

Assessment for the New Curriculum: A Guide for Professional Accounting Programs

Section 9.1

Basic Principles and Procedures for Designing Learning Outcome Measures

Section 9.1 presents criteria and basic procedures for design and validation of program-level measures. Sections 9.2 through 9.4 offer specific guidance for measurement related to each of the three major categories of learning outcomes. Section 9.5 addresses assessment related to learning to learn.

9.1.1 Criteria for Program-Level Measures

To make program-level judgments with assurance, learning outcome measures should meet several criteria:

- Measures *must* be targeted to specific program goals
- Measures should yield diagnostic feedback about the accounting *program* by revealing patterns of strength and weakness in the performance of a group or groups of students
- Reliability of measurement should be confirmed using methods such as internal consistency checks (useful for multi-item instruments measuring a single construct) and rater agreement checks ("interrater reliability," important when judging complex products or performances)
- Measures should be validated, for example through comparison of results for the same objective using different modes of assessment (e.g., paper-and-pencil analytical tests of problem-solving skill and ratings of performance in simulated or actual problem situations) or by obtaining judgments from practicing professionals for comparison with judgments made by faculty.

Initially, not all measures will meet all criteria. Over a period of years, the faculty's understanding of measurement will increase, leading to improvements in the assessment portfolio.

9.1.2 Design Process for Learning Outcome Measures

The procedure for developing measures of learning outcomes is as follows:

- Review targeted *objectives* and *performance criteria* (described in [Chapter 7](#)) and clarify if necessary
- Select or devise *measurement instruments* and *rating criteria* that *best* represent achievement of the desired performance
- Obtain *performance data* from students
- Apply rating criteria to obtain *performance ratings or judgments*
- Determine *reliability* and *validity* of measures
- *Recalibrate measurement*: review instruments and/or procedures periodically and make needed improvements

Application of this basic procedure will be described for each major category of learning outcomes.

It is not necessary to develop separate measures for each goal or objective. Instead, areas of legitimate overlap should be identified to simplify assessment and reduce costs. For example, a complex, realistic case (such as the multi-media program *Dermaceutics, Inc.*, produced by Coopers & Lybrand, or the *CableCo Chronicles*) can be used to assess professional knowledge, interpersonal skills, critical thinking and complex problem solving, ethical decision-making and professional orientation. To avoid contamination of ratings ("halo effect"), students' performance in each goal area should be rated separately, preferably by different raters. However, a single performance measure should be supplemented by other measures to strengthen interpretation of results.

9.1.3 Reliability and Validity of Outcome Measures

As already noted, a key difference between measures developed for classroom use and those used for program assessment is that greater care is often exercised in program assessment to demonstrate reliability and validity. These terms apply as follows:

Reliability is the degree to which a test minimizes errors due to problems of measurement. A test with only a few items is not as reliable as a longer one, because with fewer items it is less certain that results are not obtained by chance. A reliable measure will yield consistent results when the same sample is tested repeatedly (assuming no relevant intervention has occurred).

Reliability of multi-item tests is frequently estimated based on the internal consistency of responses, using correlational measures such as the Kuder-Richardson formula. The higher the correlation, the more likely that the test is measuring a single quality or characteristic—in this case, students' knowledge of course content. For program evaluation, reliabilities around .60 are acceptable, although higher reliability is recommended if results will be used to advise or make decisions about individual students (Erwin, 1991; Banta and Schneider, 1988).

For performance rating scales, reliability is the degree to which observers agree when rating the same performance independently. This is referred to as interrater reliability (a concept similar to "objectivity" in the accounting literature; see Section 9.3 for procedures). Interrater reliabilities should reach a minimum of 70% (Erwin, 1991).

Validity is an estimate of the degree to which the instrument measures what it is intended to measure (Erwin, 1991). Two basic forms are *content* validity and *construct* validity.

- *Content validity*: In the judgment of experts, the content of the test accurately reflects the content of the course or program on which it is based.
- *Construct validity*: Results obtained from the measure are consistent with other measures associated with the underlying trait or "construct." For example, performance on a formal assessment of oral communication skills in accounting should be positively correlated with grades in courses that require extensive use of public speaking skills and with supervisor or employer ratings of on-the-job communication skills. Scores on an exit test for accounting seniors should predict performance as a practicing accountant (sometimes referred to as "predictive" or "criterion" validity).

Standardized tests such as the CPA exam have been criticized because they do not have high construct (predictive) validity (Ferris, 1982); that is, scores on these tests are not highly correlated with success in the accounting profession. Similarly, ratings of internships may appear to be valid because they reflect actual professional performance. However, if the internship experience includes only routine tasks, it affords little opportunity for the supervisor to judge students' performance on more challenging tasks that may actually better reflect professional practice.

When performance measures are used, only a few situations can usually be assessed, so reliability and predictive validity of results are problematic. Using many performance-oriented assessments (case studies, projects, and internship evaluations) from a variety of sources gives faculty a more complete profile of students' capabilities, strengthening the reliability and validity of resulting judgments.

For further discussion of reliability and validity, see Light and others (1990) and Williams and others (1988).

9.1.4 Norm-Referenced and Criterion-Referenced Testing

Before designing or selecting a test, it is important to know whether results will be used to compare students' performance to that of other students or to determine whether each student has achieved a pre-defined level of performance.

Norm-referenced tests are designed to rank-order students' performances relative to each other. A norm-referenced test should yield a standard distribution of scores, on the

assumption that ability is normally distributed in the subject population. The use of standardized norm-referenced test allows for comparisons between students at different institutions.

Criterion-referenced tests are intended to determine how well students perform relative to the objectives and performance criteria for the course or program. A criterion-referenced test should yield a positively skewed distribution, assuming the test is valid, instruction has been effective, and students have participated actively in the learning process.

The distribution of scores is a joint function of the *difficulty level* of each test item, defined as the percentage of students who respond correctly, and the *item discrimination index*, defined as the degree to which students' performance on each item is consistent with their performance on the overall test (Erwin, 1991; McBeath, 1992). For *norm-referenced* testing, items should be written to yield an average item difficulty of about .5. For *criterion-referenced* tests, item difficulties (percent correct) should be somewhat higher; that is, the majority of students should be able to reach the criterion (usually 80% or 90% correct). If many students fail to reach the criterion, the test, instructional emphasis or methodology, and students' background preparation should be reviewed. Items with moderate difficulties should have a discrimination index of at least .30 (McBeath, 1992).

Test analysis programs can be used to obtain these figures for each item on the test. Such information is useful for validating a test and also for determining whether it should be kept in the item pool, modified, or discarded. For a summary of basic test design and validation procedures, see McBeath (1992).

Generally, program-level measures should be *criterion-referenced* rather than *norm-referenced*, since the purpose is to determine whether the program facilitates students' achievement rather than to compare students' performances to each other. Standardized, norm-referenced instruments are appropriate when comparisons between students at different institutions are desired.

9.1.5 Choosing Standardized or Locally Developed Instruments

An important decision for the assessment of knowledge outcomes is how much to rely on commercially available standardized examinations and how deeply to get involved in developing measures in-house. Each has advantages and disadvantages, summarized in Figure 9.2. Although the costs are not insignificant, faculty involvement in developing outcome measures offers important benefits, for example, enhanced curricular consistency, improved classroom testing, and greater faculty attention to application of knowledge (Banta and Schneider, 1988, pp. 78-79). Experience further indicates that:

...the programs that have invested time and effort in designing their own exams have made the most use of students' scores. These faculties have been more ego-involved in the outcomes of testing, since they made the decision about what content should be tested and by what means. (Banta, 1985, p. 27)

Class assignments and examinations should reflect curricular objectives and incorporate basic test development and validation procedures such as those described below.