DOCUMENT RESUME

ED 242 589 SO 015 440

AUTHOR Walstad, William B.
TITLE Measurement in Economic Education Research.
PUB DATE Sep 83
NOTE 36p.; Paper presented at the Joint Council on Economic Education-National Association of Economics Education Annual Meeting (San Antonio, TX, October 6, 1983).
PUB TYPE Viewpoints (120) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Affective Measures; *Cognitive Tests; *Economics Education; Educational Research; Elementary Secondary Education; Evaluation Needs; Higher Education; Item Analysis; Norm Referenced Tests; Research Needs; Standardized Tests; Teacher Developed Materials; *Test Norms; *Test Reliability; *Test Validity
IDENTIFIERS Basic Economics Test; Test of Economic Literacy; Test of Understanding in College Economics

ABSTRACT

The basic measurement topics of reliability, validity, and norms as they apply to the major norm-referenced cognitive and affective instruments in economics education are discussed to provide researchers with a framework for judging technical quality. Because the topic of measurement is neglected or given improper treatment in much research work and because few national data sets containing reliable and valid data are available to researchers, attention to the technical properties of instruments used to collect data is essential for a sound empirical study in economics education. The paper begins with a study of cognitive tests. Identified are what to look for in the reliability and validity information accompanying a standardized economics achievement test. Specific tests are discussed. The alternative to standardized tests, teacher-developed tests, is considered. Norms are also examined. Next, the author deals with affective instruments. Many researchers fail to report information on the reliability or validity of affective measures, now widely used in economics education. This measurement problem is examined in depth, and general guidance is given on ways to evaluate new measures. (RM)

MEASUREMENT IN ECONOMIC EDUCATION RESEARCH

by

William B. Walstad*

(September 1983)

*Associate professor of economics and Director, Center for Economic Education, University of Nebraska-Lincoln, Lincoln, NE 68588-0400. Paper prepared for the JCEE-NAEE Research Clinic on Econometric Techniques, San Antonio, Texas, October 6, 1983.

MEASUREMENT IN ECONOMIC EDUCATION RESEARCH

An understanding of measurement is essential for work in education. For example, a noted measurement authority has stated:

> In today's educational milieu just about 50 percent of the problems we encounter do, in fact, involve test use, test construction, or test interpretation. Consequently, just about any kind of specialist who, lacking knowledge about measurement, goes out to do battle with today's educational problems is almost certain to come back a loser. For the present and foreseeable future, educators who wish to be effective in their work simply must master the major tenets of educational measurement (W. James Popham, 1981, p. 4).

This recommendation for educators also applies to researchers in economic education. Knowledge of econometric techniques is not sufficient to do the work, we also need a firm grasp of measurement principles.

In essence, empirical work in economic education begins with measurement. We can identify research problems, specify hypotheses, and construct an elaborate research design, but we must start our empirical studies with measurement. In fact, many worthwhile research ideas have probably been abandoned for lack of available instruments to measure important inputs or outputs. The continuing work also depends on measurement. Statistical tests are based on comparisons among measures. Conclusions are drawn from the statistical analysis of the measured data. Strictly speaking, the use of poor quality instruments or the use of instruments with incomplete technical information raises doubt about the findings of a study, even if the study is done carefully in all other respects.

Since measurement is central to the research process in economic education, this chapter provides a general introduction to the topic. The major cognitive tests of economic understanding are discussed from the perspective of the technical characteristics of reliability, validity, and national norms. Suggestions are also offered for what to do when a standardized economics test

is not available. In addition, affective measures are now widely used in economic education, but many researchers fail to report information on reliability or validity of these measures. This measurement problem is examined in depth and general guidance is given on ways to evaluate new measures.

A caveat is probably in order at this point. The chapter is not designed as a substitute for the basic coverage of material presented in measurement texts (c.f., Ebel, 1979; Gronlund, 1981; Nunnally, 1978). These texts offer extensive information on many measurement topics and should serve as a background reference in the same way researchers use econometric texts. This chapter only discusses the basic measurement topics of reliability, validity, and norms as they apply to the major norm-referenced cognitive and affective instruments in economic education so researchers will have a framework for judging their technical quality.[1]

The Major Cognitive Tests

The assertion that measurement questions have been completely ignored in economic education research is not supported by the evidence, especially when one looks at the major cognitive measures. In a recent literature review, Weisbrod found that 22 perdent of the 106 papers focused on "how to define and measure outputs" (1979, p. 15), with no doubt most of the studies being of cognitive measures. Also, more than a decade ago Rendigs Fels recognized that:

> Hypothesis-testing requires measurement, quantification, and data. One reason there has been little hypothesis testing in economic education has been the lack, until recently, of objective measuring instruments (1970, p. 27).

The instruments that Fels was referring to were the: Test of Economic Understanding (TEU) for high school students; the College Level Entrance Placement (CLEP) examination in introductory economics; and, the then new Test of Understanding of College Economics (TUCE).

The TUCE became the standard cognitive instrument in economic education research with 62 of the 100 empirical studies in the Journal of Economic Education from 1969 to mid-1983 reporting its use. The TUCE was recently revised after a 12 year period and the revision (RTUCE) is available for further research work (Saunders, 1981). Similarly, the Test of Economic Literacy has replaced the TEU at the high school level (Soper, 1979). And, the new Basic Economics Test (Chizmar and Halinski, 1980) offers researchers interested in measuring achievement at the upper elementary level a test to replace the Test of Elementary Economics.

The RTUCE, TEL, and BET give economic education researchers a set of nationally normed and standardized instruments at both the college and pre-college levels.[2] Each measure, however, needs to be carefully analyzed for information on its reliability, validity, norming, and test item data before it is used by researchers. Blind use may lead to misuse and close study of each instrument can reveal where caution is necessary or where further measurement work is required. We begin with a discussion of reliability and then turn to validity, norms, and test item data.

Reliability

Reliability refers to the consistency of measurement, or the capacity of test to measure student performance accurately. Any test which contains too much randomness (error) cannot be used for making comparisons or decisions in research work. Random errors of measurement are present in any test, so it is the degree of consistency which is of interest and which we estimate when we look at reliability. What we search for are instruments that are reliable measures of student performance over time, over test conditions, and over samples of items.

Basically, there are four ways to estimate reliability. The first method provides an estimate of the stability of the test over time. This property is estimated by the correlation between test scores of the same test given at two different points in time (usually over a two-week period) without any intervening treatment or instruction. This test-retest method accounts for: (1) constancy of student response on the test over time; and, (2) the consistency of the test procedures since two test administrations are necessary.

In theory, a test only contains a sample of all possible items in the sampling domain. When "parallel" forms of a test are developed, we have two samples of test items from the test domain. By administering the two parallel tests to students and correlating their scores we can examine the property of test equivalence, where we look at how consistent scores are from one sample to another. This equivalent-forms reliability method takes into account: (1) the consistency of measurement over different samples of items; and, (2) the consistency of the test procedures since two test administrators are necessary.

When equivalent forms of a test are administered over a time period, then the property of test equivalence and stability can be estimated by correlating the test scores from the two test administrations. This combination method takes into account: (1) the constancy of student response over time; (2) the consistency test procedures; and, (3) the consistency of the test over different samples of items. We also refer to this method as an equivalent forms reliability, but with a time interval.

Finally, we can obtain information on the internal consistency of a test. Internal consistency indicates whether the items in a test are measuring a common characteristic, or whether the test is homogeneous. A common approach to estimating internal consistency is to split the test in half and correlate the scores from each half, producing a split-half reliability estimate. A more

sophisticated and recommended method to estimate internal consistency is by Kuder Richardson 20 (KR-20) or Cronbach alpha formulas. These procedures offer internal consistency estimates which are essentially an average of all possible split-half coefficients and they are popular because they require only one test administration.[3] Internal consistency estimates accounts for: (1) the consistency of the test over different samples of items; and, (2) consistency over test conditions since in theory two test scores are being correlated.

With this basic background on reliability, we can now compare the four approaches across consistency factors. As shown in Table 1, different information is provided by different estimates. A test-retest method with a time interval gives us information on the consistency of the test procedure and constancy of student response, but tells us nothing about the consistency of the test over different samples of items. The internal consistency estimates tells about the consistency of the test over different samples of items and over test procedures, but tells us nothing about the constancy of student responses over time. Only the coefficient estimate of equivalence and stability (equivalent-forms reliability) accounts for all three consistency considerations, making it the most rigorous reliability test.[4]

Insert Table 1 about here

Reliability of Economics Achievement Tests

We now have a framework to judge the reliability of the various economics achievement tests. Only internal consistency estimates are reported for the BET, TEL, and RTUCE. These estimates offer information on the consistency of the measure over different samples of items and over test procedure. We have no

information on the constancy of student response over time. In fact, since the BET, TEL, and RTUCE all have "parallel" forms, it is surprising that no data are reported to support the equivalence assertion. In this case, one test expert recommends that "a teacher should look with suspicion on any test that has two forms available and does not report information concerning their equivalence" because without evidence "the comparability of the results of the two forms cannot be assumed" (Gronlund, 1981, p. 98). The probable reason for the omission was the expense and difficulty of arranging two test administration for the large national sample. Yet, a smaller sample study could be offered as stability and equivalence evidence. So while we have some reliability information, further measurement work would help us make a more complete judgment about the reliability of the RTUCE, TEL, and BET.[5]

The internal consistency estimates for the RTUCE, TEL, and BET can still be useful. The posttest KR-20 estimates for the RTUCE were: .81 for macro form A; .76 for macro form B; .75 for micro form A; .74 for micro form B; .73 for the hybrid micro/macro form A, and .71 for the hybrid micro/macro form B. The TEL showed Cronbach alphas of .87 for each test form. The alphas for the BET were .83 for form A and .78 for form B.

What do these numbers mean? Reliability is measured on a scale from .00 to 1.00, with 1.00 indicating perfect reliability and .00 indicating no reliability. Since our estimates are somewhere in between, but over .70, are the instrument reliable as estimated by internal consistency formulas? The answer is yes for research purposes. As Nunnally (1982) states:

> It is not necessary for the reliability to be as high in instruments that are used for research in education or related fields as it is for such practical applications as assessing the progress of students in school . . . In basic research a good working rule is that the reliability coefficient should be at least .70, but it is not always necessary to have reliabilities that range into the 90s (p. 1600).

According to this working standard, none of the instruments should be used for applied work where important decisions and reliability coefficients over .90 are necessary. The BET, TEL, and RTUCE do meet the standard for research work.[6]

On many occasions researchers want to study the difference in performance from pretest to posttest on the RTUCE, TEL, or BET, where the difference score is considered to be a measure of value-added for an instructional unit or an experiment. Since difference scores are calculated from two fallible tests, difference scores will be an imperfect measure of change. The reliability of difference scores is lower than the average reliability of the two tests from which the difference score is calculated.

A formula for calculating the reliability of difference scores is:

$$r_{DD} = \frac{r_{AA} + r_{BB} - 2r_{AB}}{2\,(1 - r_{AB})}$$

Where r_{DD} is the reliability of the difference between test A (pretest) and test B (posttest); r_{AA} is the reliability of test A; r_{BB} is the reliability of test B; and r_{AB} is the correlation between pre- and posttests. For example, if a test A and test B have the same reliability (.7) and the pre- and posttest correlation is .7, then the reliability of the difference test is .00. An increase in the reliability of test A and B to .8 only increases the reliability of the difference test to .33. Another way to increase the reliability of the difference score is to decrease the correlation between the pretest and posttest, but this raises questions about test validity (c.f., Brown, 1970, pp. 88-91).

The low reliability of difference scores usually makes them an inconsistent and risky form of measurement to use for either research comparisons or important decisions about student performance. Unfortunately, there appears to

be no recognition in economic education literature that difference scores are basically unreliable measures and to date few researchers have examined this measurement problem. While a value-added measure may be useful for most traditional areas of economic research, when calculated using student test scores, it is probably a measure with low reliability.[7]

One final point needs to be mentioned before we turn to other matters. Reliability is a necessary, but not a sufficient condition for test validity. We can produce an instrument which has great internal consistency and good stability. If, however, a test does not measure the property that we wish it to measure, then the consistency or reliability of the measure is of little value. A test with a high reliability estimate does not mean that the test possesses high validity. A complete look at the RTUCE, TEL, and BET require an inspection of the validity of these tests.

Validity

The most important characteristic of a test is its validity, or the extent to which it measures what it is designed to measure. Validity is not a property that the instrument possesses, but is specific to the situation for which the instrument is intended to be used. The RTUCE, for instance, may be a valid measure of introductory college economics; it is not a valid measure of introductory college mathematics. Validity is also based on the "soundness of the interpretation" of the test results for a particular group of individuals and only the interpretation of the test data has validity, not the test instrument. Validity, therefore, is determined in the context of the situation where the instrument is used and the interpretation of the results produced. While we may use the terms "test validity" or the "validity of a test," the above qualifications should be remembered when studying the validity of the RTUCE, TEL, and BET.

As was the case with reliability, there are several types of validity for a test instrument. Content validity refers to the degree to which items on a test adequately represent a sample of the content area under study. Content validity is the most important type of validity to consider for most achievement tests because of the focus on subject matter. Criterion-related validity involves determining how well performance on the test correlates with performance on another "criterion" measure. When we correlate performance on the instrument and a criterion measure at the same point in time, we refer to this criterion-related validity as concurrent validity. When we use the performance on the test instrument to predict performance on as criterion measure given at a latter point in time, we are working with a form of criterion-related validity called predictive validity. Finally, there is construct validity which includes methods for obtaining evidence on how well an instrument measures an unobservable "construct" such as "economic understanding." Construct validity is probably the most comprehensive of the three types of validity, but proper test validation may require information on all types. We will examine the RTUCE, TEL, and BET from the various validity perspectives.[8]

Content Validity:

The establishment of content validity is most important for achievement test development and is probably the strongest feature in the development of the RTUCE, TEL, and BET. Each test contains a test specification matrix which includes information on the content domain covered by the test. The TEL and BET were developed using the content framework in the Master Curriculum Guide (Hansen, et. al., 1977) to identify the content areas which should be covered by the test. (With the BET this list was modified somewhat since certain listed in the MCG are not even taught to elementary children). For the TEL and BET, a

working committee wrote the test questions, pretested the items, and received feedback from a national advisory committee. Similar procedures were followed with the development of the RTUCE, except that no handy written framework was available to specify concepts taught in the "typical" introductory college economic course. The test content, however, closely parallels most of the basic concepts covered in a standard micro- or macroeconomics principles text.

Test questions were also developed and categorized according to a cognitive level classification. The BET and TEL cognitive classification were based on a modified form of the widely used cognitive taxonomic system developed by Bloom (1965). The RTUCE used a classification system defined by the test developers, which categorized questions as realistic, implicit application, and explicit application. Why an ad hoc specification was used for the RTUCE rather than a more widely accepted one, such as Bloom's taxonomy, remains a mystery and may be a weakness in this test. (We will return to this point when we discuss construct validity).

The content-cognitive test specification matrix for Micro form A of the RTUCE is provided in Table 2 for illustrative purposes. Establishing content validity is a complex process involving rational judgments by experts and is not simply a statistical calculation. The potential problems with the content validity of the RTUCE, or for the TEL or BET, involve the appropriateness and weighting of the content-cognitive matrix. Criticism could be directed at test committee judgment on the selection of concepts and the cognitive level at which they are tested for the groups of introductory economics students under study.

Insert Table 2 about here

For example, the RTUCE can be viewed by some instructors as not being representative of the content covered in their particular introductory

courses. Since the original TUCE faced a similar criticism, it may be worth reemphasizing a point made in the original TUCE manual:

> Whether the TUCE is a valid test depends on the purposes for which it is used. Some economics instructors will no doubt disagree with the content or objectives emphasized by the test committee. For these instructors, TUCE will not possess content validity (Fels, et al., p. 15).

This point also applies to users of the BET, TEL, and RTUCE. Researchers need to make certain that the instrument is appropriate for the situation under investigation. The BET, TEL, and RTUCE are general achievement measures and when we wish to make comparison across courses on general achievement in economics, these measures are quite valid to use. In a research investigation, there may be differences in what is emphasized in one course over another, but if the dfferences are slight, and if comparisons are to be made, then a standardized measure is still appropriate. On the other hand, the RTUCE is not a valid measure to use for grading or evaluation purposes in a course where the course content differs substantially from the content coverage of the RTUCE.

Criterion-related Validity

Criterion-related validity is often determined by correlating performance on economics test with a "criterion" instrument. The major problem, of course, with this type of validity is the selection of an appropriate criterion instrument. The better the instrument, the stronger the validity evidence. This procedure also gives researchers an empirical method for supporting validity claims rather than the judgmental approach of content validity.

None of the new economics tests provide any criterion-related evidence to support validity claims, possibly because no suitable criterion could be found at the time the test was constructed. A few suggestions, however, for future work could be made. RTUCE scores could be correlated with scores on the CLEP

economics test. Scores on the TEL could be correlated with scores on the hybrid RTUCE. Large national samples may not be available for this concurrent validity work, but some small sample studies might offer new validity evidence.

What might be of even more interest would be to use the instruments for predictive validity work. Are scores by high school seniors on the TEL useful for predicting performance, either on grades or on the RTUCE, in the introductory economics course taken a year later? Does the RTUCE have any predictive validity for later performance in upper level economics courses? Can the BET be used to predict student performance in economics in junior high school? These questions suggest areas for future predictive validity work with the BET, TEL, and RTUCE. Information on the predictive power of our measures, while controlling for background variables, may help guide curriculum work in economics.

Construct Validity:

"Economic understanding" is essentially an unobservable construct which we wish to measure, and so we must consider construct validity as well as the other types of validity. Several methods are used to establish construct validity. First, we can make predictions about how certain groups will perform on the measure and then test the groups and compare performances. We might predict, for example, that the high school student would score lower on the RTUCE than college students who have completed the introductory course, but that graduate students in economics would outperform both groups. Some limited evidence of this type is provided by the TEL since there is a statistically significant difference in test scores for groups of students classified with and without economics training. A statistically significant difference was also reported with groups who took the RTUCE as a pretest and groups who took the RTUCE as a posttest in an introductory economics course.

In addition, we can examine construct validity by making predictions about the effects of intervention or treatment. If test scores respond to instruction, then this finding is evidence to support construct validity. The best example of this evidence is provided with the BET. An analysis of variance was conducted examining BET test scores while controlling for grade level, sex, instruction in economics, and a sex interaction variable. The analysis showed that student scores in economics increased with the amount of economics instruction received. What is unique about this analysis for the BET is the attempt to control for background variables. More work may be needed here however, since no measure of general ability or reading was included in the model and grade level may not be an adequate proxy for these factors or other factors might might be especially important (i.e., socioeconomic status).

Obtaining correlations with other measures is a third way to support construct validity claims. Probably the most important construct validity work to be done with all three tests is to support the cognitive level claims for test questions. Assertions are being made that parts of each test are assessing higher level cognitive skills, but we have no evidence. An earlier study which sought to address this type of problem was conducted by Lewis and Dahl (1971) for the old TUCE. In essence, this study looked at the correlations among the TUCE or its subparts, and other measures, such as the ACT test or a critical thinking test, to assess the cognitive level construct validity of the TUCE. This topic is ripe for further work with the new economics test and the Lewis and Dahl study offers a useful starting point.

Norms

An important characteristic of a standardized economics achievement test is the availability of national norms. To be certain that norming data is useable

for comparison purposes, both the quantity and quality of the comparative data need to be judged. The usual criteria to look for are: (1) the norming sample size; (2) the recentness of the data collection; (3) the representativeness of the sample; and, (4) complete description of the test procedures.

An examination of the RTUCE, TEL, and BET indicates that the instruments meet all these criteria. The RTUCE data were collected from over 7,000 college students taking introductory economics courses in 24 colleges and universities of various sizes across the United States in the spring term of 1979. TEL norming data were obtained from over 8,500 eleventh and twelfth grade students in 92 high school classrooms in 36 states in May and June of 1977. The BET norming data were collected from over 14,000 fourth, fifth, and sixth graders in 56 classrooms in 23 states in May 1979. Thus, large sample sizes were used and the data were collected recently. Given the spread of the sampling across states and educational institutions, we also have some information that the test developers tried to obtain a representative sample of students for each level, although without random sampling procedures we may never be as certain of this judgment as we might wish to be. Detailed test procedures and interpretation information is also contained in each published test manual.

The quality of the norming data is critical to comparisons made with test scores, either for groups or individuals. With norming data we can convert a raw score to a percentile rank based on the use of the norming sample data. To illustrate, if a researcher found that a class of twelfth graders received a mean score of 26 on the TEL (form A) after an instructional unit in economics, then the researcher interprets the class performance as meaning that the class performed as well or better than 50 percent of the norming sample of 12th graders with economics instruction. Even if researchers had other data available on group performance in similar classes, the norming data could still be

used as a basis for comparison for all classes and as a way to add meaning to the interpretation of the test scores.

When using norms researchers must remember that the norms were developed from the norming sample. Norms are not statement of what should be or ought to be; they should not be viewed as standards. Norms are simply a large data set available for comparison purposes. This comparison is not with all student or classes at that age (grade) level, just the norming group.

The age of the norming data also becomes more critical over time. The older a test the greater the probability the norming sample scores are outdated for comparison purposes and reliability and validity date are more suspect. This problem is mentioned because test revision in economic education has not been frequent. The norming data for the "new" TEL is now over 6 years old and may soon be in need of revision, but a 15 year period lapsed before the TEL replaced the old TEU. The developer of the RTUCE also recognized the time problem and recommended that the RTUCE tests "be revised more frequently than the 12 years that elapsed between publication of the original TUCE and the current revision." (Saunders, 1982, p. 10). As stated earlier, empirical research is influenced by the quality of the data collected by the major test instrument, and consequently, we all have a stake in test instrument development even if we do not do the measurement work. Timely revision of major measures is essential for research work in this field.

Item Analysis

Data on the difficulty level of each item is provided in the BET, TEL, and RTUCE manuals. (Difficulty level refers to the percentage of students in the norming sample who got the item right). In addition, data are presented on the discriminating power of each item, or the ability of our item to distinguish

between students who do well on the test and those who do not. With the RTUCE, for example, the discriminating power is measured with a point biserial correlation between the mean score of those giving a correct response on an item and the mean score of the total norm group for that test.

While item data may be of interest to instructors who wish to evaluate student performance on particular items with the norming group, item data is usually of little interest to researchers. There may be items that researchers do not like or do not think show sufficient difficulty or discriminating power. A test, however, is an index and what we need to know is whether this index is an adequate measure of the construct under study. This quality is most properly assessed by the reliability and validity characteristics. Remember the maxim: judge the overall test, not individual items.

Standardized versus Teacher-made Tests

Reliability and validity have been called the "meat and potatoes" of educational measurement. In the previous discussion we identified what to look for in the reliability and validity information with a standardized achievement test in economics. In certain areas the RTUCE, TEL, and BET offer the researcher only limited information on the major technical features, most noticeably with equivalent-forms reliability estimates. Since measurement and test instruments lay the foundation for research work, we must continue to increase the amount of reliability and validity data to maintain high standards for research.

No suggestion is being made that these measures not be used because of the lack of complete information. The RTUCE, TEL, and BET are the best available instruments for research and are of good quality. The test development process is also an arduous one. Researchers who ventured into this area and produced

instruments of the quality of the RTUCE, TEL, and BET, given time and resource constraints, are to be applauded for their labors which resulted in a long-run contribution to the field.

We should also consider the alternative to standardized tests-teacher-made tests. The basic differences between the two types of tests should be reviewed before the use of a standardized instrument is rejected in favor of a "home-made" substitute. As we have illustrated, items for standardized measures are carefully written, pretested, and selected by a committee of experts; teacher-made test items are not constructed with the same level of quality. Standardized test also provide a manual with detailed reliability and validity data, norms for comparison purposes, and clear test procedures; information on the technical characteristics and test procedures is often unknown or unpublished with teacher-made tests. A standardized test can be used for comparison and research purposes; the classroom test is only sufficient for evaluating individual student performance in a particular classroom, and it is rarely of acceptable quality for use in research. In short, a teacher made test is not likely to inspire much confidence in the results of a study where they are used, and researchers should have good grounds before rejecting the use of a standardized test.[9]

A Modified Test: An Example

A teacher-made test may be appropriate when no standardized test is available for use with the group under study or when there are limits to the test period. Even in these situations researchers are better off searching for a previously developed instrument as a source for items with some reliability and validity information.

For example, in an evaluation study of an elementary school program,

Walstad (1979) used a 29-item version of the 40-item Test of Elementary Economics (West-Springfield, 1971) as the evaluation instrument since the BET had not been developed at the time the study was conducted. A shortened form of the TEE was required for several reasons. First, the TEE was originally developed for use with sixth graders and in the study the target group was fourth graders. Second, a shortened instrument was needed to fit the limited classroom testing period. Third, the norming data from over 2,500 elementary students in New England was dated and test reliability needed to be checked.

When the TEE was administered to a separate sample of 63 fourth, fifth, and sixth graders in a local school district, the reliability (KR-20) was a low .53. Eleven items had difficulty levels (percent correct) below chance (.25) or had negative high-low discrimination (percentage difference between the highest and lowest scoring groups for the correct alternative). By shortening the length of the test from 40 to 29 items, the reliability actually increased rather than decreased and was a modest .65. The item mean difficulty level was .43 and item mean discrimination level was .36. After the separate sample work, the reduced TEE was used in the study and showed an internal consistency reliability of .71, which was acceptable for research work. If the more difficult and poorly discriminating items of the original TEE had not been eliminated, the reliability estimates would no doubt have been much lower.

The improvement in reliability of the TEE by omitting items did not appear to come at the expense of content validity. A comparison was made between the original content matrix and the reduced content matrix for the TEE. All content areas in the orfginal test were still represented in the reduced test. The items eliminated were basically ones of a factual or historical nature unrelated to the study, or else the items duplicated material covered in other items. The reduced test represented the best available instrument to test the general

understanding and application of economic concepts likely to be taught to classes at these grade levels. (See Table 3 for test matrix comparisons).

Insert Table 3 about here

Conclusion on Teacher-made Tests

In instances where no previous instruments are available to offer guidance or test questions, then researchers must start from "scratch." Guidelines for good test construction can be found in most measurement texts. Ideally, the test should be pretested with a separate sample before it is used and information made available in the research report on the descriptive test statistics, reliability estimate, and how validity was determined. Teacher-made tests can be made acceptable for research (at least for exploratory work) as long as we have adequate documentation of the technical characteristics so we can judge the quality of the measure.[10]

One educational researcher has recommended that "editors and reviewers ought to routinely return papers that fail to establish psychometric properties of the instruments they use" (Willson, 1980, p. 9). This standard is a strict one and if it were applied to cognitive measures used in recent studies published in the Journal of Economic Education, then a number of studies would be returned for lack of complete information. For example, studies by Ferber, et. al., (1983); Paul (1982); and Swartz, et. al., (1980) all use a teacher-made multiple choice economics test as a measure of output, but no study provides any reliability data on the developed measure. In addition, the discussion of test validity is limited since we are not told what content areas were covered by the test. In each of these studies, for example, TUCE items were included as part or all of the teacher-made tests, but we are not told which items were selected. In other words, the reliability and validity information on the

instruments is incomplete and we are forced to accept the opinion of the researcher that the measures are sound. When teacher-made or modified standardized tests are used, then more test information is required because we have no test manual to consult.[11]

Affective Instruments

Affective instruments have been incorporated in economic education research from the first issue of the Journal of Economic Education. Every issue thereafter usually contains one or more articles with an affective measure as an input or output variable in model specifications. Constructs which have been examined in the research literature include: attitudes towards methods of instruction (McConnell and Lamphear, 1969); economic attitude sophistication (Mann and Fusfeld, 1970); student course evaluations (Villard, 1973); attitudes towards economics (Karstensson and Vedder, 1974); attitudes towards economic issues (Riddle, 1978); and, learning and instructional styles (Miller, 1982). A review of the findings on the significance of some of these affective variables at the college level is found in Siegfried and Fels (1979, pp. 930-937).

The Neglect of Documentation

Paradoxically, while there has been great interest in affective instrument and recognition of the need to use normed, reliable, and valid measures of cognitive achievement, little attention has been paid to the measurement qualities of the affective measures. It appears to be acceptable practice in economic education to develop an attitudinal measure and use it in research without documenting its technical characteristics. Henry and Ramsett, for example, examined the effects of computer-aided instruction on learning and attitudes in principles courses. All we are told about the attitude towards

economics measure is: "this score was obtained by having students complete on attitude test at the end of the course." (p. 28). This example may seem extreme; it is not, but the impulse to list of the studies in economic education that provide no documentation of the characteristics of the measure used or that provide incomplete documentation will be resisted and the task left as an exercise for the reader.

There is no substitute for an exact description of the affective measure with information on how the instrument was developed, the validation procedures, its reliability, and the samples to which it has been administered by the original developer. We would also want at least an estimate of the internal consistency reliability with the group under study. Devoting one or two paragraphs or lengthy footnotes to the description of measures in an article is not too much to ask of researchers. Since the conclusions are ultimately based on the quality of the instruments used, this consideration should be sufficient justification for subtantiating the value of the instruments.[12] In addition, affective measures are often viewed as "softer" than cognitive measures, and it might be expected that more "rigorous" documentation would be both desired and required of affective measures before they are used. The reverse has been the case to date: higher measurement standards are found with cognitive rather than affective measures, on average.

Criteria for Affective Measures

Basically, the same evaluation criteria apply to both cognitive and to affective measures. When considering validity, it is necessary to look at the three kinds of validity evidence--content, criterion-related, and construct. Achievement tests usually give most weight to content validity, but with affective measures the emphasis shifts to documentation of construct validity.

Also, criterion-related validity can be even more difficult to determine in the affective domain than in the cognitive demain because there are even fewer suitable criterion instruments. Trying to predict behavior from responses to self-report attitude measures has not met with much success. Despite these problems, we will seek evidence in all three ways to support our contention that the instrument measures what it purports to measure.

Information on the reliability of the affective measure also should be reported by researchers. In most cases, affective measure require only the reporting of internal consistency reliability. Estimates for stability through the use of a test-retest procedure may not be appropriate because of the reactive nature of the measures, and parallel forms are rarely available for affective measures. The reliability range for affective measures is often lower than for achievement measures, making a minimum standard difficult to specify, but any affective measure with a coefficient of .60 or greater is probably acceptable for research work.

As was the case with cognitive tests, the norming sample for affective measures should also be large and representative of the population under study so we have some assurance that the technical property of reliability (or validity) is being estimated with an appropriate group. Affective measures also need to be revised on a timely basis to maintain the value of the norms. Since the scores for an affective instrument are summations across different items, researchers should eschew item analysis and concentrate on the meaning and interpretation of the overall score.

The SEA: An Example

The only affective measure for economic education which approaches the standards of such cognitive measures as the RTUCE, TEL, or BET is the two-part

Survey on Economic Attitudes (SEA) consisting of 28 Likert-type statements (Walstad and Soper, 1983; Soper and Walstad, forthcoming). The first part of the SEA assesses attitudes towards economics (ATE). The second part of the SEA examines economic attitude sophistication (EAS), or the degree to which student views are in agreement with the consensus views of economists on economics issues. The instrument was nationally normed with a group of about 1,700 high school students (11th and 12th graders) in 67 schools in 35 states in May 1979. Small sample work also indicates that the instrument (at least the ATE) may be suitable for use at the college level.

A detailed description of the reliability and validity work for the SEA is provided in the previously cited work and will only be briefly described here. The Cronbach alpha was .88 for the ATE and .66 for the EAS with the large high school sample. Similar estimates for each instrument were obtained with the college samples. Although the alpha for the EAS is somewhat lower than the ATE, this difference is probably due to the difficulty of obtaining internal consistency when assessing attitudes on diverse economic issues and to the short length of the measure. Both ATE and EAS estimates, however, meet or exceed standards for research use.

Reliability is only a necessary condition for validity, so an in depth investigation was made of the content and construct validity of each measure. A working committee reviewed the topics to be included in each attitude measure and received feedback from a national advisory committee to select items and to judge overall content validity. Construct validity evidence was first obtained for the ATE and EAS by statistically testing for the expected differences in responses among known groups--high school students, introductory economics students, advanced undergraduates, and college professors. Student scores on the ATE and EAS were also correlated with scores on other measures--an IQ test,

the RTUCE, and the ACT--to examine whether each measure showed a degree of uniqueness. In addition, for the EAS, a survey test was conducted as part of construct validity work to help identify the "consensus" position of economists and economic educators on economic issues.

The SEA is just one example of a nationally normed affective measure for economic education research. The SEA is not a "perfect" measure and obviously more information on its technical characeristics may be desired by users. The development of SEA illustrates the extensiwe work necessary to document what we are measuring and how well we are measuring it and should represent a distinct improvement over the ad hoc development of most affective instruments in economic education.

Conclusion

Although measurement is central to the research process in economic education, the topic is neglected or given improper treatment in much research work. Maybe this attitude toward measurement is due to the excitement experienced by researchers in other phases of work. Somehow worrying about how we are measuring important input or output variables is just not as exciting or glamorous as the formulation of research hypotheses, the development of the research design, or the distillation of statistical results into general conclusions. Or, perhaps the omission is due to the strong economics influence on economic education; most researchers are trained in econometrics, not psychometrics. In economics research, the data (e.g., GNP, CPI, or retail sales figures) are usually collected by other organizations and individuals. Economic education research, on the other hand, requires the development or selection of measures and data collection by researchers. Few national data sets are available for researchers to used which contain reliable and valid data on

variables of research interest. Consequently, attention to the technical properties of instruments used to collect the data is essential for a sound empirical study in economic education.

An analogy drawn from home economics rather than economics illustrates the problem. Many people enjoy baking--from selecting the recipe, to combining ingredients, to drawing conclusions about the final output. Measuring the ingredients, which is a necessary part of the culinary process, is as exciting or satisfying as the other phases of the experiment. But imagine what would happen when quantities are estimated with invalid measures, or if unreliable measures were used to determine the amount of ingredients, or if the baker knew nothing about the measurement process. Then, conclusions drawn about the finished product and the experiment itself would change drastically. The analogy, as simple as it is, highlights the research problem of reflecting measurement concerns.

At present we only have a handful of valid and reliable instruments for research work. If we are to make more progress in exploring the dimensions of the economics learning, then we will need new measures and we will need to revise the old ones on a timely basis. Becker (Winter, 1983) recognized this problem in a recent review of economic education research:

> The fact that appropriate cognitive- and affective-domain instruments do not exist for a specific assessment task suggests that we should attempt to develop such instruments. Reliable and valid test instruments for all forms of learning are needed. These instruments must measure what they report to measure and do it consistently across individuals and over time. Without valid and reliable instruments, it is impossible to tell what is being measured and to make comparisons to assess results (p. 15).

A career could be made in economic education developing the needed instruments, and although the test development process is becoming more sophisticated, the opportunities for making a solid contribution to the field are great. Future progress in research requires this specialized work.

FOOTNOTES

[1]For researchers interested in criterion-referenced tests and the discussion of reliability and validity, as they apply to these test measures (see Popham, 1981). Some of the points discussed in this chapter are also presented in Walstad and Buckles (1983).

[2]There are other nationally normed economics tests produced by the Joint Council on Economic Education. One is the Junior High School Test of Economics (JHSTE) and the Test of Understanding in Personal Economics (TUPE). The JHSTE was normed in 1973 and the TUPE in 1970. They were omitted from the discussion due to their age. A new economics test for the Give and Take series is soon to be released, and will not be discussed since it was developed for a specific economics program. The CLEP is available from the Educational Testing Service, but is expensive to use in research work.

[3]Coefficient or Cronbach alpha is the basic formula for internal consistency. When test items are dichotomous the KR-20 formula can be used. In this case, Cronbach (1951) has shown that the KR-20 and alpha estimates are equivalent. The formula for coefficient alpha is:

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\Sigma Vi}{Vt}\right)$$

where, n = number of items on the test; V_t = variance of the total test; and, ΣV_i = the sum of the variance of individual items. The only difference between alpha and the KR-20 formulas is that ΣV_i is replaced by Σpq, where Σpq = sum of the variance of items scored dichotomously.

[4]In all fairness, internal consistency estimates are still sufficient for most research studies since the major source of the measurement error is probably due to item sampling. A KR-20 or Cronbach alpha estimate is also valuable because it sets on upper bound to the reliability of a test. If this estimate is not high, then the other types of reliability estimates (equivalent forms) are likely to be even lower (Nunnally, 1978, p. 231).

[5]Another problem with internal consistency estimates is that they may be inflated if the test becomes a speed test rather than a power test. A power test allows sufficient time for all students to complete a test but a speed test does not. The RTUCE, TEL, and BET are designed as power tests, but no data are presented which indicates that the time period specified in the manual is sufficient for all students, so the reliability estimates may be inflated.

[6]A number of factors can influence the reliability estimates of a test. These include test length, spread of the scores, the difficulty of the test, the objectivity in scoring, and the type of estimating procedure (see Gronlund, pp. 104-111).

[7]Whether this measurement error presents a problem for the statistical estimation depends on the methods used. (See Becker, Summer 1983, pp. 6-7).

[8]As Wolf (1982) notes: "Validation of a particular test usually requires an integration of all three types of evidence, and one cannot freely be substituted for another . . . There is a move towards viewing validity as a unitary rather than a tripartite concept (p. 1995).

[9]There are different rationales for the use of standardized and classroom tests. For a discussion of these points, see Becker and Walstad (1981).

[10]So far we have discussed only paper-and pencil-cognitive measures. Other types of cognitive measures which do not rely on paper-and-pencil reactions may be developed (e.g., observations of student economic behavior). These measures must also be shown to be reliable and valid before they are used for research work. The psychometric data may be more difficult to collect than it is for a multiple-choice standardized test, but the general standards still apply.

[11]Even studies which use a standardized measure, such as the TUCE, fail to report what form of the test was used (A or B), how they were used (pretest and posttest), or report any reliability data on the use of that instrument <u>with the sample under study</u>.

[12]The same conclusion applies to non paper-and-pencil affective measures. While it may be difficult to provide extensive reliability and validity information, this data must be furnished if we are to have any insight into what and how well the behavior is being measured. Observations and ratings are riddled with measurement error and invalidity. (See Nunnally, 1982, pp. 1596-1601).

REFERENCES

Becker, William E., Jr. "Economic Education Research: Part I, Issues and Questions." Journal of Economic Education, Winter 1983, 14(1), 10-17.

_____. "Economic Education Research: Part III, Statistical Estimation Methods." Journal of Economic Education, Summer 1983, (14)3, 4-15.

_____ and Walstad, William B. "Evaluating Student Learning of Economics: Investment-Good and Consumer Good Rationales" in S. Stowell Symmes (ed) Economic Education: Links to the Social Studies. Washington, DC: National Council for the Social Studies, 1981, 53-65.

Bloom, Benjamin S. (ed) Taxonomy of Educational Objectives: The Classification of Educational Goods, Handbook 1: Cognitive Domain. New York: David McKay, 1956.

Brown, Frederick G. Principles of Educational and Psychological Testing. Hinsdale, IL: Pryden Press, 1970.

Cronbach, L. J. "Coefficient Alpha and the Internal Structure of Tests." Psychometrika, 1951, 16, 297-334.

Chizmar, John F. and Halinski, Ronald S. Basic Economics Test: Examiner's Manual. New York: Joint Council on Economic Education, 1980.

Ebel, Robert L. Essentials of Educational Measurement (3/e). Englewood Cliffs: Prentice Hall, 1979.

Fels, Rendigs. "Multiple Choice Questions in Elementary Economics," in Keith G. Lumsden (ed), Recent Research in Economic Education. Englewood Cliffs: Prentice-Hall, 1970, 27-43.

Fels, Rendigs, et. al. Manual: Test of Understanding in College Economics. New York: Psychological Corporation, 1968.

Ferber, Marianne, et. al. "Gender Difference in Economic Knowledge: A Reevaluation of Evidence." Journal of Economic Education, Spring 1983, 14(2), 18-24.

Gronlund, Norman E. Measurement and Evaluation in Teaching (4/e). New York: MacMillian, 1981.

Hansen, W. L., et. al. Master Curriculum Grade in Economics for the Nation's Schools: Part I: A Framework for Teaching Econonomics: Basic Concepts. New York: Joint Council on Economic Education, 1977.

Henry, Mark and Ramsett, David. "The Effects of Computer-aided Instruction on Learning and Attitudes in Economic Principles Courses." Journal of Economic Education, Fall 1978, 10(1), 26-34.

Karstensson, Lewis and Vedder, Richard K. "A Note on Attitude as a Factor in Learning Economics." Journal of Economic Education, Spring 1974, 5(2), 109-111.

Lewis, Darrell and Dahl, Tor. "The Test of Understanding in College Economics and its Construct Validity." Journal of Economic Education, Spring 1971, 2(2), 155-166.

Mann, W. R. and Fusfeld, D. R. "Attitude Sophistication and Effective Teaching in Economics." Journal of Economic Education, Spring 1970. 1(2), 111-129.

McConnell, Campbell R. and Lamphear, F. Charles. "Teaching Principles of Economics Without Lecture." Journal of Economic Education, Fall 1979, 1(1), 20-32.

Miller, Jimmie C. "Technical Efficiency in the Production of Economic Knowledge." Journal of Economic Education, Summer 1982, 13(2), 3-13.

Nunnally, Jum C. "Reliability of Measurement," in Harold E. Meitzel (ed) Encyclopedia of Educational Research (5/e). New York: MacMillian, 1982, 1589-1601.

_____. Psychometric Theory (2/e). New York: McGraw-Hill, 1978.

Paul, Harvey. "The Impact of Outside Employment on Student Achievement in Macroeconomic Principles." Journal of Economic Education, Summer 1982, 13(2), 51-56.

Popham, James W. Modern Educational Measurement. Englewood Cliffs: Prentice-Hall, 1981.

Riddle, Terry. "Student Opinion on Economic Issues: The Effects of an Introductory Economics Course." Journal of Economic Education, Spring 1978, 9(2), 111-114.

Saunders, Phillip. Revised Test of Understanding in College Economics: Interpretive Manual. New York: Joint Council on Economic Education, 1981.

Siegfried, John J. and Fels, Rendigs. "Research on Teaching College Economics: A Survey." Journal of Economic Literature, September 1979, 27, 923-969.

Soper, John C. Test of Economic Literacy: Discussion Guide and Rationale. New York: Joint Council on Economic Education, 1979,.

_____ and Walstad, William B. "On Measuring Economic Attitudes." Journal of Economic Education, forthcoming.

Swartz, Thomas R. et. al. "Why Have We Ignored 'The Distribution of Benefits from College Instruction.'" Journal of Economic Education, Spring 1980, 11(2), 28-36.

Villard, Henry. "Some Reflections on Student Evaluation of Teaching." Journal of Economic Education, Fall 1973, 5(1), 47-50.

Walstad, William B. "Effectiveness of a USMES In-service Economic Education Program for Elementary School Teachers." Journal of Economic Education, Fall, 1979, 11(1), 1-12.

_____ and Buckles, S. "The New Economics Tests for the College and PreCollege Levels: A Comment." Journal of Economic Education, Spring 1983, 14(2), 17-23.

_____ and Soper, John C. "Measuring Economic Attitudes in High School." Theory and Research in Social Education, Spring 1983, 11(1), 41-54.

Weisbrod, Burton A. "Research in Economic Education: Is It Asking the Right Questions?" American Economic Review: Papers and Proceedings. May 1979, 69(2), 14-21.

West Sprinfield Public Schools Test of Elementary Economics. New York: Joint Council on Economic Education, 1971.

Willson, Victor. "Research Techniques in AERJ Articles: 1969-1978." Educational Research, June 1980, 9(6), 5-10.

Wolf, Richard M. "Validity of Tests," in Harold E. Meitzel (ed) Encyclopedia of Educational Research (5/e). New York: MacMillian, 1982, 1991-1998.

TABLE 1

Types of Consistency[1]

Reliability Property (Method)	Consistency Considerations: Test Procedure	Constancy of Response	Over Different Samples of Items
Stability (test-retest over time)	X	X	
Equivalence (equivalent-forms) no time interval	X	*	X
Stability and Equivalence (equivalence-forms with time interval)	X	X	X
Internal Consistency (KR-20 or Cronbach alpha)	X		X

*Short-term constancy of response may be reflected but not day to day constancy.

[1]Adapted from Gronlund (1979, p. 101).

TABLE 2.

Test Specification Matrices for RTUCE (Form A)

Macro Form A

Content Categories	Cognitive Categories			No. of Questions
	Recognition & Understanding	Explicit Application	Implicit Application	
A. Measuring Aggregate Economic Performance	1	7, 15	21*X*	4
B. Aggregate Supply, Productive Capacity, and Economic Growth	5, 28	8*X*, 26	11*X*	5
C. Income and Expenditure Approach to Aggregate Demand and Fiscal Policy	13, 22, 23, 27	9	2*X*, 19	7
D. Monetary Approach to Aggregate Demand and Monetary Policy	12, 17, 24*X*	3, 6, 14	20*X*	7
E. Policy Combinations and Practical Problems of Stabilization Policy		10, 25	4*X*, 16*X*, 18*X*, 29*X*, 30*X*	7
No. of Questions	10	10	10	30

Micro Form A

Content Categories	Cognitive Categories			No. of Questions
	Recognition & Understanding	Explicit Application	Implicit Application	
A. The Basic Economic Problem	1	5	13, 16*X*	4
B. Markets and the Price Mechanism	7, 9, 14	6*X*, 28*X*	2, 27	7
C. Costs, Revenue, Profit Maximization, and Market Structure	4, 11, 18	22*X*, 25*X*	26*X*, 29*X*	7
D. Market Failure, Externalities, Government Intervention and Regulation	8, 24	15*X*, 21	17*X*, 20	6
E. Income Distribution and Government Redistribution	10	19*X*, 23*X*, 30	3, 12*X*	6
No. of Questions	10	10	10	30

[1] Table 2 is from P. Saunders (1981, p. 12-14).

TABLE 3

Test of Elementary Economics Matrix

(original versus modifications*)

Concept Area	Knowledge Questions: Fact	Knowledge Questions: Definition	Comprehension Questions	Application Questions
Household	21	2	22, 25	4, 8, 19, 20, 40
Business	13, 17, 34	35	6, 16	
Government	9, 26 39			28, 37
Exchange	1, 14, 18, 33		7, 30, 31, 36	
Techncology			11, 27, 24, 38	
Market			10, 12, 15, 23	29, 32
National economy			3, 5	

*Underlined questions were omitted.