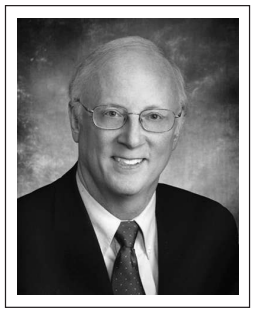

You Can't Squeeze Blood out of a Turnip: What Diagnostic Testing Is —and Isn't

Stuart Kahl, Ph.D.
CEO
Measured Progress

With local educators feeling greater pressure than ever to raise student achievement levels, there is a parallel increase in demand for diagnostic information from testing. Unfortunately, many are looking for such information from tests that were not designed to be diagnostic. "Diagnosis" is defined as the identification of the nature and cause of something. A physician diagnoses illness, while an information technologist diagnoses problems with software code or some other component of a computer system.

In education, clinical testing can diagnose cognitive or learning disabilities for individual students. However, non-clinical, diagnostic testing sheds light on specific concepts or skills that a student is having difficulty learning and determines "why"—possibly pinpointing prerequisite concepts and skills that may need to be re-taught. Two issues of concern with this latter type have to do with the specificity of the "diagnostic" information and the validity of inferences that can be drawn from it.

Growing emphasis on statewide summative tests (as well as interim tests that look, smell, and feel like the state tests), coupled with concerns about over-testing and the loss of instructional time, have led teachers to look to such tests for diagnostic information. While the tests have several important uses, diagnosis as described above is generally not one of them. Most of the tests were designed to produce total test scores, along with a few student subtest scores.

A subtest might represent only a quarter or fifth of the larger test, with items representing only a sparse sampling of a very broad subdomain (e.g., geometry/measurement in mathematics). Thus, subtest scores are not particularly reliable and no more diagnostic at the student level than a thermometer is for a medical diagnosis—it can indicate something's wrong, but it doesn't tell you what.

Summative tests should be used to raise programmatic questions that require further investigation: Why is one subgroup of students performing lower than another? Why are students performing more poorly than expected in a particular subdomain? Further investigation could involve discussions with teachers and students, review of curricula, and more testing. These tests can raise similar questions about individual students, but because inferences regarding individual students are even less reliable, deeper probing is absolutely necessary in order to acquire useful diagnostic information.

Educators may try to squeeze diagnostic information out of general achievement measures by over-interpreting single item results and speculating, based on probabilities, on hypothetical performance by students on test items unrelated to ones they actually took. While these approaches can be useful in a discussion of the general strengths and weaknesses of a large group of students, they would be no more accurate than a coin toss for individual student diagnoses.

In the next few years, I expect we're going to see more and more testing and greater and greater emphasis on test results. It is critical that educators and non-educators, parents and policy makers alike, become more assessment literate and therefore more likely to use the right assessments for the right reasons, avoid misdiagnosis, and assure that testing will be as effective as it can be in support of teaching and learning.

What do you think?
Let us know at twocents@measuredprogress.org