So you think you can dance?
Some concrete suggestions and cautions in evaluating/validating claims of college readiness

Thanos Patelis
Center for Assessment

Presentation at the National Conference on Student Assessment
San Diego, CA

June 23, 2015
Overview

• Claims of college readiness
• Framework
• Some methodologies
• Examples of results
• Cautions
• Takeaways
Claims of College Readiness (Samples)

- Achieve - Published English and mathematics benchmarks in 2004 (Achieve, 2004). The definitions of college and career readiness were in the form of the content that students should be able to do to be fully prepared to succeed in credit-bearing college courses or in certain types of occupations.

- ACT - Using final course grade data from a large sample of colleges, ACT (Allen & Sconing, 2005) modeled the probability of success in typical first-year courses as a function of ACT test scores. Success was defined as a course grade of B or higher, and for each college the ACT test score that yield a .50 probability of success was identified. The median of these scores represents the college readiness benchmark. Benchmarks were obtained for four common first-year courses: English Composition, using the ACT English score as the predictor; College Algebra, using the ACT Mathematics score as the predictor; Social Science, using the ACT Reading score as the predictor; and Biology, using the ACT Science score as the predictor.

- David Conley and EPIC - The work that stimulated the conversation around college readiness was a study to discover what must students know and be able to do in order to succeed in entry-level university courses (Conley, 2003). In a two-year study involving more than 400 faculty and staff members from twenty research universities, all members of the Association of American Universities (AAU), in extensive meetings and reviews statements representing college success were developed.

- NCEE - The National Center on Education and the Economy (NCEE) presented results of research on the requirements of community colleges. The rationale for focusing on community colleges was that they offer a gateway to four-year colleges and the workforce. Community college staff members were asked what kind and level of literacy in mathematics and English is required if that student is going to have a good chance of succeeding in the first year of a typical community college program. (NCEE, 2013a; 2013b)

- Ohio Board of Regents - In 2007, Ohio enacted a requirement as follows: “Not later than December 31, 2012, the presidents, or equivalent position, of all state institutions of higher education, or their designees, jointly shall establish uniform statewide standards in mathematics, science, reading, and writing each student enrolled in a state institution of higher education must meet to be considered in remediation-free status” (See 3345.061 (F) of the Ohio Revised Code: http://codes.ohio.gov/orc/3345.061). In December 2012, the presidents of Ohio’s public colleges and universities established a set of standards and expectation in English that included reading, writing, speaking, viewing and listening, mathematics that included mathematical processes, number and operations, algebra, geometry, probability and statistics, science in the disciplines of biology, chemistry, computer science, engineering, geology, and physics. Additionally, they established a set of college readiness indicators that they purport are supposed to guarantee remediation free status at any public post-secondary institution in Ohio. These cutoffs were set in the areas of English, reading and mathematics for the ACT, SAT, ACCUPLACER and COMPASS (Ohio Higher Ed, 2012).

- PARCC - Level 4 representing (a) in ELA/literacy Students performing at this level demonstrate a strong command of the knowledge, skills, and practices embodied by the Common Core State Standards for English language arts/literacy assessed at grade 11. They are academically prepared to engage successfully in entry-level, credit-bearing courses in College English Composition, Literature, and technical courses requiring college-level reading and writing. Students performing at this level are exempt from having to take and pass college placement tests in two- and four-year public institutions of higher education designed to determine whether they are academically prepared for such courses without need for remediation; (b) in math Students performing at this level demonstrate a strong command of the knowledge, skills, and practices embodied by the Common Core State Standards for Mathematics assessed at Algebra II or Mathematics III They are academically prepared to engage successfully in entry-level, credit-bearing courses in College Algebra, Introductory College Statistics, and technical courses requiring an equivalent level of mathematics. Students performing at this level are exempt from having to take and pass placement tests in two- and four-year public institutions of higher education designed to determine whether they are academically prepared for such courses without need for remediation.

- Smarter Balanced - Level 3 representing (a) in ELA/literacy students who perform at the College Content-Ready level in English language arts/literacy demonstrate reading, writing, listening, and research skills necessary for introductory courses in a variety of disciplines. They also demonstrate subject-area knowledge and skills associated with readiness for entry-level, transferable, credit-bearing English and composition courses; and (b) Students who perform at the College Content-Ready level in mathematics demonstrate foundational mathematical knowledge and quantitative reasoning skills necessary for introductory courses in a variety of disciplines. They also demonstrate subject-area knowledge and skills associated with readiness for entry-level, transferable, credit-bearing mathematics and statistics courses.

- Texas - In 2009, Texas with assistance from EPIC developed the Texas College and Career Readiness Standards (THECB & TEA, 2009). “The 79th Texas Legislature, Third Called Special Session, passed House Bill 1, the “Advancement of College Readiness in Curriculum.” Section 28.008 of the Texas Education Code, seeks to increase the number of students who are college and career ready when they graduate high school. The legislation required the Texas Education Agency (TEA) and the Texas Higher Education Coordinating Board (THECB) to establish Vertical Teams (VTs) to develop college and career readiness standards in the areas of English/language arts, mathematics, science, and social studies. “These standards specify what students must know and be able to do to succeed in entry-level courses at postsecondary institutions in Texas” (p. ii). As a result of this, the standards were produced and organized around four levels of specificity of (a) the key content, (b) organizing components, (c) performance expectations, and (d) examples of performance indicators. A resource was provided to explain and help schools understand the specifics of this (see www.texcrs.org).”

- Virginia - In January 2007, the Board of Education authorized VDOE to conduct studies to determine factors contributing to success in postsecondary education. In 2009 and 2010, respectively, the Virginia Board of Education approved revised Standards of Learning in mathematics and English. In 2011, Virginia developed college and career ready mathematics and English performance expectations involving Virginia’s community colleges and four-year institutions. These performance expectations define the level in entry-level, credit-bearing, college courses in mathematics and English or further career and technical training after high school (VDOE, 2011). In English, the content are in the areas of reading with vocabulary, nonfiction reading, literary reading, reading analysis and critical reading as components; writing comprised of the areas of composing, revision and editing, and documentation and ethics; and communicating comprised of speaking, listening and collaborating. In mathematics, the expectations are organized as (a) problem solving, decision making and integration, (b) understanding and applying functions, (c) procedure and calculation, and (d) verification and proof. (See: http://www.doe.virginia.gov/instruction/college_career_readiness/).
Validation Framework

*Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests.*

-- AERA et al., 2014, p. 11

### Explicit Statement of Proposed Interpretation of Test Scores

<table>
<thead>
<tr>
<th>Types of Evidence:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test Content</td>
</tr>
<tr>
<td>Response Process</td>
</tr>
<tr>
<td>Internal Structure</td>
</tr>
<tr>
<td>Relations to Other Variables</td>
</tr>
</tbody>
</table>

Consequences of Testing

The focus here is one type of validation evidence.
Relations to Other Variables

• Historically, most popular validation evidence (Jonson & Plake, 1998; Cizek, Rosenberg, & Koons, 2008)

• Offers an empirical approach using external data sources to make comparisons about the information provided.

• Takes advantage of data that exist concurrently with the assessment data in question, as well as what will exist in the near future.

• Does not imply the same results – the actual methodology matters.

• Along with content-related evidence, the evidence produced is compelling.

• Context matters – While national studies have been used to develop benchmarks and provide validation evidence, what college success means, what student experience and what performance represents is context specific.
Components involved in Relations to Other Variables

From Willingham, Lewis, Morgan, & Ramist, 1990; Patelis, 2001; 2002; Patelis, Camara, & Wiley, 2009
Components

The Predictors

- Characteristics of the test and/or grades
- Meaning of the scores
- Score variations
- Educational practices

The Criterion

- There are factors that affect the comparability of scores from tests that include purpose/use, construct/content, administration conditions, scoring, population of students tests, and opportunity to learn. (Gong and DePascale, 2013, p. 8).
  - One of these factors includes the degree of experiential specificity of the predictors used. They can range from specific representations (course grades or end-of-course test scores) to assessments representing more generalized skills (e.g., ACT and SAT) (Anastasi & Urbina, 1997; Geisinger, 2002).
Components (cont’d)

The Sample

- Institutional variation
- Types of institutions
- Students represented
- Nature of the data

- Student selection processes (i.e., admissions, course placement, opportunities to learn, etc.) can drastically impact the institutional variation, the type of institutions, and the students represented.

- If college outcomes are used as a criterion, the selection process represented by the admissions process will affect the sample and validity coefficients (Young, 2001; Ling, Patelis, & Lewis, 2004; Shen, Sackett, Kuncel, et al., 2012).

- A taxonomy of nine admissions models for post-secondary institutions can assist in understanding the institutional variation, the types of institutions, and students represented (Perfetto et al., 1999).
Relations to Other Variables

Elementary | Secondary | Post-Secondary

1. Claims of college readiness

2. College outcomes
   - Enrollment
   - Course Grades
   - Discipline Grade Point Average
   - Cumulative Grade Point Average
   - Retention/Persistence
   - Graduation
   - Faculty Ratings
   - Course Placement

3. “Link”

4. “Link”

5. Other Criteria
   - Course Grades
   - Cumulative Grade Point Average
   - Retention
   - Graduation
   - Teacher Ratings
   - Course Placements

6. Other Tests
   - 10th/11th Grades:
     - AP
     - PLAN
     - PSAT/NMSQT
   - 11th/12th Grades:
     - ACT
     - AP
     - SAT

The framework by Holland (2007) offers the basis for linking and evaluating the benchmarks with other variables.
Notes on Data

• Data can be obtained as students move forward and test scores and other outcomes are collected.
  – For tests administered this academic year, use test scores and other data from this academic year.
    • Convergent and concurrent validation
  – Next year, additional test scores and other data can be gathered and used.
    • Convergent, concurrent, and predictive validation
  – For tests administered to 10th and 11th grades, may need to wait 3-4 years before students are in a post-secondary institution.
    • Work with college/university to get data
    • Use data from the National Student Clearinghouse to get college enrollment and persistence data
  – Once students move through elementary, middle and into high school, use historical student data.
    • Back map

• Data can be obtained by administering high school tests to college students.
  – Effort involves recruiting post-secondary institutions and students, ensuring student motivation, and sample representativeness

• Use archival data (weaker argument)
  – If last year’s scores have college readiness benchmarks (e.g., PLAN, PSAT/NMSQT, AP, others), examine relationship between this year’s test scores and last years performance.
Linking Methodologies

Descriptive Information:
• Correlation coefficients
• Box plots of test scores by performance levels of college readiness
• Cross tabulations of performance levels and college readiness
• Scatter plots

Projections:
• Logistic regression
• Linear regression
• Borderline groups approach
• Equipercentile
• Decision theory
• Back mapping
In addition to quantitative representations of the relationship between two test scores or other variables (i.e., correlation coefficients) or cross tabulations comparing performance levels and being college ready or not, a visual display can be very informative.

- Using box plots showing distribution of scores at each performance level on a high school state test.
- This is a nice visual representation of the performance on tests with college readiness benchmarks and the state test.
- This represents 42% and 45% of the 10th grade students in the state taking the PSAT/NMSQT and SAT, respectively (Patelis, Behuniak, & Tucker, 2001).
Linking Methodology - Logistic Regression

• This is a very popular approach used by ACT and College Board in establishing college readiness benchmarks (as well as in providing evidence to support these benchmarks) for their tests (see for example, Allen & Sconing, 2005; Wiley, Wyatt & Camara, 2010).

• Data representing predictors (the scores associated with the test) and criteria (other tests scores or outcomes) for the same students are used.
  – While the inference will differ, this methodology can be used with concurrent data or data representing predictive type of relationship or data looking backward.

• Use generally a cut-off on the criterion that represents a dichotomy of either college readiness or college success.
  – This represents one advantage of the logistic regression in that the modeling of the relationship is focused around what is considered as college readiness or college success.

• Decide on a likelihood value(s) for estimating projected score.
  – This can be done using a policy capturing approach (e.g., Kobrin, Patterson, Wiley, & Mattern, 2012; Camara, 2015; Thacker, 2015) or logical reasoning or arbitrary values.
  – This represents another advantage of the logistic regression in that you can explicitly represent the tolerance for uncertainty.

• Calculate the regression coefficient associated with specified likelihood values between the predictors and criteria.
Logistic Regression – Example 1

• Task: Provide evidence to inform the proficiency cut-score on an 11th grade state test score.

• Used PSAT/NMSQT test scores with “steps to college readiness benchmark” for students who also took the state test. These particular benchmarks represent the scores at which a student will have a 65 percent likelihood of attaining the SAT college readiness benchmark.
  – Sample was 7,495 11th grade students in this state
  – Represented 36% of the 11th grade state test takers
  – Correlation coefficients of performance between sections on each test were substantial.
    • ELA/Critical Reading = 0.72
    • Mathematics = 0.82
    • Writing = 0.72

• Issue: The sample was not representative of the 11th graders in this state.
  – Performance on the state test was higher for the sample who took both tests.

• Cognizant of this limitation, the analyses were done to explore the results and gleam some qualified information.
Logistic Regression – Example 1 (cont’d)

- The college readiness benchmark on the PSAT/NMSQT for 11th graders and the grade 11 state test proficiency cut-score are shown.

- As can be seen, the grade 11 proficiency cut-score is well below the college readiness benchmark – even based on this relatively higher performing sample of 11th graders.

<table>
<thead>
<tr>
<th></th>
<th>Reading</th>
<th>Mathematics</th>
<th>Writing</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSAT/NMSQT College-readiness benchmark for juniors (20 – 80)</td>
<td>50</td>
<td>50</td>
<td>49</td>
</tr>
<tr>
<td>Grade 11 State Proficiency Cut-Score</td>
<td>40</td>
<td>40</td>
<td>40</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Projected PSAT/NMSQT Scores for the Grade 11 Proficiency Cut-Score</th>
<th>Reading</th>
<th>Mathematics</th>
<th>Writing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic regression (0.50)</td>
<td>40</td>
<td>43</td>
<td>31</td>
</tr>
<tr>
<td>Logistic regression (0.65)</td>
<td>42</td>
<td>45</td>
<td>35</td>
</tr>
</tbody>
</table>
• Data representing predictors (the test scores associated with the test) and criteria (other tests scores or outcomes) for the same students are used.

• Using the cut-score on the predictor add/subtract one standard error of measurement of the predictor to/from the cut-score.

• For those students who score within one standard error of measurement around the mean on the predictor, calculate the mean on the criterion.

• Can also make the calculations the other direction using the benchmark of the criterion and standard error of measurement of the criterion to select the students and calculate the mean on the predictor.
• The projected PSAT/NMSQT scores for the grade 11 proficiency cut-score are below the college readiness benchmark.

• The borderline group method results are very similar to the logistic regression method.

<table>
<thead>
<tr>
<th>Projected PSAT/NMSQT Scores for the Grade 11 Proficiency Cut-Score</th>
<th>Reading</th>
<th>Mathematics</th>
<th>Writing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Borderline group</td>
<td>41</td>
<td>43</td>
<td>36</td>
</tr>
<tr>
<td>Logistic regression (0.50)</td>
<td>40</td>
<td>43</td>
<td>31</td>
</tr>
<tr>
<td>Logistic regression (0.65)</td>
<td>42</td>
<td>45</td>
<td>35</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PSAT/NMSQT College-readiness benchmark for juniors (20 – 80)</th>
<th>Reading</th>
<th>Mathematics</th>
<th>Writing</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSAT/NMSQT College-readiness benchmark for juniors (20 – 80)</td>
<td>50</td>
<td>50</td>
<td>49</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Grade 11 State Proficiency Cut-Score</th>
<th>Reading</th>
<th>Mathematics</th>
<th>Writing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade 11 State Proficiency Cut-Score</td>
<td>40</td>
<td>40</td>
<td>40</td>
</tr>
</tbody>
</table>
Task: How does the cut-score on the 10\textsuperscript{th} grade state test compare to another test’s college readiness benchmark?

Used three years of PSAT/NMSQT test scores with college readiness benchmarks for students who also took the state test. These benchmarks represent the minimum scores at which students have a 65\% chance of obtaining a 2.67 freshman year grade point average (FYGPA).

- Samples ranged from 8,794 to 15,307 who took both tests
- Represented approximately 11-20\% of the 10\textsuperscript{th} grade state test takers
- Correlation coefficients of performance between sections on the test were substantial.
  - Reading = 0.74-0.79
  - Mathematics = 0.79-0.85

Issue: The sample was not representative of the 10\textsuperscript{th} graders in this state.

- Performance on the state test was higher for the sample who took both tests.

Cognizant of this limitation, the analyses were done to explore the results and gleam some qualified information.
Logistic Regression & Borderline Group – Example 2

- The projected 10th grade state test scores were above the “meeting the standard” cut-score on the state test, but within the range associated with the performance level. The projections were actually near the center of the range of the performance level.
  - This was expected based on the higher performance of the sample taking both the state test and the PSAT/NMSQT in 10th grade compared to all students taking the state test in 10th grade.
  - The findings, considering the sampling bias, supported the rigor of the performance level associated with “meeting the standard”.
- The borderline and logistic regression provided similar results.

<table>
<thead>
<tr>
<th></th>
<th>Reading</th>
<th>Mathematics</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSAT/NMSQT College Readiness Benchmarks for 10th graders</td>
<td>42</td>
<td>44</td>
</tr>
<tr>
<td>Cut-Off for Meeting the Standard on 10th Grade High School Test</td>
<td>674</td>
<td>487</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Reading</th>
<th>Mathematics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Borderline</td>
<td>Logistic Regression</td>
<td>0.50</td>
</tr>
<tr>
<td>Yr 1</td>
<td>738</td>
<td>733</td>
</tr>
<tr>
<td>Yr 2</td>
<td>736</td>
<td>727</td>
</tr>
<tr>
<td>Yr 3</td>
<td>732</td>
<td>728</td>
</tr>
</tbody>
</table>
Cautions

• With a number of new tests being introduced making claims of college readiness, gathering validation evidence to support these claims is crucial.

• While there are a number of types of evidence and the assessment program’s goals and theory of action should drive the type of evidence needed, gathering empirical evidence based on relations to other variables is a popular, effective approach that, if done well, can offer a compelling argument to support the claims being made.

• The components involving the predictors, the criterion, and the sample being used must be clearly specified and their impact on the results noted.
  – A perfect study does not exist.
  – The limitations represented by the decisions to use certain predictors, criteria, and samples must be noted.
  – But, some evidence, properly documented and qualified, is better than no evidence.
  – Better yet, a body of evidence, properly documented and qualified, is better still.

• While there are limitations imposed by decisions on which predictors and criteria to use (assuming they’re reliable), in studies examining the relation to other variables, there is nothing more impactful than the sample used.
• **Timing of data**
  
  – If you use the ACT, SAT, or data with multiple time points, it is crucial to keep track of the timing.
  
  – For example, students can take the SAT seven times a year starting in 9th grade.
  
  – If you were to examine the relationship of the 10th grade state test to the SAT college readiness benchmark, it is important to select an administration or set of administrations and the number of times the SAT was taken.
    
    • Score variations may arise depending on the time interval associated with which administrations selected and repeaters.

![](diagram.jpg)

• **ACT and SAT data generally represent the latest test scores for a cohort of students graduating high school on a certain year.**
  
  – Need to get date when tests taken or data for specific administrations.
  
  – Note that there are score variations by administration. So, if possible, replicate analyses by time intervals.
Cautions (cont’d)

• Sample
  – Missing data based on sampling can affect the magnitude of the validity coefficients.
    • This can happen when the criterion is another test where not all students participate (e.g., ACT or SAT).
  – Selection systems or attrition can affect the score variability and influence the results.
  – Qualify interpretation, if the participate rates are different between predictors and criteria.

• Educational Practices
  – When examining the relationship of state test scores to other tests, school selection or course assignment practices can produce differential experiences that could create differences in the regression coefficients between the predictor and criterion by school.
  – Examine (if at all possible) differences in the relationship between predictors and criteria by school.
    • Could utilize intraclass correlation coefficient – similar to what one would do when using a multi-level or hierarchical linear model analysis

• Institutional Variation
  – Since post-secondary institutions will have different admissions models, the relationship in state test scores to college outcomes will be influenced by the impact of these institutional practices.
  – Replicate analyses across types of institutions or utilize some kind of multi-level or hierarchical linear modelling.
  – Example: Validation study examining the relationship of the SAT on college freshman GPA found different raw and adjusted correlation coefficients for all institutions (0.35 raw and 0.53 adj.) vs. those that were more selective institutions accepting less than 50% of the applicants (0.39 raw and 0.58 adj). Kobrin, Patterson, Shaw, Mattern, & Barbuti, 2008.
Cautions (cont’d)

• Methods

  – In using the logistic regression methods, limitations can arise when the proportions of the dichotomy in the criterion (e.g., percent at/above the benchmark is 90% vs. below is 10%) are substantially different.
    • The accuracy of the regression coefficients will be inflated.
    • Selecting a higher likelihood in calculating the projected score may help.
    • At least, examine it and qualify the findings.

  – Methods relying on observed scores without modeling the relationship between the predictor and criterion (e.g., borderline group or equipercyntile) may be affected by drastic differences in the distributions of scores with respect to differences in the degree of non-normality between the distributions.
    • It is a good idea to examine the distributions.
    • Sampling bias can create the differential distributions.
    • Even if you use multiple methods to examine relationship or make projections, lean toward selecting the results that are produced from methods more robust in the face of issues with the distributions (e.g., logistic regression).
Takeaways

• For assessments programs making claims of college readiness, should implement validation studies involving examining the relations to other variables.
  – Part of body of evidence that contribute or attend to theory of action of the assessment program.
• There is a wide array of other variables that can be selected.
  – Utilize third-party sources (faculty ratings, colleges/universities, National Student Clearinghouse)
• Care must be exercised in selecting the variables.
  – Care and time needed in selecting variables and the time intervals associated between predictors and criteria.
• Sampling issues offer the most potential for impacting the results.
• Educational practices and institutional variation necessitate examination of results by school and institution.
  – Replication or some way of accounting for school effects should be exercised.
• Utilize multi-method approach to help in converging or triangulating on the results.
• Some evidence, even if qualified, is better than no evidence.
  – Use what you have; purposefully sample within a state to maximize representation; examine (or at least weight) by school/institution.
• Build analyses and studies into ongoing operations of the assessment program.
• Testing vendor can do this, but be an informed consumer.
References

Camara, W. J. (2015). Employing empirical data in judgmental processes. Presentation at the National Assessment on Student Assessment, San Diego, CA.
http://www.ncee.org/college-and-work-ready/
Validation Methods 25
For more information:

Center for Assessment
www.nciea.org

Thanos Patelis
tpatelis@nciea.org