

The Fourth Edition of the Test of Understanding of College Economics

William B. Walstad*

Ken Rebeck**

(December 7, 2007)

*Corresponding author: Department of Economics, 339 College of Business Administration, University of Nebraska-Lincoln, Lincoln, Nebraska 68588-0402. Tel: 402 472-2333; Fax: 402 472-9700; e-mail: wwalstad1@unl.edu

**Department of Economics, St. Cloud State University, St. Cloud, Minnesota 56301-4498. Tel: 320 308-2968; e-mail: kcrebeck@stcloudstate.edu

Paper prepared for presentation at the American Economic Association annual meeting, New Orleans, Louisiana (January, 2008) in the session titled *Assessment of Economics in U.S. High Schools and Undergraduate Programs* chaired by Wendy A. Stock with discussants Alan Krueger, Wendy A. Stock and Tisha L. N. Emerson.

The Fourth Edition of the Test of Understanding of College Economics

William B. Walstad and Ken Rebeck*

This edition of the *Test of Understanding of College Economics* (TUCE-4) is a revision for a test that was developed forty years ago, and has a long history of use by teachers and researchers in the economics profession. The previous editions and their uses have been described in earlier studies (e.g., Rendigs Fels, 1967; Phillip Saunders, Fels and Arthur L. Welsh, 1981; Saunders, 1991) and in research in economic education (e.g., William E. Becker, 1997).

As with past editions, the TUCE-4 has two main objectives: (1) to offer a reliable and valid assessment instrument for students in principles of economics courses; and (2) to provide norming data for a national sample of students in principles classes so instructors could compare the performance of their students on a pretest and a posttest with this national sample. Separate exams were prepared in microeconomics and macroeconomics. Both exams consist of 30 multiple-choice items and can be administered within the time constraints of a single class period for most course formats. What follows is a description of the revision process, the content and cognitive specifications, the norming sample, and the statistical characteristics of the TUCE-4.

I. The Revision Process

This revision of the TUCE was once again a joint effort of the Committee on Economic Education of the American Economic Association and the National Council on Economic Education (NCEE), which provided the funding for the TUCE-4 revision from a Spencer

*Walstad: Department of Economics, University of Nebraska, Lincoln, NE 68588-0402 (wwalstad1@unl.edu); Rebeck: Department of Economics, St. Cloud State University, St. Cloud, MN 56301-4498 (kcrebeck@stcloudstate.edu). This paper is based on the contents of the TUCE-4 *Examiner's Manual* (Walstad, Michael Watts, and Rebeck, 2007).

Foundation grant. The committee members responsible for selecting, writing, and editing the questions on the TUCE-4 were Stephen Buckles (Vanderbilt University); William Bosshardt (Florida Atlantic University); Rae Jean Goodman (U.S. Naval Academy); Paul Grimes (Mississippi State University); Claire Melican (then with the NCEE); Walstad; and Watts (Purdue University). Walstad was the general project director, Watts served as the chair of the revision committee, and Melican was the NCEE administrator for the project. Rebeck reviewed questions and analyzed test data as the associate project director. The NCEE recruited instructors and classes for the norming sample, and organized and collected the norming data.

The test revision committee began working in spring 2004. Content and cognitive specifications, described below, were completed in July 2004, and the committee met in August 2004 to produce the first draft of the TUCE-4. After further review and revisions by the committee, a second draft was field tested as a pretest at the beginning of the spring 2005 semester. The microeconomics test was administered to 660 principles students at six universities, and the macroeconomics test was administered to 1,820 students at seven universities. For comparative purposes, each test was also administered to students taking intermediate theory courses in microeconomics or macroeconomics (40 in micro; 43 in macro).

The results from the pretest field testing were analyzed to identify and replace a relatively small number of items with problems. Those revisions yielded a third draft of the TUCE-4 for a “posttest” field testing conducted at the end of the spring 2005 semester. The third draft of the micro exam was administered to 635 principles students at six universities and the third draft of the macro exam was administered to 1,879 principles students at seven universities.

During this time period, comments on the third draft were provided by a national panel of distinguished economists. The members of this review panel were: Ted Bergstrom (University

of California-Santa Barbara); Daniel Hammermesh (University of Texas at Austin); Alan Krueger (Princeton University); W. Douglas McMillin (Louisiana State University); Arthur J. Relnick (Federal Reserve Bank of Minneapolis); Paul Romer (Stanford University); and Michael Salemi (University of North Carolina-Chapel Hill). Most members of the panel reviewed either the micro or macro exam, but a few reviewed both exams.

The comments from this national panel and the data analysis from the spring posttesting were reviewed by Watts, Walstad, and Melican at a July 2005 meeting. Test items with weak item statistics were eliminated, and other questions were deleted or revised to address concerns raised by the national panel, often to strengthen a particular distractor or wording in a question stem. The full TUCE-4 revision committee participated in writing replacement questions and revising these questions. This resulted in the 35-item fourth draft of the TUCE-4, which was used for the fall 2005 national norming.

After a statistical analysis of the norming data, a decision was made by Walstad, Watts, and Rebeck to eliminate five items from each test. The eliminated items turned out to be more difficult than expected or were unlikely to be measuring the intended concepts as well as the remaining items on each form. The coverage of the content by the final set of 30 items on each form was still within the guidelines established by the revision committee.

II. Content and Cognitive Specifications

The revision committee adopted the following six content categories for microeconomics and set recommended percentage ranges for the allocation of test items: The Basic Economic Problem (7 percent); Markets and Price Determination (22 percent); Theories of the Firm (28 percent); Factor Markets (10 percent); The (Microeconomic) Role of Government in a Market Economy (23 percent); and, International Economics (10 percent). These categories are

basically the same as those found on the TUCE-3, although some of the general descriptions are new, with the older lists of concepts moved to the parenthetical listings of topics for greater format consistency across topics. The stability in general content categories is also reflected in the test items. There are, in fact, only seven entirely new questions on this exam, with 23 items taken from the third edition, though often revised. This “default” position of staying with items from earlier editions, unless there were reasons to change based on minor revisions in the content specifications or problems with item statistics, was explicitly endorsed by the test revision committee. That was done partly because the committee viewed the TUCE-3 micro exam as still generally strong and viable, but also because the time and budget constraints for developing and, especially, field testing new and substantially revised items were very tight.

As always, it was difficult to find items acceptable to large numbers of economists teaching at different colleges and universities, and which also exhibited good item statistics. In a few cases a new item was tried on the field test but not used, and the old item (sometimes revised) was kept if the item statistics or comments from external reviewers suggested problems with the new question. There are, however, some questions on new topics, and in new formats, on the micro TUCE-4 exam. For example, there is a question on game theory, which is now covered in virtually every leading principles textbook. And for the first time on the TUCE, one question features a simple graphical model.

The six content categories for the macro TUCE-4 and the percentage of items falling in each were: Measuring Aggregate Economic Performance (13 percent); Aggregate Supply and Aggregate Demand (25 percent); Money and Financial Markets (13 percent); Monetary and Fiscal Policies (28 percent); Policy Debates & Applications (10 percent); and, International Economics (10 percent). The macro specifications were revised more because of the greater

changes that have occurred in the content and teaching of macroeconomic principles courses since the last revision of the TUCE. For example, aggregate supply and demand models are used in most principles courses and textbooks, but not always, and some recent textbooks written by prominent economists have made a major point to not use them. There has also been some de-emphasis in the coverage of “competing schools” (classical, Keynesian, monetarist, new classical, post-Keynesian, etc.), and in calculating various multipliers. Changes in monetary policy rules and regimes, and in empirical and theoretical models of such topics as economic growth, also have affected the content of most macro principles courses and textbooks. This revision reflects those changes and whatever content consensus there is for a course on macroeconomic principles. As a result, there are 10 new items on the TUCE-4 macro exam, and extensive revisions on most of the 20 other items taken from third edition of the TUCE.

The main purpose of the micro and macro content specifications is to ensure that items on the test cover the core content in a “typical” principles course. If that is done successfully, the total raw score on the exam provides a useful measure of students’ general understanding of basic economics principles. Content classifications of individual test items are often difficult to do, however, because questions often cover more than one concept or principle. In cases where the correct alternative deals with a concept or principle in one category and incorrect alternatives deal with concepts or principles in other categories, test items were generally classified in the category corresponding to the correct alternative. For one item on the micro test and three items on the macro test, the interaction between the alternatives and the situation posed in the stem was sufficiently complex to justify listing the questions in two different content categories.

An international category is included on both tests. The last three questions on the micro test focus on international concepts with a micro orientation (comparative advantage, trade

barriers, and exchange rates), while the last three questions on the macro test focus on international concepts with a macro orientation (balance of payment, exchange rate systems). Test scores discussed below represent the entire 30-item micro or macro TUCE-4, but the greater statistical detail included in the *TUCE-4 Examiner's Manual* allows for computation of norm references with or without the international content category.

Individual questions in each content category vary in difficulty, so no attempt should be made to generalize about the economic understanding of students on a particular concept or principle based on answers to a single question or few questions. It is worth restating that the TUCE-4 is an assessment instrument for measuring the general understanding of principles of economics, not a test of understanding each concept or principle on the test in isolation. Individual instructors or researchers who find that the content specification categories or weightings of these tests are not appropriate for their courses should use the detailed item analysis discussed below to help interpret their results, or perhaps modify the TUCE exams for use with their students. Modifications may, however, affect the validity and reliability measures of the test, and may change the value of the national norms.

This edition of the TUCE uses the same cognitive categories as the previous edition. Each of the 30 items on each test was categorized as Recognition and Understanding (RU, 20 percent of each test), Explicit Application (EA, 53 percent), or Implicit Application (IA, 27 percent).¹ EA or IA items account for 80 percent of the items on each test, a larger proportion

¹Benjamin S. Bloom (1956) proposed a cognitive scheme with six categories (knowledge, comprehension, application, analysis, synthesis, and evaluation) that is widely used. The TUCE is a modified version: RU is a combination of Bloom's first two categories; EA and IA may address one or more of the other three categories (synthesis omitted).

than the 67 percent of the items on the third edition. It is, however, consistent with the general purpose of all previous editions of the TUCE, which sought to emphasize the application of basic concepts and principles over simple recognition of terms and recall of information. In addition, several points should be remembered about cognitive classifications. First, classifying test items by cognitive type is not precise, as is the case with content classifications. Whether the cognitive processes used by students to answer these questions correspond to the level assigned to each question cannot be known with certainty; and any question for which a student has seen the correct answer can become a recall question, regardless of its classification. Second, there is no direct relation between the difficulty of test items and their cognitive level. Item difficulty, as measured by the percentage of correct responses, can vary across all cognitive levels.

III. Norming and Test Characteristics

A total of 70 colleges and universities participated in the TUCE-4 norming during the 2005 fall semester. As with arguably any low-stakes testing of this magnitude, some students who took the pretest did not take the posttest, and some students who took the posttest did not take the pretest. The *Examiner's Manual* reports results for both matched (the group of students who took the TUCE as both a pretest and posttest) and aggregate or unmatched samples, and researchers will have access to the entire unmatched data sets. The results reported below will focus on the matched sample of students.

Data was collected with the purpose of obtaining a sample of students that varied across student characteristics. Using Carnegie-style classifications, 57 percent of the students came from master's colleges and universities, 25 percent from doctoral-granting institutions, 11 percent from baccalaureate colleges and 7 percent from associate's colleges. Males made up 58 percent of the sample, and most students, 61 percent, were either freshmen or sophomores. Non-

white students represented 31 percent of the sample. Only 6 percent of the sample were economics majors, but many students across the nation do not (and some are not allowed to) declare economics as a major until completing part or all of the principles sequence.

The NCEE sought to collect norming data for the TUCE-4 from a large national sample of colleges and universities, but the norming sample is not a *random* sample. Some instructors at institutions who were initially asked to participate in the national norming chose not to do so, and some instructors at institutions who agreed to participate failed to provide complete data. There is, however, a broad representation of students taking principles of economics in the matched samples for the TUCE-4. The 3,255 students taking the micro test and 2,789 students taking the macro test came from 43 to 44 U.S. institutions of higher education.

Table 1 presents the average pretest scores and average posttest scores achieved by the matched sample of students during the 2005 norming, along with the standard deviations, for the micro TUCE-4 and macro TUCE-4, each score out of a possible 30 points. The mean scores for both tests show two desirable traits. First, both tests were able to capture changes in understanding from pretest to posttest, with average scores increasing 3.38 points (11 percentage points, or 36 percent of the pretest score) on the micro exam and 4.39 points (15 percentage points, or 45 percent of the pretest score) on the macro exam. These percentage increases were slightly higher than, although very close to, those found for the third edition of the TUCE.

TABLE 1 – PRETEST AND POSTTEST PERFORMANCE ON THE TUCE-4

TUCE Form	Pretest Score (Standard Deviation)	Posttest Score (Standard Deviation)
Micro [n = 3,255]	9.39 (3.32)	12.77 (4.68)
Macro [n = 2,789]	9.80 (3.48)	14.19 (5.29)

Second, the average posttest scores, 12.77 and 14.19 on the micro and macro exams, respectively, were sufficiently low in percentage terms (43 and 47 percent) to allow for capturing an increase in scores without many students reaching the ceiling. This suggested that TUCE-4 exams are, as intended, difficult measures of college principles-level economics concepts.

The TUCE is a norm-referenced measure, as opposed to a criterion-referenced measure, of economics knowledge. A score of less than 50 percent correct does not necessarily represent a failing level of knowledge in a particular course. If an instructor decides the TUCE covers the basic concepts he or she covers in class and is therefore a valid measure to be used for that class, then individual or group scores should be compared to the results of this national norming. For instance, utilizing the distribution of test scores found in the *Examiner's Manual* would show that a pretest score of 10 on the macro exam would put a student at the 63rd percentile, while a posttest score of 10 would put the student at the 27th percentile.

Test scores are considered reliable if they are a consistent measure of the underlying construct being tested. One common measure of reliability is Cronbach's alpha, which in theory can range from a low of zero to a high of 1. The alpha estimates for the TUCE-4 norming were .70 for the micro exam and .77 for the macro exam. The estimate for the micro exam was significantly lower than the TUCE-3 estimate of .81 (33 items) but for the macro exam the estimate was slightly higher than the TUCE-3 estimate of .75 (33 items). The lower estimated internal consistency for the TUCE-4 micro exam might reflect the greater difficulty of the exam, and the somewhat expanded range of topics the exam now covers, such as game theory. Nevertheless, these reliability estimates at or above .70 suggest that scores from each exam reflect knowledge of the underlying construct with reasonably good internal consistency.

The item statistics can also be used to evaluate the quality of the TUCE-4. A test item should possess three traits that norming statistics can identify: 1) all three distractors should be considered plausible by students who do not know the correct answer, 2) more students should identify the correct answer at the end of a principles course than at the beginning, and 3) students who know more about the underlying construct – principles-level economics – should be more likely to select the correct answer than students who know less. The item statistics for TUCE-4 questions suggest that all items possess the three traits mentioned above. For each item, all distractors drew responses from students at the time of the pretest. At the time of the posttest, more students selected the correct option than at the time of the pretest. The item discrimination coefficients that show the correlation between the item score (0, 1) on the item and the overall test were all positive (.11 to .41). Furthermore, most of the incorrect options drew fewer responses at the time of the posttest.

Background data collected from students cross-tabulated with the test scores also provides evidence to support the validity of the test. The background data include fifteen factors that cover a wide range of student characteristics: gender; age; year in school; race or ethnicity; communicate better in English than another language; communicate equally well in English and another language; type of institution attended; enrollment status; grade point average (GPA); academic major; number of economics courses taken; plans to take more economics courses; expected grade in the economics course at the time of the pretest; expected grade in the economics course at the time of the posttest; and, the number of calculus courses taken.

The results show higher posttest scores compared with pretest scores across *all* fifteen student characteristics and within each category of each characteristic. The findings indicate that performance on the test is responsive to economics instruction, regardless of other characteristics

of students, and suggest that there is *construct* validity to the TUCE-4. The increase in test scores from pretest to posttest are most likely due to economics instruction and are not likely attributable to some other factor or characteristic associated with each student.

The results from the TUCE-4 norming also showed expected differences in economic understanding within item categories. Students with characteristics that would be expected to be associated with higher levels of economic understanding did in fact have higher TUCE-4 scores at the posttest. For instance, with both the micro and macro exams, the average posttest scores increased with students' self-reported GPAs, their expected grade in the course, the number of economics courses taken, and the number of calculus courses taken. Also, those students who planned to major in economics had higher average scores than students with other major plans.

Of particular note as validity evidence is the expected course grade. At the time of the posttest, students probably have an accurate assessment of their final grade. These expected grades and TUCE scores should show a strong positive relationship. They did across all grade categories. In the macro course, for example, students who expected to get an A had a mean score of 17.20 (standard deviation: 5.72) compared with a mean score of 14.14 (4.87) for students expecting a B and a mean score of 12.27 (4.24) for students expecting a C.

IV. Conclusion

The development and revision of the TUCE-4 produced a valid and reliable measure for assessing student achievement in many principles of economics classrooms. The test data collected from administering it in principles classes can be used by instructors and departments to compare with national norms. The TUCE-4 also should be valuable for advancing research in economic education because it provides a standardized test that can be used to assess student achievement in principles of economics across different institutions or classes.

REFERENCES

- Becker, William E. 1997. "Teaching Economics to Undergraduates." *Journal of Economic Literature*, 35(3): 1347–1373.
- Bloom, Benjamin S., ed. 1956. *Taxonomy of Educational Objectives*. New York: David McKay.
- Fels, Rendigs. 1967. "A New Test of Understanding in College Economics." *American Economic Review, Papers and Proceedings*, 57(2): 660–666.
- Saunders, Phillip. 1991. "The Third Edition of the Test of Understanding of College Economics." *American Economic Review, Papers and Proceedings*, 81(2): 32–37.
- Saunders, Phillip, Rendigs Fels, and Arthur L. Welsh. 1981. "The Revised Test of Understanding of College Economics." *American Economic Review, Papers and Proceedings*, 71(2): 190–194.
- Walstad, William B., Michael Watts, and Ken Rebeck. 2007. *Test of Understanding of College Economics: Examiner's Manual* (Fourth Edition). New York: National Council on Economic Education.