# *The Disconnect Between Recent Research and Education Policy: Basing High-Stakes Decisions on Tests That Have Not Proven to Be Valid - Abridged*

## Introduction

High stakes tests have been a fixture in American education for many years. With the initiation of the No Child Left Behind (NCLB) legislation, such testing at the K-12 level increased substantially. While many states developed their own tests, some turned to standardized testing companies like ACT to provide them with components for a state test.

According to the ACT website, there are 22 states with state-funded ACT partnerships, ten states that have almost universal ACT testing (one with 97%, the rest with 100%), and five states offering the ACT and/or the WorkKeys as components of their official NCLB test. Almost 20% of ACT volume is from state testing.

Due in part to such partnerships, ACT's expansion during the past ten years has been substantial. The use of the ACT has expanded greatly, with 1,666,017 Class of 2012 graduating seniors taking the test, an increase of 49.2% from the Class of 2002. From 2004 to 2009 (the latest year for which data are available), ACT's revenue from its testing programs and related services increased by over 63%, from approximately $155,000,000 to $253,000,000. College Board's SAT's expansion has been much more modest, though still substantial, with 24.4% Class of 2012 more graduating seniors taking the test (1,664,479) than for the Class of 2002 (1,337,999).

The community colleges provide another standardized test stronghold. Comparable use trends over time are not published for the ACT-produced COMPASS and College Board's ACCUPLACER. However, their use is very widespread. According to a 2008 survey of assessment practices, 62% of community colleges use ACCUPLACER, and 46% use COMPASS (some institutions use both, depending on the subject).

Similarly, no expansion or sales volume data are available for the WorkKeys post-secondary workforce readiness test. ACT did report that they have issued 1.5 million National Career Readiness Certificates and have "statewide or regional credentialing initiatives in more than 40 states." The certificates are issued for passing all three WorkKeys tests at a certain level. In fact, the ACT website recently announced that "The American Council on Education (ACE) has recommended that postsecondary educational institutions award college credit in applied critical thinking to individuals who earn the ACT National Career Readiness Certificate (NCRC™). ACE recommends that three credit hours be awarded to recipients who earn specific levels of the credential."(i.e., silver, gold, or platinum score levels). (ACT National Career Readiness Certificate Recommended for Three Semester Hours of College Credit). Inasmuch as SAT and ACT performance alone does not alone earn college credit, such a recommendation is quite unprecedented.

So the college placement companies, particularly ACT, are experiencing substantial usage and growth from secondary and post-secondary institutions, and have sufficient credibility to warrant the expenditure of at least hundreds of millions of dollars. Yet several national institutions with strong independent credentials have conducted and published studies in the past two years which would call into question the current usage of such tests. This report will review the major independent studies focused on the ACT and COMPASS/ACCUPLACER tests in the past two years. It also will examine the research behind WorkKeys. Since SAT has not been the subject of such recent independent studies of its validity for placement use, it will not be a focus of this report. Also, vendor studies will not be considered, since vendor studies conducted to promote vendor products are inherently biased and not very credible - no vendor will knowingly publish a study negative to their product.

## ACT

The ACT test was the subject of a large research study produced by the National Bureau of Economic Research (NBER). "Improving College Performance and Retention the Easy Way: Unpacking the ACT Exam" was released in June 2011, conducted by Eric Bettinger and Brent Evans of Stanford and Devin Pope of the

University of Chicago. Their abstract stated that: "Colleges rely on the ACT exam in their admission decisions to increase their ability to differentiate between students likely to succeed and those that have a high risk of under-performing and dropping out. We show that two of the four subtests of the ACT, English and Mathematics, are highly predictive of positive college outcomes while the other two subtests, Science and Reading, provide little or no additional predictive power." Also, "By introducing noise that obscures the predictive validity of the ACT exam, the reading and science tests cause students to be inefficiently matched to schools, admitted to schools that may be too demanding -- or too easy -- for their levels of ability."

ACT's surprising response, as reported in EdWeek, came from, Jon Erickson, the current (then interim) president of ACT's Education Division, who stated that: "ACT is an achievement-based test that is used for multiple goals and purposes beyond just admissions or predicting overall student success, such as college GPA or retention.". Erickson's dismissing of the role of the ACT in "admissions or predicting overall college success" is surprising, since ACT's own website describes the ACT test as "the college admissions and placement test."

The most important validation for the NBER study came, ironically, from ACT's own research. "ACT College Readiness Benchmarks, Retention, and First-Year College GPA: What's the Connection?" was a 2005 study conducted by ACT in support of their benchmarks. It unintentionally cast doubt on the value of the benchmarks by showing how most students who did not meet them persisted successfully in their first year of college. But more importantly, it showed that meeting the science benchmark was a weak predictor of college persistence and GPA, and it left out references to the reading benchmark entirely.

Reading is, of course, critical to academic success. But this does not mean that a test labeled "reading" is necessarily a good measure of reading. In fact, reading already is measured by all ACT subject tests. And science is much too broad a field to be comprehensively tested in 35 minutes.

One study alone, no matter how large, cannot refute the value of a long-established test. But when ACT's own research validates that study's findings, and when ACT's own president of Educational Services acknowledges the limitation, and when other studies also validate the findings, then the test should not be considered suitable for either college admission screening or statewide student assessment.

**COMPASS/ACCUPLACER**

The community college placement tests from ACT and College Board have been the subject of two major studies released in February, 2012 from the Community College Research Center at Columbia University's Teachers College. Each study found that more than one-fourth of the students placed in remedial classes due to the use of the two tests should have been able to pass college-credit bearing courses with a grade of B or better.

From the conclusion of Judith Scott-Clayton's "Do High-Stakes Placement Exams Predict College Success?" (based upon the COMPASS test): "…overall the correlation between scores and later course outcomes is relatively weak, especially in light of the high stakes to which they are attached. Given that students ultimately succeed or fail in college-level courses for many reasons beyond just their performance on placement exams, it is questionable whether their use as the sole determinant of college access can be justified on the basis of anything other than consistency and efficiency." Additionally, she commented that: "...the incremental validity of placement tests relative to high school background predictors of success is weak, even in math."

From the abstract of Clive Belfield and Peter Crosta's "Predicting Success in College: The Importance of Placement Tests and High School Transcripts" (based upon both the COMPASS and ACCUPLACER): "We find that placement tests do not yield strong predictions of how students will perform in college. Placement test scores are positively—but weakly—associated with college grade point average (GPA). When we control for high school GPA, the correlation disappears...In contrast, high school GPAs are useful for predicting many aspects of students' college performance."

Following the media publicity over the studies, representatives from both ACT and College Board acknowledged that misplacement is likely if the schools use only the test score without additional information from other academic indicators, particularly high school grades, and that preparation for the tests will reduce misplacement. Unlike the ACT and SAT, the COMPASS and ACCUPLACER tests are often used with minimal consideration for a student's high school grades. Additionally, they often are used in conjunction with the ACT and SAT, with the latter tests used as a screener for requiring the taking of the COMPASS or ACCUPLACER. Given the strong relationships between performance on the general college admissions test and the community college test (e.g., students who do well on the ACT do well on the COMPASS), it is very unlikely that the latter will provide a significant level of additional information over the former.

So while there generally is a positive relationship between community college classroom performance and COMPASS or ACCUPLACER scores, the relationship is weak, and may cease to exist if prior grades already have been considered. The test should not be used as the sole or even primary or indicator for placement. The tests predict too little future student classroom performance and are much less predictive than prior student classroom performance. And there is no point to using them after an ACT or SAT has been administered.

**WorkKeys**

WorkKeys research has been substantially more limited, with no major independent research entities conducting prediction research. Almost all studies that have been found have been limited to content analyses and ACT-sponsored/partnered case studies.

All major tests are subject to content validity analysis in the course of their development. Both ACT and the American Council on Education (ACE) conducted such analyses on WorkKeys. While this is useful and necessary for test development, it is no substitute for seeing if the test actually works in practice.

The WorkKeys content is supposed to measure skills that are more focused upon the workplace. The items are workplace-oriented, e.g., test questions related to ringing up a sale in a store. But the "common sense" attitude that such test question will predict workplace performance better than, for example, correctly identifying a number sequence may be totally flawed. For example, the ACT math test predicts science classroom performance far better than the ACT science test. Yet the science test items are more closely related to science classroom activities. Such common sense-defying realities of test item validity are one reason why prediction studies are done.

ACT also has showcased an array of its own case studies from its product users. While these technically are prediction studies, they also are marketing materials. As mentioned in the introduction, vendor studies are not very credible - a vendor will not seriously criticize its own product.

A potentially independent prediction study of WorkKeys was found, but it was not at all a research study in the scientific sense. There may be other such studies out there, but apparently none with the institutional credibility or scientific protocol of the NBER or Columbia studies.

Human behavior is hard to predict, particularly by using a relatively brief test. The best ACT subject tests (English and Math) predict only a small percentage of the variance in future classroom performance. Workplace behavior is even harder to predict than future classroom performance, since classroom performance already is in part measured by test performance, and workplace behavior generally is not. So test items relating to employment may only predict a tiny percentage of the variance in important workplace behaviors.

There probably is a positive relationship between workplace performance and WorkKeys performance. Any systematic use of a measuring tool such as a test should improve job placement. If a workplace had no other measures, using WorkKeys probably would be useful. But many factors can predict future workplace performance. The important questions to ask are "how much can they predict" and "can another predictor do better?"

Anecdotes, test content validations, and case studies from vendors will not suffice. We still do not know with any certainty how good WorkKeys is, how much workplace behavior it predicts, and whether it is any better than already available information; e.g., ACT tests and high school grades in Illinois. We unbiased scientific research studies to provide such answers.

**Conclusion**

Standardized tests can be very useful in identifying the overall performance of students at a school. We can be rather certain that students at a school with a composite average ACT score of 25 will be higher performing overall than students at a school with 20. We can even assume that it is likely (though not as certain) that an individual student with an ACT composite score of 25 will do better in college coursework than one with a score of 20. However, recent research, which clearly is supported by ACT's own research, shows that some ACT tests (math and English), are useful predictors of college performance, and some (reading and science) are not. Given the expanding use of the ACT reading test in state assessments under NCLB, there is a serious disconnect between the research and state policy. Use of the reading test (rather than, for example, the English test) for such purposes is not supported.

Recent studies (as well as past ones) also clearly demonstrate that the COMPASS and ACCUPLACER are only minimally useful, and useful only as an adjunct to prior classroom performance. Their current overly-frequent use as a single indicator of college readiness is highly inappropriate and unsupported by any research. Institutions also need to examine whether other already-administered tests (whether currently administered ACT's or SAT's, or a future Common Core assessment) would be equal or superior to the task.

Finally, the validity of WorkKeys in terms of screening students for the workplace is unsupported except by superficial and potentially biased research. Unbiased scientific research is needed on whether the test has sufficient value in predicting workplace performance. As with COMPASS and ACCUPLACER, research also is needed on whether the test provides any additional prediction information over already-mandated tests or high school transcripts. Until that time, there is insufficient scientific evidence justifying the use of WorkKeys.

Public education students already are seriously over-tested and money is very limited. Spending many millions of dollars and hours of student and staff time on yet more tests cannot be justified without first making sure that the tests really provide additional useful information.

Students and schools will continue to have to pay attention to such tests as the ACT reading and science, the Compass, the ACCUPLACER, and the WorkKeys when high stakes decisions are being made upon them. But given the availability of good tests like the ACT English and math, as well as the as-yet-not-refuted SAT tests, we have much better tests available to answer those high-stakes questions.

This report is not meant to be an exercise in psychometric nit-picking. The high stakes tests discussed in this report provide measurements that can be seriously life-altering for the students who take them. When there is compelling recent research from nationally established institutions that challenge the validity of such tests, and when the findings from those studies validate each other (in the case of the Columbia studies) or are validated by the vendor themselves (the NBER study), we need to step back and reconsider the use of such tests. Similarly, we must reconsider the use of measures like WorkKeys that have not been independently validated. The educational community needs to connect placement policy with contemporary research.