the RP group

Research • Planning • Professional Development
for California Community Colleges

Validating Placement Systems Comprising Test and Multiple Measure Information

Craig Hayward

May 2017

www.rpgroup.org

Executive Summary

Placing students into math, English, reading, and ESL sequences has a tremendous impact on students' trajectories and their ultimate probability of success. The purpose of this research brief is to suggest effective ways of validating the placement decisions made by community colleges based on information generated by assessment and placement systems which include multiple measures of student capacity. The information is generally applicable to a variety of placement systems; however, the particular focus is on scenarios in which students take an assessment test and also provide high school performance information, such as GPA, as a multiple measure.

This paper argues that there are at least two distinct aspects or phases of validation that should be conducted: (1) psychometric properties of the test instrument and of the predictive validity of non-test measures of student capacity and (2) validating the decisions made based on the output of the placement system. The second aspect of validation is the focus of this brief. Test validation includes a number of important and well-known metrics such as reliability, content validity, construct validity, and bias assessment. The validation of the actual placement decisions made on the basis of tests and other information is less well-developed, however. The current California Community Colleges Chancellor's Office standards call for validating placement decisions by asking students (and faculty) whether they felt they were properly placed. Such a form of validation is far from definitive as students are not experts in assessment and placement and, in such a situation, are subject to cognitive biases such as system justification and confirmation bias. This research brief describes how to validate placement decisions using several objective key metrics including throughput, throughput rate, predictive validity, and disproportionate impact.

Introduction

The placement of incoming college students into an initial English or math course (developmental vs. college level) has important implications for students' likelihood of enrollment, persistence, and completion (Bailey, Jeong, and Cho, 2010; Fong, 2016; Fong and Melguizo, 2016; Hayward and Willett, 2014; Melguizo, Kosiewicz, Prather, and Bos, 2014). There is a growing consensus that including additional sources of information beyond placement test scores reduces error in placement decisions. For example, accuracy of placement can be improved by incorporating high school performance information, such as GPA and course grades earned in high school (Belfield and Crosta, 2012; Geiser and Santlices, 2007; Fuenmayor, Hetts, and Rothstein, 2012; Ngo and Kwon, 2015; Scott-Clayton, 2012; Scott-Clayton, Crosta, and Belfield, 2014; Willett, Hayward, and Dahlstrom, 2008; Willett, 2013). The importance of including additional sources of information has been codified in California's legislation, policies, and standards. Specifically, California Community Colleges are mandated to use at least one additional independent source of information when using a test to place students into English, math, reading, and ESL coursework (California Community College Assessment Association, 2001).

There are two distinct aspects of validating placement systems: (1) validating the test or measure itself and (2) validating the decisions being made as a result of the output of the placement system. Validating a placement test itself requires that the test evince good psychometric properties such as reliability, content validity, lack of bias, and construct validity. The desirable psychometric features of tests are addressed in the current standards and in numerous other publications (e.g., Carlson, Geisinger, and Jonson, 2014). While tests have an established process of psychometric validation in statewide standards, multiple measures do not. Indeed, the Multiple Measures FAQ explicitly indicates that multiple measures do not require validation (CCCAA, 2005). Multiple measures,[i] such as high school performance information, are typically not tests and therefore do not have the same psychometric properties as placement tests (it is possible to use a second, uncorrelated test as a multiple measure, but this rarely happens in practice). One way for multiple measures to demonstrate their validity is to show that they indeed enhance the predictive validity of a placement system above and beyond the performance of the test alone. A number of research projects, including the Multiple Measures Assessment Project (MMAP), have demonstrated that the correlation between high school GPA and success in initial community college math and English classes can meaningfully enhance the predictive validity of placements systems that are based on tests alone. These findings establish in a general way the predictive validity and usefulness of high school performance information (Willett et al, 2015; Scott-Clayton, 2012; Westrick and Allen, 2014).

Despite growing consensus on the predictive validity and importance of multiple measures of student capacity in placement decisions, there is relatively little explicit guidance on how to validate the decisions made by placement systems that utilize multiple sources of information. The *Standards Policies and Procedures for the Evaluation of Assessment Instruments Used in the California Community Colleges* mentions multiple measures only three times - and each mention is only in passing (CCCAA, 2001). No detailed information on validation of multiple measures per se is provided.[ii] Yet it is essential to realize that validation of placement systems is not conducted on the parts of the placement system in isolation. Rather, one validates the

decision made as a result of the totality of the information analyzed to arrive at this decision. Essentially, we are asking, "Is this the best possible placement decision for this student given the available information?" (Kane, 2006). A placement decision based on information from a test along with multiple measures must be validated as a single decision, even though information from two sources (e.g., common assessment and MMAP decision rules) is considered; the test and the multiple measures are both part of the overall placement system.

An example may help illustrate why we cannot validate the individual parts of a complex placement system. Imagine there are two students; each answers the same questions on an objective English placement test correctly. The test results suggest that the students should be placed in developmental education, one level below the transferable gateway English class. One student, however, has a 3.3 high school GPA; the other student has a 2.3 high school GPA. Since these students are being assessed at a college that uses the statewide MMAP decision rules, one student will be cleared for transfer-level English because his or her high school GPA is higher than 2.6. The other student's placement will be confirmed at one-level below transfer. Placing students with the same test results in two different levels will confound a validation process that only considers test scores. It is necessary to factor in the effect of the multiple measures in order to understand how the placement system is actually working and why students with the same test results are being placed differently.

The current Californai Community Colleges Chancellor's Office (CCCCO) standards call for validating placement decisions by asking students (and faculty) whether they felt they were properly placed. This form of validation, described as consequential validity, is far from definitive; students are not experts in assessment and placement and their judgment in such a situation is subject to cognitive biases such as system justification and confirmation bias (Jost, Banaji and Nosek, 2004; Nickerson, 1998). Nonetheless, it is the de facto standard for validating placement decisions across the California Community College system. In the scenario above, assuming both students are appropriately placed, a consequential validation would likely show that students and teachers agree that the students were placed appropriately. However, it is problematic that identical information from the test was used to place students into different levels. Without explicitly incorporating multiple measures into the decision validation process (i.e., cut-score validation), the validation process is necessarily incomplete. In a test-centric validation process, students with the same test results should not be placed into different levels. Because the placement decision is influenced by the multiple measure as well as by the test score, it is necessary to validate the overall decision. Since students' *actual* placement is based on multiple pieces of information about student capacity, validating only the test component of the placement system will yield incongruous results.

The validation that we are discussing in this brief is similar to the cut-score validation process described in the CCCCO standards. However, here we outline a process for validating the placement decision that is based on both test and multiple measures information. In other words, we are outlining a process for validating the use of *all* placement information. In this brief, we address the use of a test instrument (one that is already psychometrically validated), along with other information, known as multiple measures, to place students in math, English, ESL, and reading sequences. Colleges can use this document to validate the decisions that are being made (or are potentially being made) as the result of the combination of placement test and multiple measures information.

As California community colleges are in the process of re-examining their assessment and placement designs and practices, it is appropriate at this time to suggest that multiple measures should have some criteria for validation. Since multiple measures are potentially widely variable given the local control of colleges to design and implement their placement systems, what metrics can be used for their validation? Two available metrics that can evaluate a placement system that incorporates multiple measures are: (1) predictive validity and (2) throughput optimization.

Multiple Measures Evaluation Metric 1: Predictive Validity

Predictive validity is the extent to which a score on a measure predicts scores on a certain criterion, such as success in a key course. In order to demonstrate predictive validity accurately, a multiple measure must be reliably associated with an outcome of interest, such as success rates in a course or throughput rates for a sequence (i.e., the rate of students completing the transfer level course in the sequence regardless of where they begin in the sequence). The association should be statistically significant, and it should be strong enough to improve the accuracy of placement decisions.

In a typical case, multiple measures information will be combined with results from a psychometrically validated placement test. To analyze the predictive validity of a multiple measure, college staff should first look at the association of the multiple measure on its own with a relevant criterion such as success in a target course. The particular statistical procedure will depend on the measurement type of the multiple measure (e.g., categorical, ordinal, ratio).[iii] The association between the multiple measure and the criterion should be tested at each level of the sequence, beginning with the most advanced course. Transfer level placement is the most critical placement for students, as it is this level of placement that must be assessed for disproportionate impact (cf. CCCAA, 2001).

If multiple measures are being used disjunctively[iv] with the test, the demonstration of a statistically significant association is evidence that the multiple measure may be used to improve accuracy of placement. While there are no absolutes in what constitutes a "good enough" correlation to be useful, the current standards call for a correlation between predictor and outcome of at least 0.35 (CCCCA, 2001). However, if the goal is to improve upon what has come before, an uncorrected statistical association equivalent to at least a correlation coefficient of 0.15 (i.e., $r \geq 0.15$) represents a reasonable guideline for a minimum practical level of predictive validity.[v] It may also be the case that combining several distinct measures reveals unique contributions from each measure, increasing the overall multiple correlation. In such cases, it may be beneficial to use all of the available metrics and information.

If multiple measures are being combined with the test information to produce a single, blended placement (i.e., a compensatory model), the next step in determining predictive validity involves assessing whether the multiple measure is able to increase the predictive validity of the test on its own. This step is accomplished by conducting a multivariate analysis, such as logistic regression or decision tree analysis. If using a logistic regression, the multiple measure should have a significant Beta coefficient, even with the test coefficient in the model. The overall proportion of variance explained (as indicated by the Cox and Snell or Nagelkerke pseudo R^2) should be

significantly higher for the model that includes the multiple measure and the test score than it is for the model that predicts the criterion exclusively from a test score.[vi]

Multiple Measures Evaluation Metric 2: Optimizing Throughput

Throughput describes the number or proportion of a cohort of students who complete a gateway (or "gatekeeper") math or English course (i.e. the transfer-level course at the end of the sequence).[vii] Validating placement systems based on throughput rates has certain advantages over the criterion validity approach. Once a measure is actually being used to place students, it becomes harder to show predictive validity because the range of scores at any given course level becomes restricted. For instance, students who are placed into transfer-level coursework will tend to have higher scores on the multiple measure because students with lower scores will predominantly be placed into lower-level courses. This restriction of variance in scores will tend to artificially lower correlation estimates. Throughput rates, on the other hand, examine sequence completion as the ultimate aim of developmental coursework and can readily be used to evaluate placement systems that are already in effect.

If changes are made to the parameters of the placement system (e.g., a different high school GPA is used), they should result in increased throughput rates. Changes that reduce throughput rates would generally be considered unacceptable without justification. For example, if a college implements a new placement process and begins placing more students into developmental education classes, it would be required to show that, as a result of this change, a higher proportion of students were able to successfully complete the transfer-level course than before the change. If throughput rates showed a decline, then the new placement system would be considered to be less valid than the prior system. The college would then need to revert to its prior system, or try a new approach that promised to improve throughput rates.

The predicted throughput rate (i.e., the eventual success rate in the college level course) of a cohort of starting students, given their placement, becomes a key piece of information to consider when evaluating whether to place a student into developmental or into college level work.[viii] If a student's expected throughput rate is lower when s/he is placed into remediation than when placed directly into transfer, this would indicate that, in terms of boosting the probability of success in the sequence as a whole, there is no benefit to this student in being placed into remediation.[ix]

Disproportionate Impact

It is important to conduct disproportionate impact (DI) analyses for both transfer level course success rates and for throughput rates. Per the current standards, disproportionate impact needs to be examined for various student subgroups including by ethnicity, gender, age, and disabled students programs and services (DSPS) status. Predictive validity and throughput rate should be disaggregated by these groups to determine if any groups are being adversely impacted by the placement system. If predictive validity models do not work as well for some subpopulations or if some subpopulations are disproportionately less likely to complete transfer-level courses, additional research will be required to examine why this pattern exists and what can be done to

correct it (e.g., using different or additional assessment measures). Use a standard measure of DI, such as the proportionality index, the percentage gap method, or the 80% rule to ascertain whether DI exists (Institutional Effectiveness Partnership Initiative (IEPI), 2016). If DI is found, the college must develop a plan to address the disproportionate impact or else explain why it is in some way unavoidable.

Summary of Key Metrics Used to Validate Placement Decisions

- Throughput: The total number of students who successfully complete or are projected to successfully complete the transfer-level gateway or gatekeeper course.
- Throughput rate: The number of students who successfully complete the transfer-level gateway or gatekeeper course at the end of a course sequence divided by the number of students in the initial cohort.
- Predictive validity: A demonstration that a placement test or multiple measure is associated with outcomes of interest such as course success rates. Assessed by a measure of association such as a correlation coefficient.
- Disproportionate impact: A condition where some students' access to college-level coursework and ultimately their academic success is hampered by inequitable practices, policies, and approaches to student placement.

Research and Planning Group for California Community Colleges

The Research and Planning Group for California Community Colleges (RP Group) strengthens the ability of California community colleges to discover and undertake high-quality research, planning, and assessments that improve evidence-based decision-making, institutional effectiveness, and success for *all* students.

Project Team

This report was prepared by the MMAP Research Team. The primary contact is Craig Hayward, who can be reached at chayward@rpgroup.org.

www.rpgroup.org

References

Bailey, T., Jeong, D., & Cho, S. (2010). *Student progression through developmental sequences in community colleges* (CCRC Brief). New York: Columbia University, Teachers College, Community College Research Center.

Behizadeh, N. & Englahard, G. (2011). Historical view of the influences of measurement and writing theories on the practice of writing assessment in the United States. *Assessing Writing*, 16, 189-211.

Belfield, C. & Crosta, P. M. (2012). *Predicting success in college: The importance of placement tests and high school transcripts* (CCRC Working Paper No. 42). New York: Columbia University, Teachers College, Community College Research Center.

Burton, N.W. & Ramist, L. (2001). *Predicting success in college: SAT® studies of classes graduating since 1980* (College Board Research Report No. 2001–002.) New York: College Board.

California Community College Assessment Association (2005). *Assessment Q & A*. Retrieved from CCCCO Matriculation Handbook on 4/23/17: http://extranet.cccco.edu/Portals/1/SSSP/Matriculation/MatriculationHandbookRevSeptember2011.pdf

California Community College Assessment Association (2001). *Standards Policies and Procedures for the Evaluation of Assessment Instruments Used in the California Community Colleges*. Retrieved from CCCCO Matriculation Handbook on 4/23/17: http://extranet.cccco.edu/Portals/1/SSSP/Matriculation/MatriculationHandbookRevSeptember2011.pdf

Carlson, J. F., Geisinger, K. F., and Jonson, J. L. (Eds.). (2014). *The Nineteenth Mental Measurements Yearbook*. Lincoln, NE: Buros Center for Testing.

Fong, K. (2016). *A Look at Developmental Math Education through Student Behavior: Students' Progression through their Sequence and Students' Choices through their Assessment and Placement Process*. Paper presented at the annual conference of the Research and Planning Group, San Diego, California. Retrieved August 13, 2016 from http://rpgroup.org/system/files/A%20Look%20at%20Developmental%20Math%20Education%20through%20Student%20Behavior.pdf

Fong, K. & Melguizo, T. (2016). Utilizing Additional Measures of High School Academic Preparation to Support Students in their Math Self-Assessment. *Community College Journal of Research and Practice*, June 2016. DOI: 10.1080/10668926.2016.1179604

Geiser, S. & Santelices, M.V. (2007). *Validity of high-school grades in predicting student success beyond the freshman year: High school record vs. standardized tests as indicators of four-year college outcomes*. Research and Occasional Paper Series (CSHE.6.07). Berkeley, CA: University of California, Center for Studies in Higher Education.

Hayward, C. & Willett, T. (2014). *Curricular Redesign and Gatekeeper Completion: A Multi-College Evaluation of the California Acceleration Project*. Berkeley, CA: The Research and Planning Group for California Community Colleges.

Institutional Effectiveness Partnership Initiative (2016). *Using Disproportionate Impact Methods to Identify Equity Gaps*. Retrieved from https://s3-us-west-1.amazonaws.com/pdwareh/wp-content/uploads/2017/01/Disproportionate_Impact_Equity_and_Placement-201701051.pdf Accessed on 5/7/17.

Jost, John T., Banaji, Mahzarin R., & Nosek, Brian A. (2004). A Decade of System Justification Theory: Accumulated Evidence of Conscious and Unconscious Bolstering of the Status Quo. *International Society of Political Psychology*. 25 (6): 881–919. doi:10.1111/j.1467-9221.2004.00402

Kobrin, J. L., Patterson, B. F., Shaw, E. J., Mattern, K. D., & Barbuti, S. M. (2008). *Validity of the SAT® for Predicting First-Year College Grade Point Average* (College Board Research Report No. 2008-5). New York: The College Board. Retrieved from: http://research.collegeboard.org/rr2008-5.pdf.

Fuenmayor, A., Hetts, J.J, & Rothstein, K. (2012, April). *Assessing assessment: Evaluating models of assessment and placement*. Paper presented at the annual conference of the Research and Planning Group, Pomona, CA.

Kane, M. T. (2006) Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17-64). Westport, CT: American Council on Education/Praeger.

Melguizo, T., Kosiewicz, H., Prather, G., & Bos, H. (2014). How are community college students assessed and placed in developmental math? Grounding our understanding in reality. *Journal of Higher Education*, 85, 691-722.

Messick. S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 134-103). New York, NY: American Council on Education and Macmillan.

Messick, S. (1998). Test validity: A matter of consequences. *Social Indicators Research*, 45, 35-44.

Mislevy, R. J. (2006). Cognitive psychology and educational assessment. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 257-305). Westport, CT: American Council on Education/Praeger.

Mislevy, R. J. (2007). Validity by design. *Educational Researcher* 36, 463–469.

Nickerson, Raymond S. (1998). Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology*. 2 (2): 175–220. doi:10.1037/1089-2680.2.2.175

Ngo, F. & Kwon, W.W. (2015). Using multiple measures to make math placement decisions: Implications for access and success in community colleges. *Research in Higher Education*. 56: 442. doi:10.1007/s11162-014-9352-9

Otte, G. & Mlynarczyk, W. (2010). *Basic Writing*. West Lafayette, IN: Parlor Press and WAS Clearinghouse.

Scott-Clayton, J. (2012). *Do High-Stakes Placement Exams Predict College Success?* CCRC Working Paper No. 41.

Scott-Clayton, J., Crosta, P.M., & Belfield, C.R. (2014). Improving the Targeting of Treatment: Evidence from College Remediation. *Educational Evaluation and Policy Analysis*, vol. 36, no. 3.

U.S. Department of Labor. Employment and Training Administration (1999). *Testing and Assessment: An Employer's Guide to Good Practices*. Washington, DC: Author. Retrieved from: http://uniformguidelines.com/testassess.pdf

Westrick, P.A. & Allen, J. (2014). *Validity Evidence for ACT Compass® Placement Tests*. RR2014-2.

Willett, T. (2013). *Student Transcript-Enhanced Placement Study (STEPS) Technical Report.* Berkeley, CA: The Research and Planning Group for California Community Colleges.

Willett, T., Hayward, C., & Dahlstrom, E. (2008). *An early alert system for the remediation needs of entering community college students: Leveraging the California Standards Test. Report* 2007036. Encinitas, CA: California Partnership for Achieving Student Success.

Willett, T., Hayward, C., Nguyen, A., Newell, M., Bahr, P., Hetts, J., Lamoree, D., Sorey, K., & Duran, D. (2015). *Multiple Measures Assessment Project (MMAP) Spring 2015 Technical Report*. Sacramento, California: Research and Planning Group for California Community Colleges.

---

[i] Multiple measures are defined in Title 5 section 55502 (i): "Multiple measures" are a required component of a district's assessment system and refer to the use of more than one assessment measure in order to assess the student. Other measures that may comprise multiple measures include, but are not limited to, interviews, holistic scoring processes, attitude surveys, vocational or career aptitude and interest inventories, high school or college transcripts, specialized certificates or licenses, education and employment histories, and military training and experience. The requirement to use multiple measures when using a placement test is in section 55522.

[ii] The clearest guidance regarding multiple measures can be found in the 2005 Frequently Asked Questions document, which indicates that multiple measures themselves do not have to be validated (CCCCO, 2005). However, any type of *placement* validation must still take the influence of a multiple measure into account.

[iii] For example, if the multiple measure is on a ratio level of measurement (e.g., high school GPA) and the criterion is pass/fail (coded as 1/0), the technique would a point-biserial correlation.

[iv] Disjunctive placement is when a student is placed at the highest placement recommended by either the test or the multiple measure. It is contrasted to compensatory placement or conjunctive placement.

[v] A correlation of 0.15 is typical of observed correlations between placement tests and later success with a "C-" or greater in in college coursework. It therefore represents a fair estimate of a minimum expected level of utility.

Certainly, if an instrument can demonstrate a predictive validity of 0.35, it should be considered a highly useful predictor. However, lower levels of predictive validity could still be useful *particularly when combined with other assessment information*. The recommendation of 0.15 as a floor comes from the literature as well as the MMAP team's research into the correlations of placement tests (including ACCUPLACER) and success in college coursework. MMAP analyses have found predictive validity correlations for ACCUPLACER in the range of 0.10 to 0.21 (see, e.g., http://bit.ly/MMAPatCAP).

In practice, a range of predictive validity correlations are typically observed, as shown in the MMAP research. Hughes and Scott-Clayton (2011) found correlations in line with what the MMAP research team found for ACCUPLACER's predictive validity: 0.10 to 0.13 for English and 0.23 to 0.25 for math. Burton and Ramist (2001) report results from a meta-analysis of studies predicting cumulative undergraduate GPA and graduation probabilities using SAT scores and high school grades. The findings indicate that the average adjusted correlation of verbal and math SAT scores with cumulative college GPA is 0.36 compared to a correlation of 0.42 for high school grades with college GPA. The College Board's research shows that the unadjusted correlation between the SAT and college grades is between 0.26 and 0.33 (Kobrin, Patterson, Shaw, Mattern, and Barbuti, 2008). Another meta-analysis found an average correlation of 0.17 across a number of post-secondary predictor-outcome combinations (McManus, Dewberry, Nicholson, Dowell, Woolf and Potts, 2013). Taken as a whole, research on predictive validity would suggest that 0.35 is near the upper limit of observed predictive validity correlations and as such, a lower value in line with the observed correlations of existing tests such as the ACCUPLACER, would better serve as a minimum standard of utility.

Ultimately, the context of the application is going to determine if a given correlation is useful or not. It is more important to determine if a given predictor can improve student outcomes, in combination with other predictors than to set a minimum threshold and reject many potential measures outright. Even a measure with a relatively low predictive validity could meaningfully improve overall predictive validity if it is uncorrelated with the other predictor(s) in the model (i.e., it is explaining unique variance). From a more practical, applied perspective, the US Department of Labor produced a document that contains the following practical guidelines for the utility of tests based on the strength of their predictive validity: (1) above 0.35–very beneficial, (2) 0.21 - 0.35–likely to be useful, (3) 0.11 - 0.20–depends on circumstances, and (4) below 0.11–unlikely to be useful (U.S. Department of Labor, 1999).

Fundamentally, the usefulness of a test or measure should be determined by whether it improves upon existing practice rather than by comparing it to a necessarily arbitrary cut-off score. That being said, it may be useful to establish a predictive validity floor (e.g., 0.15) below which it is unlikely that a measure will make a meaningful contribution to improving the overall accuracy of placement.

[vi] Other model fit metrics such as the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) can also shed light on the balance between predictive validity and parsimony.

[vii] Note that both Reading and ESL sequences may have transfer-level English as their gateway course (students with goals that explicitly do not include completion of transfer-level English can be treated as a distinct subpopulation).

[viii] The expected first-year throughput rate of a cohort of incoming students can be calculated by adding the expected one-year throughput of those placed directly into transfer level [(N transfer-placed * enrollment rate in first term * Past success rate in class) + (number enrolling in second term * success rate in second term)] to the throughput rate of those placed into one-level below developmental coursework [(N placed one level below who enrolled in first term * historical one year completion rate of transfer-level course)].

[ix] One caveat to consider when optimizing transfer-level throughput concerns students whose educational goals require college-level but not transfer-level math and/or English. Such students could be identified by their educational goals and/or programs of study and then excluded from the throughput validation analysis.