CALIFORNIA SOCIAL WORK EDUCATION CENTER
UNIVERSITY OF CALIFORNIA, BERKELEY
SCHOOL OF SOCIAL WELFARE

Using the Results from the California Common Core Evaluations: What Training Evaluation Data Can (and Can't) Tell Us - JUNE 2008 -

Making valid inferences from tests:

Tests are not valid in and of themselves. Test results are valid for specific purposes and certain kinds of decisions. What is a "valid" decision depends on the purpose the test was designed to serve.

What can the Common Core Evaluations tell us?

The tests included as part of the Common Core have been designed to provide feedback for program improvement and evidence of training effectiveness for trainees as a group. They are also part of establishing a standard method of evaluating training effectiveness in response to Federal requirements in the Program Improvement Plan (PIP) for California. The Common Core Evaluations can tell us:

- Whether or not trainees *as a group* improved their knowledge of key concepts following training (from pre to posttest).
- Whether trainees *as a group* understood key elements related to recognizing physical abuse and sexual abuse, and could make a valid decision based on their analyses of the elements present in a given case.
- Whether or not the course is at the right level of difficulty and whether or not key concepts are being covered adequately and consistently (when combined with demographic information and other sources like participant comments and observation by subject matter experts).

What can't the Common Core Evaluations tell us?

When building a test, the test author(s) can't ask every question that could possibly be asked about the content of the training, just as curriculum writers can't cover everything there is to know about child welfare practice in classroom training. There is always a trade-off between the time available for the test and how completely all of the training content is covered.

A test is a *sample* of the many possible questions that could be asked. A few notes about tests:

- Tests reflect people's knowledge and ability, but also have some random error. For example, people get some items right by guessing, and get some items wrong through fatigue or carelessness. The more items are included on a key concept, the less likely it is that someone would appear to understand based on a lucky guess, or not understand based on careless errors. Thus, the longer the

test and the more items per concept, the more precise and error-free the estimate of a person's ability will be.

- Tests designed to make decisions about program effectiveness (and the knowledge of trainees in general) have fewer items and may have only one or two items per concept. If an estimate of an individual's knowledge is a little bit off, it is not as critical to the way these tests are used, as it would be if the person might be fired or have to go through additional training based on a low score.

- Tests where there might be personnel consequences are referred to as high stakes tests. They are typically quite long (as long as 3 hours for a 1.5 day training), and must go through a rigorous development and validation process in order to withstand legal challenges. These tests must meet standards such as those published by the American Educational Research Association, American Psychological Association and National Council on Measurement in Education, and the Uniform Guidelines on Employee Selection Procedures (EEOC). There also are a number of policies and procedures that must be in place for test administration and use, including:

  - Notice of testing and due process procedures
  - Policy and procedures for accommodating persons with disabilities
  - Policy and procedures for test security and identification of test takers
  - Policy for repeat testing
  - Policies on record keeping and reporting

  The Common Core tests are being developed to meet professional standards of reliability, validity, item functioning and fairness. However, they currently are not designed to have the extensive content coverage and numbers of items that would be required to make judgments about an individual's knowledge. **Therefore, the results of these tests should NOT be used to 1) make inferences about individuals' knowledge and skills or 2) make personnel decisions.**