Principles for Measures Used in the CAEP Accreditation Process

Peter Ewell

May 29, 2013

The purpose of this document is to present a set of principles to be used by the CAEP Commission and the Data Task Force in evaluating proposed measures associated with the draft Standards now under review. Fifty-nine measures have been proposed, grouped under the five standards and as a free-standing group of eight measures to be reported annually to CAEP by accredited programs. The Data Task Force is charged with prioritizing these measures and assessing their suitability for use in making accreditation decisions. These principles were designed to provide guidance in this task. They may also be of benefit to the leaders of teacher preparation programs in determining what kinds of evidence to use for program review and improvement.

The fifty-nine measures presented under the CAEP standards and the eight annual indicator measures are of quite different kinds. As a result, they have different properties and applications that affect how the principles are applicable. Types of measures include examinations, surveys, observations, programmatic measures, curricular features, and case studies.[1]

The following principles are offered to provide guidance for determining the adequacy of measures proposed for use in the CAEP accreditation process.[2]

1. Validity and Reliability. All measures are in some way flawed and contain an error term that may be known or unknown. In general, the greater the error, the less precise—and therefore useful—the measure. But the level of precision needed depends on the circumstances in which the measure is applied. To be used in accreditation decisions, measures need to be founded upon reliable measurement procedures, but they also need to be designed to operate under less-than-ideal measurement conditions. Many of the measures used to assess the adequacy of teacher preparation programs such as licensure examination scores meet these rigorous standards but many of the more qualitative measures proposed do not. Even the most rigorous measures, moreover, may not embrace the entire range of validities—construct, concurrent, and predictive. "Validity," moreover, as its meaning has evolved has come to embrace the appropriateness of the use to which the measure is put ("consequential validity" as in Messick, 1995). This means that Value Added Measures (VAM) need to be carefully evaluated because their original application was to examine classroom and school performance, not the effectiveness of the programs that prepared the teachers in those classrooms and schools.

2. Relevance. The measures advanced ought to be demonstrably related to a question of importance that is being investigated. This principle implies validity, but it goes beyond it by

[1] Appendix A provides fuller descriptions of these different types of measures.
[2] Appendix B provides examples of the application of these principles to three different measures—licensure passage rates, completer employment rates, and case studies.

also calling for clear explanation of what any information put forward is supposed to be evidence of and why it was chosen. This principle also implies that there is a clear and explicable link between what a particular measure is established to gauge and the substantive content of the Standard under which it is listed.

3. Verifiability. The validity of any measure advanced should be subject to independent verification. This partly a matter of whether the process of creating the current value of the measure is replicable, and if repeating the process would yield a similar result. This principle implies reliability, but goes beyond it to require transparency and full documentation—whether sufficient information is available to enable any third party to independently corroborate what was found.

4. Representativeness. Any measure put forward should be typical of an underlying situation or condition, not an isolated case. If statistics are presented based on a sample, therefore, evidence of the extent to which the sample is representative of the overall population ought to be provided, such as the relative characteristics of the sample and the parent population. If the evidence presented is qualitative—for example, case studies or narratives, multiple instances should be given or additional data shown to indicate how typical the examples chosen really are. CAEP holds that sampling is generally useful and desirable in generating measures efficiently. But in both sampling and reporting, care must be taken to ensure that what is claimed is typical and the evidence of representativeness must be subject to audit by a third party.

5. Cumulativeness. Measures gain credibility as additional sources or methods for generating them are employed. The resulting triangulation helps guard against the inevitable flaws associated with any one approach. The same principle applies to qualitative evidence whose "weight" is enhanced as new cases or testimonies are added and when such additions are drawn from different sources. In sum, the entire set of measures used under a given Standard should be mutually reinforcing.

6. Fairness. Measures should be free of bias and be able to be justly applied by any potential user or observer. Potential sources of bias might be introduced by the values or beliefs of those applying the measure, such as the conviction that a particular result should be observed. Other sources of bias are situational, such as the limited perspective of an untrained observer undertaking a classroom observation or applying a rubric. In this sense, fairness is a special case of reliability: a fair measure will return the same result even if applied by different observers under different circumstances or at different points in time.

7. Stakeholder Interest. A sound set of measures should respect a range of client perspectives including the program, the student, the employer, and the state or jurisdiction. Taken as a whole, a set of measures should potentially support the establishment of an informed dialogue among the appropriate parties. A statistic on the employment rates of program completers, for example, can be summarized from the student point of view as the probability of being placed,

from the program's point of view as a placement rate, and from an employer's point of view as the proportion of needed job openings filled each year. To reflect stakeholder interests, moreover, proposed measures should be neither arcane nor overly academic.

8. Benchmarks. Without clear standards of comparison, the interpretation of any measure is subject to considerable doubt. Measures can be compared across programs, against peers, against established "best practices," against established goals, against national or state norms, or over time. For every measure under each Standard, CAEP should be able to indicate an appropriate benchmark against which a given program's performance can be judged.

9. Vulnerability to Manipulation. All measures are to some extent vulnerable to manipulation. This is one reason to insist upon triangulation and mutual reinforcement across the measures used under each Standard. For example, program graduation and licensure passage rates depend a great deal on which students are included in the denominator. Because the incentives to perform well on such measures are considerable, programs may identify ways to construct these denominators that yield maximum values on these measures regardless of what they are actually doing.

10. Actionability. Good measures, finally, should provide programs with specific guidance for action and improvement. Many promising measures fail simply because they are too expensive, too complex, too time consuming, or too politically costly to implement. Often, the simplest are best, even if they seem less technically attractive. Value Added Measures, for example, are conceptually compelling, but they demand a good deal of investment to work and interpret. This principle also suggests that any measure should be able to be disaggregated to reveal underlying patterns of strength and weakness or to uncover populations who could be served more effectively. Finally, the measures provided should be reflectively analyzed and interpreted to reveal specific implications for the program.

Appendix A

Types of Measures

The fifty-nine measures presented with the CAEP draft standards are of different kinds, as described below:

- Examinations. Prospective teachers take examinations in the course of their training and in order to be licensed to practice. Dimensions of interest for examinations include their content coverage, the level at which this content is tested, the depth at which various kinds of knowledge is probed (which affects the duration of the examination), how items are structured (e.g. constructed response or multiple choice), and whether or not responses can be compared across test-taking populations (degree of standardization). The results of examinations can be reported on an absolute basis or in the form of Value-Added Measures (VAM).

- Surveys. Students in teacher preparation programs are frequently surveyed as they progress and after they are employed. Dimensions of interest for surveys strongly resemble those for examinations except that items are self-reported. Another important coverage dimension involves the extent to which survey items are directed at actions or behaviors, or are self-reports on knowledge or skill outcomes. This is important because students are generally more accurate commentators on the former than the latter. Surveys are also administered to the employers of teachers and to their students to help provide evidence of the effectiveness of teacher training.

- Observations. Observations of teacher candidates in field placements and of newly employed program graduates are also used as quality measures. Dimensions of interest parallel those of surveys but employ an element of peer judgment embodied in a trained observer using a well-developed observational protocol.

- Statistics. Various behavioral statistics are used as outcome measures for teacher training programs. The most prominent examples are completion rates and job placement rates. Dimensions of interest include the outcome of interest (the numerator), the population to which the rate applies (its denominator), and the time period over which the calculation is run (e.g. “one-and-a-half times catalog length of program” or “within one year”).

- Curricular Features. The CAEP standards address various aspects of teacher training curricula, so some of the proposed measures address descriptive aspects of the programs themselves such as the extent to which students are taught assessment methods (e.g. 1.g.) or reflect upon professional expectations (1.h.). Dimensions of interest here are the aspect of the program in question and the extent or duration of coverage.

- Case Studies. Where quantitative measures are unavailable, the CAEP Standards call for qualitative investigations termed "case studies" (for example, "case studies of districts in which a large number of program graduates are employed"). Dimensions of interest include the question to be investigated through the case study, baseline conditions, the intervention or phenomenon of interest, the goal of the intervention, observed results, and implications for action.

Appendix B

Application of Principles

To illustrate how these principles can be used, this appendix applies each of the principles to three measures: licensure passage rates, employment rates, and case studies of districts where a large number of program graduates are employed.

Validity and Reliability

- Licensure Passage Rates. Although state licensure examinations differ by provider, the two main test vendors have established rigorous standards of test development and regularly report statistics on validity and reliability. As a result, this measure fully meets the principle.

- Employment Rates. These are also well defined measures that are valid and reliable so long as they are properly calculated. Because of the latter, they should be examined carefully for threats to validity and reliability such as exclusions from the denominator or changed conditions over repeated annual measures.

- Case Studies. Case studies are a bit more problematic with respect to this principle because their validity depends on the type of information collected and the extent to which the same procedures are used across cases and over time. This will naturally be a peer judgment.

Relevance

- Licensure Passage Rates. Insofar as licensure tests faithfully reflect content knowledge and knowledge of pedagogical practice, program completers' performance on them constitutes relevant information about the quality of the program.

- Employment Rates. Many things can affect employment rates that are not directly related to program quality including local job market conditions or the general state of the economy. As a result, this measure does not fully meet the principle.

- Case Studies. So long as the topics addressed in the case study are selected to reflect important dimensions of program performance—for example, the ability of program graduates to effect learning growth in their pupils or their demonstration of ethical and professional practices—the principle of relevance is met.

Verifiability

- Licensure Passage Rates. Licensure examinations are designed and administered by testing organizations that can be queried about the way these examinations operate. Similarly, the state authorities responsible for administering them can be more or less transparent in how they calculate passage rates. Both of these will affect verifiability.

- Employment Rates. Under ideal conditions, these are collected by state agencies on behalf of EPPs using surveys or unit record wage databases using known calculation rules. If these rules are documented, states' application of them can be independently audited to determine compliance, thus rendering them verifiable. Where surveys are conducted by EPPs themselves, such documentation may or may not be present.

- Case Studies. The contents of case studies can also, in principle, be audited if their construction is fully documented. But this is a good deal more complicated than for program completion rates. As a result, case studies will probably not fully meet the principle.

Representativeness

- Licensure Passage Rates. The examined population is the population of interest with the denominator representing the parent population. So long as these populations are correctly constituted, the measure is representative.

- Employment Rates. These will typically be representative but when measured over time, they may not be if labor market conditions change.

- Case Studies. These must be carefully examined to determine if the districts chosen are typical of many districts to which the program sends graduates as employees. This could be done if the requisite documentation was supplied.

Cumulativeness

- Licensure Passage Rates. Although these are precise, they are sole source measures so there is not much ability to triangulate results. One indication of "weight of evidence" might be recording performance over time or correlating test score performance with other evidence of academic achievement on the part of program graduates such as grades in course examinations or portfolios.

- Employment Rates. The same situation largely applies here as to licensure passage rates. A possible exception is if this information is collected both by surveys and by tapping wage record databases.

- Case Studies. Case study results will be bolstered by the presence of other measures such as teacher observations or student surveys that show similar conclusions. They also can be examined for consistency over time. Finally, the construction of case studies themselves is important: if they involve mutually reinforcing lines of inquiry or investigation, their credibility is strengthened.

Fairness

- Licensure Passage Rates. These are supplied by third parties, so there should normally be little opportunity for bias. Where this could occur is when EPPs themselves report these rates.

- Employment Rates. If states collect these data on behalf of EPPs using surveys or by tapping wage record databases, the measure should be unbiased. Again, if EPPs themselves conduct the surveys, bias could enter.

- Case Studies. Because they are conducted by EPPs entirely and are used to advance a particular quality claim, case studies are unlikely to be entirely unbiased.

Stakeholder Interest

- Licensure Passage Rates. These can be looked at from a number of stakeholder perspectives including the candidate who wants to know her or his chances of success, the program as a measure of the quality of preparation in the areas tested, and the employer who wants to know how well potential employees are prepared.

- Employment Rates. The case is similar to licensure passage rates.

- Case Studies. Stakeholder interest for case studies will vary enormously depending upon the information collected. So determining the extent to which the principle is met would entail a systematic review of the contents of the case study from the point of view of each stakeholder.

Benchmarks

- Licensure Passage Rates. These are all publicly reported as required by Title II but differences in state cut scores for the ETS examinations renders meaningful comparisons difficult except within a given state. If all results were calculated using a standard cut score, benchmarking would be possible in these states. In the states where licensure examinations are provided by a non-ETS provider, within-state comparisons are possible but cross-state comparisons are not.

- Employment Rates. Benchmarking is completely possible if all programs use the same calculation rules in constructing the statistic.

- Case Studies. Comparative benchmarks are in principle possible here but comparisons of qualitative studies are much more difficult to accomplish. So this principle will be less fulfilled for this measure than for most others.

Vulnerability to Manipulation

- Licensure Pass Rates. These are calculated and reported by third parties so programs have no opportunity to tamper with them except by restricting who is tested, which would violate representativeness.

- Employment Rates. These can be subject to manipulation through exclusions from the denominator, which should be carefully checked for.

- Case Studies. Because they are entirely qualitative, case studies are vulnerable to bias in the evidence that is presented and discussed. For this reason, they will likely never fully meet the principle.

Actionability

- Licensure Passage Rates. The ability to take action on the results of this measure depends a good deal on the amount of information on test performance that is available. If sub-scores on these examinations are provided, there is some diagnostic information to inform action. Similarly, disaggregating the tested population to determine who passed and who did not can aid intervention.

- Employment Rates. The case here is similar. Disaggregating the population can determine who is not completing and this would aid intervention. There is no analog to sub-scores for employment rates, but some information on when and in what jobs and circumstances graduates obtain employment might inform action.

- Case Studies. Actionability will depend entirely on the contents of the case and how thoroughly this is discussed and actionable implications drawn. Actionability can be aided by constructing the case study in a way that explicitly emphasizes actionable conclusions.