an essay (or constructed-response) test, students generate responses that have the potential to show originality and a greater depth of understanding of the topic. The essay also provides a written record for assessing the thought processes of the student.[1]

Despite the claimed differences for each format, little empirical work exists to support the suppositions. If a multiple-choice and an essay test that cover the same material measure the same economic understanding, then the multiple-choice test would be the preferred method for assessment because it is less costly to score and is a more reliable measure of achievement in a limited testing period.[2] If, however, an essay test measures unique aspects of economic understanding, then the extra examinee time and substantial scoring costs may be justified.

The research evidence from other subjects suggests that there is little difference in the knowledge, skills, or abilities measured by multiple-choice and essay (or constructed-response) tests. A study of Advanced Placement (AP) tests in seven college subjects (calculus, computer science, chemistry, biology, history, French, and music) concluded that "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 a subjectively scored section for which this is not true" (Howard Wainer and David Thissen, 1993 p. 116). Similarly, an investigation of the AP exam in computer science found "little support for the stereotype of multiple-choice and free-response formats as measuring substantially different constructs (i.e., trivial factual recognition vs. higher order processes)" (Randy Bennett et al., 1991 p. 89). A review of studies in four domains (writing, word knowledge, reading, and quantitative) was more equivocal about the value of constructed response but concluded that "if differences do exist for any domain, they are very likely to be small" (Ross Traub, 1993 p. 38). Finally, a study

[†]*Discussants:* William Greene, New York University; Jane Lillydahl, University of Colorado.

*Department of Economics, University of Nebraska, Lincoln, NE 68588, and Department of Economics, Indiana University, Bloomington, IN 47405, respectively. We appreciated the help from Rick Morgan, Walter MacDonald, and Gary Marco at ETS in providing AP data and test information. All responsibility, however, for errors or omissions in this study rests with the authors.

[1]See Chapters 15 and 16 in Phillip Saunders and Walstad (1990) for further discussion of the differences and uses in economics.

[2]This may explain why the SAT and GRE are multiple-choice tests.

predicting the GPA of first-year college students found that "the essay added essentially nothing" to what was predicted from high school GPA, SAT scores, and a multiple-choice test of writing skills (Brent Bridgeman, 1991 p. 319).

I. AP Economics Exams

To investigate the relative value of achievement information from multiple choice and essay tests in economics, data were obtained from the 1991 College Board's AP exams in microeconomics and macroeconomics. Each exam had a multiple-choice and an essay section. The multiple-choice section consisted of 50 items to be answered in 60 minutes. The essay section of each exam gave students 30 minutes to write a response to one essay question (macro exam) or two essay questions (micro exam). These AP exams were specifically designed to cover economic content similar to that found in college Principles courses. The exams were prepared and evaluated by test developers at the Educational Testing Service (ETS) and a national committee of seven academic economists. High-ability students generally took one or both of these exams during the final month of a high-school course in the college principles of economics.

Grades for an AP exam range from 1 to 5 (1 = no recommendation; 2 = possibly qualified; 3 = qualified; 4 = well qualified; and 5 = extremely well qualified). The AP grade was awarded using information from each test section. The multiple-choice score was calculated by summing the number right and then correcting for guessing (subtracting one-fourth of the number wrong). Essays were scored in a single reading on a 0–9 scale using a well-defined scoring rubric. About 30 experienced college economics professors and high-school economics teachers were hired to travel to a common site and score the essays. The composite score was determined by giving a two-thirds weight to multiple choice and a one-third weight to the essays. This composite score was then transformed to discrete 1–5 grades based on ranges set by the Chief Reader and ETS staff.[3]

TABLE 1—AP SCORES AND GRADES

Score/ grade	Mean	SD	Correlations MC	E	C
Micro (n = 3,966):					
MC	26.85	11.13			
E	8.49	3.64	0.69		
C	46.38	18.07	0.97	0.84	
G	2.99	1.28	0.95	0.81	0.98
Macro (n = 4,876):					
MC	27.09	10.41			
E	4.03	2.16	0.64		
C	45.90	18.08	0.95	0.85	
G	3.03	1.25	0.93	0.82	0.97

Notes: MC = multiple choice; E = essay; C = composite score; G = AP grade. The last three columns report r correlations for scores and rho correlations for AP grades

II. Analysis

Data for this analysis were obtained from ETS on 3,966 (87 percent) of the 4,539 students who took the micro exam and 4,876 (89 percent) of the 5,476 students who took the macro exam. The sample means and standard deviations for the multiple choice, essay, composite scores, and AP grades are found in Table 1.[4] Also shown in Table 1 are the Pearson correlations among the scores that were used to determine the AP grades. The coefficients of correlation for the multiple-choice and essay scores were relatively high (0.69 micro and 0.65 macro), but far from unity. Although these correlations suggest that multiple-choice and essay

[3]For test information and sample questions, see ETS (1992) or College Entrance Examination Board (1992). Slight changes were made to the AP economics exams in 1993 that increased the exam time by 30 minutes and the number of questions in each section, but weights for determining the composite score remained the same (0.67 multiple choice; 0.33 essay).

[4]There was no statistically significant or practical difference in our sample means and the total-group means reported by ETS (Carole Bleistein et al., 1991).

questions are not measuring the same construct, it is also possible that measurement errors of various types attenuated the correlations.[5]

The observed correlations between multiple-choice and constructed-response sections of AP exams tend to be higher in more quantitative subjects and foreign languages, most likely because of less measurement error. In 1991, for example, the correlations were 0.86 in chemistry, 0.83 in calculus, 0.80 in biology, 0.80 in German, and 0.73 in Spanish, but they were only 0.55 in U.S. history, 0.42 in U.S. government, and 0.49 in English. The average correlation across the 26 AP exams given in 1991 was 0.69, which is about the same as in economics.[6]

Unlike the correlations between the multiple-choice scores and the essay scores, the correlations between the multiple-choice scores and the composite scores were almost unity (0.97 micro and 0.95 macro). The linear relationships between the multiple-choice and composite scores are evident in the regression of the multiple-choice scores on the composite scores:

$$\underset{}{\text{Micro} =} \underset{[0.182]}{4.114} + \underset{[0.006]}{1.574}\ (\text{MC score})$$

($R^2 = 0.94$, $n = 3{,}966$); and

$$\underset{}{\text{Macro} =} \underset{[0.182]}{1.2129} + \underset{[0.006]}{1.650}\ (\text{MC score})$$

($R^2 = 0.90$, $n = 4{,}876$). (Numbers in brackets are standard errors.) The multiple-choice score explained at least 90 percent of the variability in the composite score in each exam. These results suggest that the one-third contribution of the essay to the composite score adds minimal information about the student for determining the AP grade beyond what is already contained in the multiple-choice score. Including covariates for sex, race, ethnic origin, and grades on the AP English literature exam added no statistical significant explanatory power to the composite-score regressions. Grades on the AP calculus exam had a statistically significant effect, but the change in the adjusted R^2 was trivial ($+0.001$) in each equation. The primary determinant of the composite score is the multiple choice score.

A related question to the regression analysis is how well the multiple-choice test score predicts the discrete AP grade instead of the continuous composite score. The results from estimating micro and macro AP grades based on the multiple-choice scores for each test using the ordered-probit procedures in LIMDEP are reported in Table 2. The multiple-choice score correctly predicted 78 percent of the micro grades and 72 percent of the macro grades, where the predicted grade had the highest estimated probability at each multiple-choice score. Although these results are impressive, they are less than perfect and depend on the weighting that the multiple-choice test received in the determination of the composite score. The selection of cutoff scores for grading also introduced variability.[7] Including other background variables in the ordered-probit equations did not change the results.[8]

[5]A true score correlation is estimated by psychometricians by dividing the observed correlation by the square roots of estimates of the test reliabilities. Multiple-choice reliabilities were 0.89 (micro) and 0.87 (macro). Essay reliabilities ranged from 0.51 to 0.75 (micro) and from 0.49 to 0.67 (macro). The correlations corrected for attenuation ranged from 0.84 to 1.0 (micro) and from 0.86 to 1.0 (macro) (see Bleistein et al., 1991).

[6]The correlations were obtained in personal communication from Gary Marco at ETS. See also Wainer and Thissen (1993 p. 114).

[7]See Bleistein et al. (1991) for grade boundary reliabilities.

[8]Further analysis was conducted to explain the difference between actual and predicted AP grades using three methods to construct predicted grades: (i) the ordered probit results from Table 1; (ii) applying the percentage distribution of AP grades to the multiple-choice score distribution; and (iii) proportionally adjusting the cutoff points used for the 90-point composite score to the 60-point multiple-choice score. We were unable to identify factors that explained the three deviation scores in our residual analysis.

TABLE 2—ORDERED PROBIT RESULTS FOR AP GRADE

Variable	Micro[a]	SE	Macro[a]	SE
Constant	−5.439	0.124	−4.537	0.093
MC score	0.364	0.007	0.307	0.005
μ_1	2.511	0.081	2.162	0.067
μ_2	6.090	0.138	4.921	0.098
μ_3	9.104	0.192	7.435	0.136
Log-likelihood:	−2,000.14		−3,013.81	
X^2	8,513.65		9,213.65	

AP grade	Predicted micro AP grade 1	2	3	4	5	*N*
1	550	115	—	—	—	665
2	61	531	118	—	—	710
3	1	122	899	125	—	1,147
4	—	—	138	638	104	880
5	—	—	—	95	469	564
N:	612	768	1,155	858	573	3,966

AP grade	Predicted macro AP grade 1	2	3	4	5	*N*
1	586	130	—	—	—	716
2	74	539	174	—	—	787
3	1	157	1,010	250	—	1,418
4	—	2	236	889	132	1,259
5	—	—	—	225	471	696
N:	661	828	1,420	1,364	603	4,876

Note: For an explanation of the μ's, see William H. Greene (1993 pp. 672–74).

[a]All entries in these columns are statistically significant at the 1-percent level.

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 measure. These results may arise from test definition, construction, or weighting of the composite score (see Bennett et al., 1991; Traub, 1993; Wainer and Thissen, 1993). Preparing and grading one or two essay questions, however, accounted for between one-third and one-half of the testing cost and took more examination time, which raises the question of whether the added cost of the essay component is worth the marginal information benefit. Although the essay section may be valuable for reasons that outweigh cost or information considerations, we could not find empirical evidence to support its use in determining the composite AP grades in microeconomics and macroeconomics.

REFERENCES

Bennett, Randy E.; Rock, Donald A. and Wang, Minhwei. "Equivalence of Free-Response and Multiple-Choice Items." *Journal of Educational Measurement*, Spring 1991, *28*(1), pp. 77–92.

Bleistein, Carole; Batleman, Mark and Flesher, Roberta. "Test Analysis: AP Microeconomics, Macroeconomics." Statistical Report No. SR-91-108, Educational Testing Service, Princeton, NJ, 1991.

Bridgeman, Brent. "Essays and Multiple-Choice Tests as Predictors of College Freshman GPA." *Research in Higher Education*, June 1991, *32*(3), pp. 319–32.

College Entrance Examination Board. *Advanced placement course description: Economics*. New York: College Entrance Examination Board, 1992.

Educational Testing Service. *The 1990 advanced placement examinations in economics and their grading*. Princeton, NJ: Educational Testing Service, 1992.

Greene, William H. *Econometric analysis*, 2nd Ed. New York: Macmillan, 1993.

Saunders, Phillip and Walstad, William. *The principles of economics course: A handbook for instructors*. New York: McGraw-Hill, 1990.

Traub, Ross E. "On the Equivalence of the Traits Assessed by Multiple Choice and Constructed-Response Tests," in R. E. Bennett and W. C. Ward, eds., *Construction versus choice in cognitive measurement*. Hillsdale, NJ: Lawrence Erlbaum, 1993, pp. 29–44.

Wainer, Howard and Thissen, David. "Combining Multiple-Choice and Constructed-Response Test Scores: Toward a Marxist Theory of Test Construction." *Applied Measurement in Education*, Spring 1993, *6*(2), pp. 103–18.