



## Methods of assessment

In this third section of the *Inquiry Brief*, the program faculty describes in detail the assessment **methods** cited in the rationale. These are the methods by which the faculty found the evidence that supported, or failed to support, its claims of candidate learning and accomplishment. The particular assessment forms and rubrics the faculty may have developed are presented in Appendix F.

The faculty also describes the research design it has employed to secure the evidence. *Was the evidence based on all the students and graduates of the program? Some representative sample? If it was based on a sample, how was the sample drawn and determined?* The faculty members also describe how the research design addresses rival explanations for the results and how they will address potential aggregation errors and other threats to the validity of their findings.

The methods section also describes any assessments and measures that will provide corroborating evidence for the faculty's main findings and any other evidence that has a bearing on any rival or alternative explanations of their findings. Faculty might show that the sample was truly representative of the program student body, that what look like ceiling or halo effects were really the outcome of a mastery learning regime, etc.

The design of the faculty's investigation must support the faculty's interpretations of the results of its assessment system and the appropriateness of the uses to which it puts them. The faculty members must consider several factors: evidence about the content of the assessments, the assessment criterion relationships, the theoretical and scholarly basis of the construct they assessed, and the uses to which they put the assessments.

In the *Inquiry Brief*, a program faculty will invariably provide evidence of the quality of student learning in the program. Typically, programs use some combination of the categories of evidence presented in the chart following this page. However, each program is encouraged to present novel and tailored evidence of student learning, in place of or in addition to, these categories.

### Qualitative assessments and measures

When a program faculty uses qualitative assessments and measures, those writing the *Inquiry Brief* describe the methods of procuring the evidence and give a rationale for them, just as with any quantitative assessment. The program faculty would present precisely the procedures it employs: for example, team-recorded observations; interview protocols with students, alumni, faculty, administrators, employers; representations of student products or artifacts; interpretations of student journals, lessons, field notes, and audio/video presentations.

### Linking to Quality Principle I

Whether qualitative or quantitative, each source of evidence must have a clear link to a component of *Quality Principle I*. Without such links, the measures may still have value, but only in documenting the context of the program or providing corroboration for subsidiary claims in the *Inquiry Brief*.

### Categories of evidence

Most program faculties actually have a fairly limited number of sources of evidence with which to make their case for the claims about *Quality Principle I*. The types of evidence fall into the following five categories:

1. Course grades
2. Standardized test scores (entrance, exit, and license scores) from the program completers or the graduates' own students
3. Ratings of candidates and program completers (by students, alumni, employers of graduates, portfolios, work samples, cases, impressions, and recollections)
4. Rates of hiring, promotion, certification, graduate study, professional awards, publications, etc., when the decisions are made by third parties in the areas of *Quality Principle I*
5. Case studies of students and alumni competence

## **EXAMPLE: Types of evidence**

### **Grades**

1. Candidate grades and grade point averages in each component of *Quality Principle I*: subject matter; pedagogy; and teaching skill

### **Scores on standardized tests**

2. Student scores on standardized license or board examinations in any of the areas of *Quality Principle I*
3. Student scores on admission tests for graduate study in the areas of *Quality Principle I*
4. Standardized scores and gains of the program graduates' own pupils

### **Ratings**

5. Ratings of portfolios of academic accomplishment
6. Third-party rating of the program's graduates (employers, principals, etc.)
7. Ratings of in-service, clinical, and Professional Development School teaching
8. Ratings by cooperating teachers and college/university supervisors, of practice teachers' work samples

### **Rates which indicate candidate competence**

9. Rates of completion of courses and program
10. Graduates' career retention rates
11. Graduates' job placement rates
12. Rates of graduates' professional advanced study
13. Rates of graduates' leadership roles
14. Rates of graduates' professional service activities

### **Case studies and alumni competence**

15. Evaluations of graduates by their own pupils
16. Alumni self-assessment of their accomplishments
17. Third-party professional recognition of graduates (e.g., NBPTS)
18. Employers' evaluations of the program's graduates
19. Graduates' authoring of textbooks, curriculum materials, etc.
20. Case studies of the graduates' learning

## **Multiple measures**

Because each kind of evidence (grades, surveys, portfolios, standardized tests, etc.) can be misleading, it is important that the faculty commit to include several measures that converge, triangulate, and indicate true student learning. The faculty should also take steps to reduce factors that which affect the validity of the faculty's interpretations. (See Comment, at the end of this section, on issues of reliability and validity.)

At least two measures are generally needed for each component of *Quality Principle I* and the methods of investigating the reliability and validity of the measures must be described and reported.

The methods section of the *Inquiry Brief* gives a complete account of the measures and the faculty's case for the reliability and validity of the measures.

In the case of qualitative measures, the faculty should present the triangulation methods used to reduce error and increase the trustworthiness, dependability, and authenticity of the measures.

## **COMMENT**

### **Validity issues**

There are validity issues for each category of evidence.

**Rates.** Hiring rates, for example, based upon the hiring district's own evaluation of the subject matter knowledge, pedagogical knowledge, and caring teaching skill (*Quality Principle I* components), may not be as much an indicator of student accomplishment in times of teacher shortages, such as are expected in the decade ahead, as they would be in times of teacher oversupply. In times of shortage, hiring rates may indicate

very little about quality because virtually everyone is hired. The rate of first choice hires, for example, may prove to be a more persuasive indicator of student accomplishment.

Similarly, some categories of evidence may be relatively meaningless if the rates are low or less than the normative rates. The rates may indicate something important about the program's quality, however, if the rates are significantly higher than the norm – for example, if nearly all the program's graduates become certified by the National Board for Professional Teaching Standards.

Passing rates on the currently available teaching license tests, for example, are surprisingly high,<sup>1</sup> but some passing-scores are set as low as the 25<sup>th</sup> percentile of actual cohort performance and with fewer than half the test's items answered correctly in some cases. Retention, program completion, and graduation rates average 50 percent in most cases.

Rates have meaning in the TEAC framework only if they are based upon an evaluation by a third-party of some aspect of *Quality Principle 1* that also provides for normative comparison.

**Survey data** particularly that derived from survey forms created by those without special expertise in instrument development are known to be affected by a number of extraneous factors. For example:

- the order in which questions were presented,
- the context in which questions appeared,
- whether the questions weed out those with no opinion (filtering),
- the range and order of choices,
- whether middle categories were provided, and
- whether the format was open or closed.

Survey results need to be examined for their reliability and validity, as do course grades.

**Course grades** are meant to be a measure of subject matter understanding, but their validity is threatened by the fact that they are frequently measures of other matters that may have only a tangential or no relationship with the student's mastery of the subject matter of the course.

Some of the common threats to the validity of course grades occur when they become influenced by other factors and become as a result measures of these other factors. In contemporary higher education, it is fair to say that grades may be, in varying degrees, measures of any, or all, of the following:

Punctuality: when faculty members take points off for late work or give extra points for early work

Gain or growth: when faculty members base the grade on the degree of improvement over the course of the semester

Place in a distribution: when faculty assign grades on the curve, or some predetermined percentage formula, so that the grade indicates only the student's percentile or rank in the class

Dishonesty: when faculty or the university lower the grade for cheating, plagiarism, etc. with the result that a low grade is uninterpretable because it may signify a low level of understanding or a low level of honesty

Extra or additional achievement: when faculty give extra points for more work that may not be qualitatively superior to the prior work, but is simply quantitatively more than other students have done

Attendance: when faculty members deduct points for unexcused absences

Writing skill: or some prior expertise separable from the subject matter as when neatness, rhetoric, or format count

---

<sup>1</sup> Pass rates of 100 percent are becoming common, but many programs achieve them by using the state's license test as a program admission test or screening test for latter stages of a program. High pass rates in this instance are of little use as indicators of program quality.

Reduced spread: when faculty members inflate the grades or reduce the variance (as in the quip, “the best way to turn C students into B students is to put them in graduate school”)

Motivation and perseverance: when students receive the last grade of several unsuccessful attempts at the subject matter, or when effort is rewarded

Group membership: when faculty members introduce examples and analogies that speak to some groups of students more than others, or when there is cultural, racial, or gender bias in the teaching format

Political statement: when faculty are sensitive to the student's military draft or immigration status, scholarship and grant conditions, graduate or undergraduate status, race, and gender, etc., and take these into favorable or unfavorable consideration in the assignment of course grades

The inference that grades, or any other measures of learning, are valid can be based on a number of considerations and investigations:

Are the grades the faculty members give consistent and correlated with other known measures of student learning (e.g., standardized tests of the same content)?

Are they based on the appropriate content so that they measure only what they are supposed to measure?

Are they correlated with and predict later accomplishment that depends on student learning?

Are they related to other factors that one would expect, in theory, to be related to what the grade measures (e.g., intelligence, prior grades, aptitudes, specialty training, beginning or end of the program accomplishment, motivation)?

In general, the correlations about .50 provide confidence that the measure is valid for the purposes to which it is put.

### **Reliability issues**

An investigation of the *reliability* of course grades or any other quantitative measure of student learning might entail the following:

The computation of an *alpha* or *kappa* coefficient when the grades are thought to be measuring a single attribute.

Correlations between two different administrations of a test that determined the grade;

- or between even and odd items on the test;
- or between the first and second half of the test;
- or a correlation between equivalent versions of the test;
- or the stability of the mean grades and standard deviations across several administrations of the test to comparable groups;
- or published reliability statistics from test manuals.

Along the same lines, faculty members might explore the reliability of their grades through correlations of the grades from each half of the transcript for a random sample of students; or correlations between grades in the same course in two semesters from a sample of professors.

Or they might examine whether the variance in the distribution of a faculty member's grades (0-4), or the variance in the average grade in selected courses, is contained within one point or a letter grade.

In general, correlations about .80 yield confidence that the measure is trustworthy and dependable.

**EXERCISES** on Reliability, Validity, and Organizing Data  
from *TEAC Exercise Workbook, 2010*  
pages 63-66

**Exercise 43: Evidence of reliability:** Which of the following approaches would yield evidence that the faculty would find compelling about the reliability of the evidence for the claim, “*our students know how to apply technology in the classroom*”? Circle the number(s) of the approaches that your faculty would find credible.

1. For a 10 item rating form completed by methods instructors, a coefficient alpha is provided, with a value of .82.
2. The faculty observes that the means of a 10 item rating form completed by methods instructors across four sections of the course are almost identical.
3. Two methods instructors rate a sample of students in the program independently, and the level of agreement between the ratings is perceived to be high.
4. The level of agreement of the two methods instructors cited in option 3 above is assessed with a correlation coefficient – and is found to be .85.

List other evidence that would convince the faculty that the measures were reliable.

**Exercise 44: Validity:** The faculty is interested in knowing whether the 10-item scale used to assess the program’s claim concerning technology was valid as a useful tool to verify the claim. Circle the number(s) of the approaches for assessing validity that your faculty would find credible.

1. Since the measures were found to be reliable, the issue of validity is no longer relevant. If the measures are reliable, they are surely valid.
2. The students’ scores on the ten-item scale on technology are correlated with the ratings they received in student teaching on “uses technology effectively.” The correlation between these two measures is .75.
3. The faculty reviewed the ten items on the technology scale and determined that the items covered all of their intentions about what students should learn about technology in their program. The scale was judged to have content validity.
4. The ratings on the scale discriminated between those students who used technology well in student teaching and those who did not – a finding yielded by a discriminate analysis of the evidence.

List other approaches that would yield evidence that the faculty would find compelling about the validity of the ten-item scale.

**Exercise 45: Measures truly relied on:** Review the following novel and idiosyncratic measures uncovered in TEAC audits and consider the evidence upon which the program faculty truly rely:

- Candidates equal or exceed majors in grades in the disciplines (teaching subjects)
- Faculty noted the exceptionality of those as students who later were board certified
- High faculty agreement in rating quality of random samples of students by name only
- A&S departments hire candidates as graduate teaching assistants (GTAs)
- Local superintendents waive interviews for recommended students
- Higher state scores in schools with higher densities of program graduates
- Candidates are the first choice and accept their first choice in employment
- Candidates are first choice of cooperating teachers for student teaching assignments
- Lawful patterns of correlations among internal and external measures of the available measures of competence

- Work samples with student/pupil learning data
- Authentic artifacts (viz., technology, video)
- Comparisons of retention of program's students in teaching with other programs
- Regents or NAEP examination scores for candidates
- Reporting assessments at various stages in the program to show reductions in variance over time
- On-demand ratings by faculty of students, video-taped lessons show lawful correlations with internal & external measures
- Pupil evaluations of student teachers

**Exercise 46: Organizing your data:** With your colleagues, try organizing a spreadsheet like the one below for a sample of your students. Fill in the column headings for as many data sources as you have.

Each row contains the data for one and only one unique student in your sample. Each column contains something you know about your students that is important to the quality of your program. (Example data sources are provided below.) The cells in the spreadsheet contain information (qualitative and/or quantitative) about each student.

**Student characteristics**

Student	Year	Option	Level	Gender	Race	Major	Site	Etc.
1.								
2.								
3.								
N								

**Admissions indicators**

Student	SAT score	ACT score	Rank in H.S.	H.S. Grades	Interview	Writing Sample	Etc.
1.							
2.							
3.							
N							

**Grade point indices**

Student	GPA in methods	GPA in major	GPA in clinical	Grades in techn.	Grades in MC	Etc.	
1.							
2.							
3.							
N							

**Local program measures and ratings**

Student	Field experience	Coop. teacher rating	College supervisor rating	Portfolio artifacts	Self-ratings	Etc.	
1.							
2.							
3.							
N							

**License tests and other external measures**

Student	Praxis I	Praxis II	GRE	Etc			
1.							
2.							
3.							

N							
---	--	--	--	--	--	--	--

**Post-graduate and employer surveys (and the like)**

Student	Rating of prgm.	Rating of courses	Rating of faculty	Years teaching	Employer Rating	Pupil state tests	Etc.
1.							
2.							
3.							
N							