RESEARCH & PRACTICE IN ASSESSMENT

VOLUME TWELVE | WINTER 2017
www.RPAjournal.com
ISSN# 2161-4210

A PUBLICATION OF THE VIRGINIA ASSESSMENT GROUP

RESEARCH & PRACTICE IN ASSESSMENT

VOLUME TWELVE | WINTER 2017
www.RPAjournal.com
ISSN# 2161-4210

CALL FOR PAPERS

Research & Practice in Assessment invites you to submit a high-quality manuscript related to various higher education assessment themes, and that adopt either an assessment measurement or an assessment policy/foundations framework. Manuscripts should be submitted electronically through the RPA submissions portal and must comply with the RPA submission guidelines. Click the links below for more information.

RPA Submission Portal: https://rpa.scholasticahq.com/for-authors

RPA Submission Guidelines: http://www.rpajournal.com/authors/

RESEARCH & PRACTICE IN ASSESSMENT

The goal of *Research & Practice in Assessment* is to serve the assessment community as an online journal focusing on higher education assessment. It is dedicated to the advancement of scholarly discussion amongst researchers and practitioners in this evolving field. The journal originated from the Board of the Virginia Assessment Group, one of the oldest continuing professional higher education assessment organizations in the United States. *Research & Practice in Assessment* is a peer-reviewed publication that uses a double-blind review process. Approximately forty percent of submissions are accepted for issues that are published twice annually. *Research & Practice in Assessment* is listed in Cabell's Directory and indexed by EBSCO, ERIC, Gale, and ProQuest.

Published by:
VIRGINIA ASSESSMENT GROUP | www.virginiaassessment.org

Publication Design by Patrice Brown | Copyright © 2017

TABLE OF CONTENTS

FROM THE EDITOR

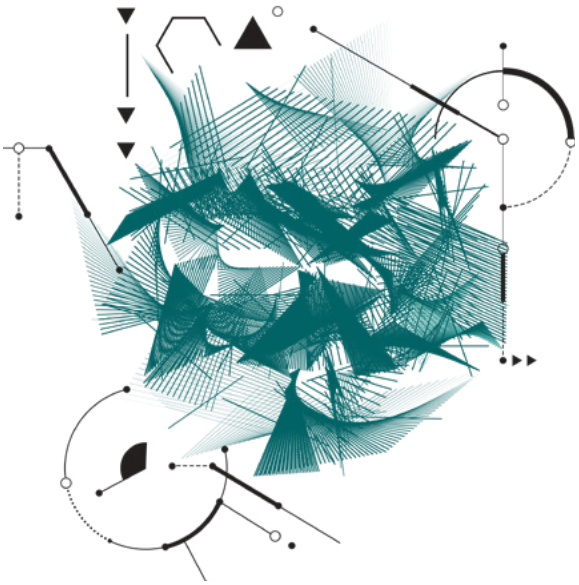
Advancing Assessment

Faculty, assessment practitioners and scholars, student affairs educators, and administrative leaders all have the opportunity to advance assessment efforts. Some may apply new techniques to collect better data about student learning and development. Others may provide the necessary leadership and support to enable assessment activities to occur. While still others may examine ways to improve uses of data for learning improvement. These collective efforts, in classrooms, on campuses, across institutions, in the United States, and throughout the world exemplify progress in understanding, measuring, and improving student learning and development. The contributions presented in this issue of *Research & Practice in Assessment* demonstrate important advancements in the practice and scholarship of assessment.

The Winter 2017 issue of RPA includes five peer-reviewed articles that exhibit the importance of data-informed practices as well as the importance of attitudes and approaches toward assessment. Using readily available data from IPEDS, Myers and Myers examine the impact of dual enrollment credit on graduation rates. Fuller and Lane present an analysis of the *Student Affairs Survey of Assessment Culture*, a measure of assessment cultures within divisions of student affairs. The importance of clearly defining terminology in assessment practices is exemplified by Gansemer-Topf, Downey, and Genschel through their work in operationalizing "academic undermatching" in college selection. Stevenson, Finan, and Martel also study assessment climate using a developmental approach. Roberts, Nardone, and Bridges examine writing proficiency by gender and race to determine if differences exist.

Lau reviews *Real-Time Student Assessment: Meeting the Imperative for Improved Time to Degree, Closing the Opportunity Gap, and Assuring Student Competencies for the 21st Century Needs*, by Peggy Maki, a text that challenges faculty and practitioners to utilize real-time assessment techniques to improve learning and development in real-time.

This issue also includes three Notes in Brief exemplifying impressive assessment practices. Smith, Finney, and Fulcher provide readers with actionable steps for incorporating implementation fidelity in their own assessment practice. Steinke and Fitch demonstrate a method for addressing the challenge of bias when assessing student work. Welch, Karpen, Cross, and LeBlanc examine practical uses of Bloom's Taxonomy in assessment work. I hope you find the scholarship in this issue will advance your own assessment efforts.



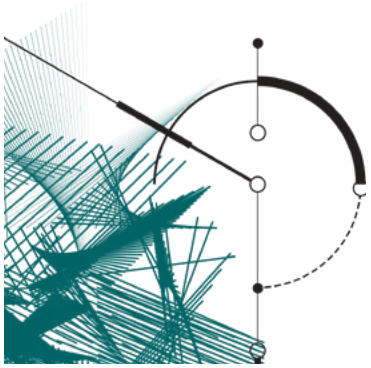
Regards,

Katie Busby

University of Mississippi

Abstract

Previous studies have found that freshmen who enter college with dual enrollment credits earned during high school have higher 6-year graduation rates. Yet, we do not know if institutional graduation rates benefit in the aggregate from their practice of accepting dual enrollment credits among incoming freshman cohorts. In this study, we used institutional panel data from the Integrated Postsecondary Education Data System and the 2006, 2007, and 2008 incoming freshman cohorts to address this policy issue. Based on regression results from generalized linear models, we found a contradictory pattern for the relationship between the institutional practice of accepting dual enrollment credits and graduation rates. Among the lesser selective institutions, those that accepted dual enrollment credits among their incoming freshmen realized higher 6-year graduation rates. But among the more selective institutions, this same practice was associated with lower 6-year graduation rates.



AUTHORS

Carrie B. Myers, Ph.D.
Montana State University

Scott M. Myers, Ph.D.
Montana State University

Dual Enrollment Policies and Undergraduate Rates in the United States: An Institutional and Cohert Approach Using the 2006–2014 IPEDS

Graduation rates from four-year colleges in the United States have risen significantly in the past 40 years but recent data indicate that these improvements have stagnated. Indeed, 6-year graduation rates for males since 2006 have stalled at around 55-56% and those for females at about 60-62% with even lower stalled rates for racial minorities (NCES, 2016). As a result, higher education institutions are increasingly being held accountable for institutional-level graduation rates that are assessed as indicators of best practices, institutional success, and major inputs in performance funding models (Heck, Lam, & Thomas, 2014; Rabovsky, 2014).

The Integrated Postsecondary Education Data System (IPEDS) used in this research was itself developed to help institutions comply with the 1990 federal accountability policy, "Student Right-to-Know and Campus Security Act" (SRK). This act requires 2- and 4-year institutions eligible for Title IV funding to assess and report yearly graduation rates for an incoming freshman cohort, where successful graduation is defined as within 150% of normal time—three years at a 2-year institution and six years at a 4-year institution. These graduation rates are often seen as measures of institutional effectiveness. Russell (2011) argued that this assessed accountability vis-à-vis graduation rates is embedded in the "college completion agenda" promoted by stakeholders such as the College Board and the Bill and Melinda Gates and the Lumina Foundations.

These developments lead institutions to continually seek and implement best policies and practices that assessment research has shown to increase graduation rates, especially among students who are traditionally at a higher risk of incompletion. The goal of these policies is to encourage greater college preparedness and readiness that are strong predictors of college success where the activities that best achieve this are high-school students taking college-level courses and earning college credits prior to enrollment (Struhl & Vargas, 2012). Advanced Placement (AP) courses historically served this purpose, and

CORRESPONDENCE

Email
cbmyers@montana.edu

college admission policies often use AP scores as a proxy for college preparation and future achievement (Jackson, 2010). More recently, higher-education institutions have partnered with high schools to offer their pre-college students the ability to earn college credit through dual enrollment (DE) courses, and when passed, these courses are accepted by some higher-education institutions as college credits. This practice is consistent with a main recommendation by Conley (2005) for improving the college readiness and success of students. Namely, there needs to be better articulation between high school and college curricula.

These institutions accept DE credits as college credits partly because it has been shown that if a DE course is well structured and provides authentic college-level learning and socializing experiences then students who take such a DE course may experience increased college preparedness and better college outcomes (Allen, 2010; Blackboard Institute, 2010; Karp, 2012). Indeed, at the individual level, research has found that students who enter college with DE credits have better college outcomes—net of selection effects—including higher GPAs, 1-year persistence rates, and, most importantly, 6-year graduation rates. Further, the effects of DE were stronger for at-risk students and those in less selective institutions (An, 2013b, 2015; Lerner & Brand, 2006).

What has not yet been assessed is whether DE credits influence graduation rates above and beyond these individual-level effects. That is, does an institution benefit as a whole when they have an admissions policy that accepts earned DE college credits among incoming freshmen?

What has not yet been assessed is whether DE credits influence graduation rates above and beyond these individual-level effects. That is, does an institution benefit as a whole when they have an admissions policy that accepts earned DE college credits among incoming freshmen? To the best of our knowledge, our research is the first to statistically assess the link between DE credit policy and graduation rates at the *institutional* level. We situate our study within the logic of Astin's Input-Environment-Output model (Astin, 1991; Astin & antonio, 2012). In this study, we focus on 4-year institutions and institutional cohort graduation rates from the incoming freshman classes of 2006 – 2008. We adopt this approach to guard against the substantial differences between 2-year and 4-year institutions (Newell, 2014) and to allow our findings to be compared to other studies who had identical selection factors and used cohort rates, as these are required by the SRK and are the only common metrics for comparing rates across the wide array of 4-year institutions in the United States (Bound, Lovenheim, & Turner, 2012; Hess, Schneider, Carey, & Kelly, 2009). Ideally, our research would also test the relative effects of DE versus those of AP policies. However, in 2006–2008 over 98% of our 4-year institutions had an admissions policy of accepting AP credits, and IPEDS does not contain any specifics about required minimum AP exam scores to receive college credit. Therefore, within a regression format we cannot assess the effects of having an AP policy as this variable is virtually a constant across our institutions.

Conceptual Approach and Research Hypotheses

Much like the AP program, the goal of a well-structured DE course is to provide high-school students with a more rigorous classroom experience, the opportunity to experience an authentic college-level course, and to earn college credits prior to enrolling in a higher-education institution. The most recent data from the 2010–11 academic year showed that 82% of public high schools had students who took DE courses, with over 1.44 million students enrolled in just over 2 million DE courses (Thomas, Marken, Gray, Lewis, & Ralph, 2013). About 16% of first-time, full-time students entered 4-year institutions in Fall 2008 with college credits from at least one DE course (Shapiro, Dundar, Yuan, Harrell, & Wakhungu, 2014).

Research has found that students who take well-structured and authentic DE courses are exposed to a potentially more rigorous, accelerated, and college-level curriculum that prepares them academically and socially for college, reduces the need for remedial courses, accelerates the earning of credits, and lowers the financial costs (Allen, 2010; An, 2013a; Bound, Lovenheim, & Turner, 2010; Karp, 2012; Lerner & Brand, 2006). Thus, DE courses provide not only an accelerated college preparation and credit but also a potentially more authentic and socializing college experience (Speroni, 2011a). A longitudinal qualitative study by Karp (2012) revealed that high-school students who take DE courses at a community college generally reported that these courses were an authentic college experience and allowed them to learn about the expectations and roles of a college student and actively practice these behavioral expectations. For these reasons, research has found that students who earned DE

credits in high school had better first-year and overall GPAs, better course sequencing, less major switching, more credits earned in the first year, and shorter times to degree completion. As a result, these students experienced greater retention and graduation rates after adjusting for a range of demographic, SES, and academic covariates (Allen, 2010; An, 2013b, 2015; Jackson, 2010; Speroni, 2011a, 2011b). The influence of DE is largely due to better college preparation, entering with college credits, the socializing and learning effects of a DE course, and greater academic motivation and engagement among these students (An, 2013a, 2015; Speroni, 2011a).

But our research seeks to answer a more macro question: do institutions benefit in the aggregate by accepting DE credits? That is, is it possible that institutions attain boosts in graduation rates from admitting students with DE credits above and beyond the summed superior outcomes among these students themselves? This institutional phenomenon would be possible if the presence of students with DE credits provides meta-individual processes and environments that benefit their fellow students without DE credits. We posit that peer effects in educational settings is the meta-individual environment and conceptual link between an institution's policy of accepting college credit for DE courses and its graduation rates. Our conceptual use of peer effects is based on Coleman et al. (1966) who found that a student with access to higher-achieving peers performed better academically, which led to the school performing better aggregately. Recent literature in higher education argues students who are exposed to higher-achieving and better-prepared peers will have enhanced academic outcomes as they have access to peer networks that provide a combination of social, human, and cultural capital resources (Booij, Leuven, & Oosterbeek, 2015; Conley, Mehta, Stinebrickner, & Stinebrickner, 2015; Estell & Perdue, 2013; Nechyba, 2006). Our review above on the relationship between DE and college behaviors and outcomes suggests that those who enter college with DE credits were more college ready, motivated, and successful than those without DE credits, and thus could fit the label of "higher-achieving peers."

The reader may link our approach to the logic of Astin's Input-Environment-Output model that is often used in educational research to organize longitudinal data and study academic outcomes (Astin & antonio, 2012). In this model, students enter college with specific Input (I) backgrounds and academic characteristics and capabilities that are partly a reflection of institutional characteristics (e.g., selectivity) given that students self-select themselves into these institutions (Tinto, 2012). Once in college, they enter into academic, social, and co-curricular Environments (E) that influence their learning and progress. Environments derive largely from institutional policies, decisions, resources, and practices that shape the educational settings and experiences in which students come into contact. The I-E-O model places an emphasis on academic and co-curricular engagements and interactions and the environmental influence of a student's peer group or "the characteristics of the student's peer group" (Astin, 1991, p. 92). At the end of this process, the Outcomes (O) or the consequences of these environments and inputs are measured, where the behavioral outcomes are usually academic progress and completion.

We posit that peer effects in educational settings is the meta-individual environment and conceptual link between an institution's policy of accepting college credit for DE courses and its graduation rates.

For our study, students either do or do not bring with them DE credits, experiences, and DE-related benefits (i.e., inputs). Above we argued that it would be expected that institutions that accepted DE credits would also enroll more incoming freshmen with these credits. Thus, those institutions who accept DE credits may have a peer environment (E) created by these students who bring with them the benefits that stem from DE. This peer environment may also benefit students who themselves bring with them DE credits. As the I-E-O model places an emphasis on interactions in these environments, we reason that these institutions who accept DE credits would realize superior outcomes (O) in terms of 6-year graduation rates compared to their institutional counterparts who do not accept DE credits.

The IPEDS data do not collect the requisite data to directly test peer effects so our approach is heuristic. The main limitation of the IPEDS survey is that it measures with a binary variable whether the institution has a policy of accepting DE credits (yes or no). The IPEDS submission does not query the institutions beyond this binary measure, such as the percentage of students who enter the institution with DE credits. However, we do assume that schools that accept DE credits will enroll a larger percentage of freshmen with DE credits

compared to institutions that do not accept DE credits and therefore are more likely to have an environment that fosters peer effects. We make this assumption for two reasons. First, research has found that a motivating factor in taking a DE course was how college credits influence a quicker time-to-degree pace and a lower financial burden. As such, students in DE courses reported being aware of college and university policies regarding the acceptance of credits earned through DE and AP courses (Smith et al., 2007). No research exists for DE, but the College Board (2014) found over half of the AP students surveyed reported they would be less likely to apply to a college or university that did not give credit for AP exam scores. Second, the College Choice Model recognizes that parents and students make choices strongly based on the price of college and the extent to which their student will achieve success at the chosen college (Niu & Tienda, 2008). This indicates that students with DE credits may be more likely to consider a college that accepts DE credits due to the lower cost and higher utility realized from the positive effects of entering college with college credits.

This indicates that students with DE credits may be more likely to consider a college that accepts DE credits due to the lower cost and higher utility realized from the positive effects of entering college with college credits.

The research we cited earlier showed that incoming freshmen with DE credits were, on average, more academically and socially prepared for college, more academically motivated and engaged once in college, and more likely to graduate compared to those without DE credits. Thus, situating peer effects theory with the logic of the I-E-O model suggests that students who enter without DE credits would benefit academically from the exposure to and interaction with students with DE credits and what these students bring with them to college above and beyond college credits (e.g., socializing experiences, academic preparation, motivation, and engagement). If so, then institutional graduation rates should be higher at institutions that accept DE credits above and beyond the cumulative individual-level effects of entering college with DE credits. We propose the following hypotheses:

- Hypothesis 1: Institutions that accept DE credits will have higher 6-year graduation rates compared to those that do not accept DE credits.
- Hypothesis 2: For peer effects to be supported, we hypothesize that the graduation rates among institutions who accept DE credits will be greater than the cumulative contributions of individual-level effects.

Lastly, emerging research has found that the positive benefits of college credits earned in high school on college outcomes were often greater for those at less selective institutions (An, 2013a, 2013b, 2015). Our third hypothesis is as follows:

- Hypotheses 3: The positive effects of a DE policy on graduation rates will be greater at less selective institutions.

Methods

Data

We used institutional-level panel data from the 2006–2009 and 2012–2014 Integrated Postsecondary Education Data System (IPEDS) to track the 6-year graduation rates of the incoming 2006, 2007, and 2008 cohorts of freshmen. We chose to analyze the most recent three cohorts for which IPEDS final release graduation data were available to better reduce any biases or distinctive results that may emerge from a single cohort. IPEDS collects data from post-secondary institutions in the United States (the 50 states and the District of Columbia) and other jurisdictions, such as Puerto Rico. Participation in IPEDS is a requirement for the institutions that partake in Title IV federal student financial aid programs such as Pell Grants or Stafford Loans during the academic year.

The IPEDS definition of "cohort" refers to full-time, first-time, degree-seeking students. We followed prior studies on graduation rates by limiting our institutions to those that were eligible for Title IV funding, enrolled at least 50 full-time freshmen in 2006–2008, granted bachelor degrees, were not-for-profit, were 4-year institutions, provided graduation data in the 2012–2014 Graduation Rate Survey, had complete data on the institutional measures, and had a Barron's selectivity ranking (Bound et al., 2010; Hess et al., 2009). Our analytic sample included 1,370 institutions that met these specifications for all cohorts, which resulted in 4,110 institution observations.

Focal Study Variables

The study variables are presented in Table 1. The 6-year graduation rate was measured in 2012, 2013, and 2014 and represented the percentage of the 2006, 2007, and 2008 cohort, respectively, who earned their bachelor's degree within 150% normal time. We used the 6-year institutional rate given its inclusion in federal acts, as a common measure of institutional effectiveness, and use in past research. The focal independent variable was measured in 2006, 2007, and 2008 for the three freshman cohorts and indicated (1 = yes; 0 = no) whether the institution had a policy of granting college credit for DE courses passed while in high school. Lastly, we rated each institution's academic admissions selectivity with the Barron's Selectivity Score that ranged from 1 = noncompetitive to 6 = most competitive. This selectivity score classified the admissions competitiveness of each institution using criteria such as median SAT or composite ACT scores, GPA of the incoming freshman cohort, class rank, and acceptance rates.

Institutional Covariates

A parsimonious set of institutional covariates other than selectivity consistently predicted most of the differences in graduation rates (Hess et al., 2009; Shin, 2010). We used this set as covariates and tagged them to the entry year of each cohort. They included: (a) the 1-year retention rate for each cohort representing the percentage of incoming freshmen who returned for their second year; (b) four categories of the log of yearly institutional expenditures (in U.S. dollars) per full-time equivalent (FTE) student that tapped instructional, academic support, student service, and research; (c) the percent of undergraduate students at the institution that received federal aid, which is often used as a proxy for the extent of low-income students at the institution as Pell Grants comprise the largest share of federal aid (NCES, 2016); (d) the log of the ratio of FTE undergraduate students per full-time faculty with instructional duties; (e) control indicating whether the institution was private or public; (f) the 2005 Carnegie classification that coded the institution as doctoral, master's, or baccalaureate with doctoral institutions serving as the reference category; and (g) the log of the percent of full-time freshmen that were classified as non-White, non-Asian. The variables that were log transformed were done so to reduce issues of skewness that were identified with regression diagnostics, to allow for nonlinear relationships with graduation rates, and to make the model more efficient when estimating standard errors. These log transformations are common in studies that examine institutional covariates whose values vary considerably across institutions (Griffith & Rask, 2016; Webber & Ehrenberg, 2010). We also used the broad category of non-White, non-Asian to allow our findings to be compared to other research on graduation rates that use the IPEDS data.

Thus, situating peer effects theory with the logic of the I-E-O model suggests that students who enter without DE credits would benefit academically from the exposure to and interaction with students with DE credits and what these students bring with them to college above and beyond college credits (e.g., socializing experiences, academic preparation, motivation, and engagement).

Statistical Analyses

There were statistical concerns inherent in the data that prevented the use of ordinary least squares regression estimations. First, the White test and plots of residuals versus fitted values indicated the presence of heteroscedasticity. Second, a variety of tests including Cook's d, studentized residuals, probability plots, and DFBeta revealed several institutions to be influential observations. However, additional diagnostics using variance inflation factors and tolerance diagnostics found no multicollinearity among the independent and control variables. Furthermore, the distribution of the graduation rate measures did not violate assumptions of normality. Third, our analytic sample included 1,370 institutions across the 50 states and the District of Columbia—suggesting a nested or clustering data structure even though IPEDS does not employ any nested or cluster sampling. Pennsylvania had the most institutions included at 105 whereas Wyoming had only a single institution in the analytic sample. Therefore, it was possible that institutional rates of graduation were correlated within states given the role of states in funding and legislating higher education and the wide and growing disparities in these funding levels (Mitchell, Palacios, & Leachman, 2014). Also, since we had three measures of graduation rates for individual institutions, these outcomes were undoubtedly correlated as well.

To correct for all these issues, we estimated quasi-likelihood regression parameters with generalized linear models (GLM) and general estimating equations (GEE). We choose this

Table 1

Description of Study Variables by Cohert: IPEDS 2006–2014 (n=4,110)

Variables	Range / Coding	M	SD
6-year graduation rate	1%–99% of incoming freshmen graduating within 6 years	53.95%	18.21
Dual enrollment policy	0 = do not accept DE credits; 1 = accept DE credits	0.82	---
Selectivity	1 = noncompetitive to 6 = most competitive	3.41	1.09
Institutional Controls			
1-year retention rate	5%–100% of incoming freshmen returned in 2nd year	74.29%	11.78
Private institution	0 = no; 1 = yes	0.62	---
Doctorate (Reference)	0 = no; 1 = yes	0.20	---
Master's	0 = no; 1 = yes	0.41	---
Baccalaureate	0 = no; 1 = yes	0.39	---
% of undergraduate students that are non-White, non-Asian	0%–100% are non-White, non-Asian	19.19%	22.04
% of undergraduates receiving federal aid	1%–100% receive federal aid	31.93%	18.60
FTE student/faculty ratio	4 FTE student/faculty to 128 FTE student/faculty	16.13	7.29
Instructional expenses per FTE student	$0 per FTE student to $75,776 per FTE student	$8,435	$6,926
Research expenses per FTE student	$0 per STE student to $98,726 per FTE student	$1,627	$6,159
Academic expenses per FTE	$0 per FTE student to $54,320 per FTE student	$2,271	$2,645
Support expenses per FTE	$0 per FTE student to $47,221 per FTE student	$2,768	$2,065

approach as the GLM/GEE model assumes that the data are longitudinal and the repeated outcome measures (i.e., graduation rates) are correlated within institutions over time. To further handle the longitudinal and correlated data and the statistical issues reported above, we conducted two additional procedures. First, we adjusted all equations with an exchangeable working correlation structure among the observations due to state-level clustering and correlated outcomes, which was preferable to an autoregressive (AR-1) structure as the outcomes were measured only one year apart. Second, we calculated the standard errors with an asymptotic covariance matrix to produce robust error estimates that provided much more conservative estimates of Z- and p-values. Finally, we followed prior econometric recommendations and research and transformed our fractional response outcome of graduation rates into the log of odds ratios. This is necessary for outcomes that are bounded by 0 and 1 as untransformed variables may return regression equations that predict values less than 0 and greater than 1 (Baum, 2008).

The quasi-likelihood GLM equation was:

$$1n\left(\frac{y_{it}}{1-y_{it}}\right) = G\left(a_0 + b_1 DE_{it} + bx_{it} + u_{it}\right)$$

Where y_{it} was the 6-year graduation rate of school i as of year t for students who entered the institution as a full-time first-year student six years earlier. This outcome was then modeled as a function of the estimated coefficients for whether the institution has a policy of accepting DE credits ($b_1 DE_{it}$), a vector of institutional-level control variables (bx_{it}), and a random error term (u_{it}). G indicates that these parameters were estimated with quasi-likelihood equations.

We used this equation to estimate three regression models. The first two models attempted to capture the chronological nature of the I-E-O approach: (a) a baseline model that included only DE policy and cohort year and (b) a full model that added in all the institutional controls that measured institutional characteristics that would influence Inputs as well as the Environments encountered by students. The third model adds an interaction term crossing DE by selectivity. For GEE, the appropriate model fit statistic is QIC (quasi-likelihood under the independence model criterion) as the more common AIC is not available for GEE since it is not likelihood based (Hardin & Hilbe, 2012). When comparing QIC statistics between models a smaller value indicates which model better captures the data. We used both QIC and QICu to address model fit because QICu adds a penalty parameter for the number of variables in the model, which awards better fitting and more parsimonious models.

We did find that less selective institutions benefit the most from accepting DE credits, but the negative effects of DE at the more selective institutions directly contradicted our expectations.

Results

The figures in Table 1 show the descriptive statistics for the three cohorts of freshmen. On average, institutions had a 6-year graduation rate of about 54%, and 82% of these institutions had a policy of accepting DE credits among their incoming freshmen. We also examined whether graduation rates, DE policy, and institutional covariates differed across the three cohorts (results available upon request). They did not. For example, for the 2006, 2007, and 2008 cohorts the graduation rates were 53.6%, 54.1%, and 54.1%, respectively. For DE policies, 82% of institutions accepted DE credits for the 2006 and 2007 cohorts whereas 83% did so for the 2008 cohort. Finally, there were no significant differences in the values of the institutional covariates for the three cohorts 2006–2008. Given our focus on DE policy and institutional selectivity, we also report our descriptive statistics across these two characteristics. In Table 2, we found small-to-moderate differences between institutions that accept DE credits and those that do not. Those that did not accept DE credits have higher graduation rates, perhaps because they are more selective, private, doctoral-granting, and have more institutional resources (Tinto, 2012). We found larger differences across selectivity levels with respect to graduation rates and DE policy. Most of the institutional covariates also varied with selectivity where the quality and quantity of the covariates increased with selectivity. These results are consistent with federal data that has found selectivity to be the strongest predictor of graduation rates (NCES, 2016; Tinto, 2012). Further, variations in our set of institutional characteristics have been found to explain about 75% of institutional differences in graduation rates (Hess et al., 2009; Shin, 2010). Thus, it was important to control for these institutional covariates to conservatively test the hypotheses and minimize omitted variable bias.

A set of regression estimates are in Table 3. The baseline results in Model 1 showed that 6-year graduation rates were lower among institutions that had a policy of accepting DE credits among the incoming freshman cohorts of 2006–2008. To test the robustness of this finding, we included all the institutional covariates in Model 2 where these covariates removed the negative association between DE and graduation rates shown in Model 1. Indeed, the DE coefficient failed to retain its directional effect size and reach statistical significance in Model 2. In Model 3, we tested whether the effect of DE varied by institutional selectivity. Here we found a negative and significant interaction between DE and institutional selectivity. An examination of the QIC and QICu fit statistics across the three models indicated that Model 3 was the best fitting model. Further, in Model 3, the QICu fit value approximated the QIC fit value (within 3% of each other) suggesting that the model was correctly specified.

Table 2

Mean Values of Study Variables by DE Policy and Selectivity: IPEDS 2006–2014 (n=4.110)

	DE Policy		Selectivity by Competitiveness (1 = non to 6 = most)					
Variables	Yes	No	1	2	3	4	5	6
6-year graduation rate	52.81%	64.20%	32.22%	40.52%	49.44%	63.61%	75.16%	86.79%
Dual enrollment	---	---	0.83	0.93	0.96	0.88	0.73	0.48
Selectivity	3.31	4.15	---	---	---	---	---	---
Institutional Controls								
1-year retention rate	73.64%	80.00%	63.55%	67.49%	71.55%	79.87%	87.22%	91.01%
Private	0.60	0.79	0.50	0.49	0.55	0.68	0.76	0.86
Doctorate (Reference)	0.19	0.24	0.04	0.09	0.17	0.31	0.36	0.51
Master's	0.43	0.25	0.42	0.53	0.53	0.36	0.17	0.05
Baccalaureate	0.38	0.51	0.54	0.38	0.34	0.33	0.47	0.44
% Non-White	21.00%	21.30%	59.93%	30.66%	19.52%	13.94%	11.29%	9.80%
% Receiving federal aid	39.08%	30.10%	44.21%	40.77%	34.63%	30.52%	24.48%	20.91%
FTE student/faculty	17.84	14.33	20.09	21.22	17.62	15.35	12.61	10.89
Instructional expenses per FTE	$7,481	$10,374	$4,957	$4,044	$4,400	$6,348	$9,368	$18,383
Research expenses per FTE	$1,911	$3,575	$1,397	$1,282	$1,444	$3,303	$5,449	$8,533
Academic expenses per FTE	$5,831	$7,465	$5,111	$4,958	$6,120	$7,377	$9,083	$13,259
Support expenses per FTE	$3,035	$4,386	$2,175	$2,496	$3,669	$4,361	$5,506	$7,477

For these reasons, it is suggested to interpret the best-fitting model (Harden & Hilbe, 2012). One main assumption of a GLM/GEE approach is that the underlying correlation structure is correctly chosen. For more support of our decision to use an exchangeable structure, we reestimated Model 3 with four different types of correlation structures: exchangeable, AR-1, unstructured, and independent. The results (available upon request) showed that the QIC and QICu statistics confirmed that exchangeable was indeed the best-fitting structure to the data.

In calculating the simple slopes from Model 3 (Aiken & West, 1991), the effects of DE on graduation rates were 0.49 for noncompetitive institutions, 0.29 for less competitive institutions, 0.09 for competitive institutions, -0.11 for very competitive institutions, -0.31 for highly competitive institutions, and -0.51 for the most competitive institutions. These simple slopes were statistically significant at $p \leq .05$ except for institutions that are competitive or very competitive. Therefore, we found that the effect of DE on 6-year graduation rates were positive for the lesser selective institutions and negative for the more selective institutions.

Table 3

Generalized Linear Model Regression Coefficients for 6-Year Graduation Rates: IPEDS Cohorts of 2006, 2007, and 2008 (n=4,110)

Variables	Model 1	Model 2	Model 3
Dual enrollment	-0.61***	-0.07	0.69***
	(0.12)	(0.06)	(0.09)
Dual enrollment *x* selectivity	---	---	-0.20***
			(0.02)
Cohort	0.01	0.01	0.01
	(0.01)	(0.01)	(0.01)
Selectivity	---	0.41***	0.54***
		(0.02)	(0.03)
Institutional Covariates			
1-year retention rate	---	0.05***	0.03***
		(0.00)	(0.00)
Private	---	0.31***	0.26***
		(0.06)	(0.05)
Doctorate (Reference)	---	---	---
Master's	---	-0.14***	-0.14***
		(0.04)	(0.04)
Baccalaureate	---	-0.15**	-0.17**
		(0.06)	(0.07)
% Non-White	---	-0.15***	-0.14***
		(0.04)	(0.03)
% Receiving federal aid	---	-0.01***	-0.02***
		(0.00)	(0.00)
FTE student/faculty	---	-0.11*	-0.09*
		(0.05)	(0.04)
Instructional expenses per FTE	---	0.13***	0.11***
		(0.02)	(0.02)
Research expenses per FTE	---	0.06***	0.04***
		(0.01)	(0.01)
Academic expenses per FTE	---	0.07	0.03
		(0.04)	(0.03)
Support expenses per FTE	---	0.10**	0.11**
		(0.03)	(0.03)
Intercept	0.23	-1.11	-1.89
QIC fit index (smaller is better)	4201.87	3877.23	3871.19
QICu fit index (smaller is better)	4172.00	3760.00	3754.00

Note: The 6-year graduation rates are log transformed. Robust Standard errors are in parentheses.
*p< .05. **p< .01. ***p<.001 (two-tailed)

Next, because our estimated model was nonlinear, a straight-forward interpretation of these slopes was not possible. Therefore, we calculated the marginal effect sizes for these significant interactions between DE and selectivity by following the two-step procedure in Webber & Ehrenberg (2010). First, we obtained a predicted value for 6-year retention rates for each institution. Second, we then took the difference between the averages of these predicted rates between institutions that accepted DE credits and those that did not. We found that among noncompetitive intuitions, 6-year graduation rates were 4.9 percentage points higher among those that had a policy of accepting DE credits. For less competitive institutions, accepting DE credits returned a 6-year graduation rate that was 3.5 percentage points higher than those without such a policy. Yet, for highly competitive and most competitive institutions, those that accepted DE credits had a 6-year graduation rate that was 2.8 and 4.4 percentage points lower, respectively, than similarly selective institutions without such a DE policy.

Summary

In this study, we assessed the effectiveness of the institutional practice of accepting dual enrollment (DE) credits among incoming freshman cohorts where the effectiveness outcome was measured with 6-year graduation rates. We positioned our study within the logic of Astin's I-E-O modeling where the Environment created by an institution's DE policy was theorized to be that of peer effect. We first found that institutional graduation rates were significantly lower at institutions that accept DE credits among incoming freshmen. However, our second finding was that our control variables removed the significant link between DE policies and lower graduation rates. Third, though, the best-fitting model found that the direction of the effects of DE on graduation rates depended on institutional selectivity where DE had a positive effect on these rates among the lesser selective institutions but a negative effect among the more selective institutions. Further, the positive effects were the greatest for the least selective and the negative effects are greatest for the most selective institutions.

Whether peer effects are the driving force can only be answered by future research that contains data on multiple levels (individuals and institutions) collected with multiple methods (qualitative and quantitative). Yet our findings do suggest that they may be a contributing mechanism and that research needs to more fully assess this possibility.

These results both supported and contradicted our research hypotheses. We expected that institutional policies of accepting DE credits would be associated with higher graduation rates, and we also expected that less selective institutions would benefit the most from earned college credits among their incoming freshman cohorts. We did find that less selective institutions benefit the most from accepting DE credits, but the negative effects of DE at the more selective institutions directly contradicted our expectations. There are no known studies at the institutional level between DE and institutional graduation rates. So, whereas our findings contradicted our hypotheses they did not contradict any existing research as that research does not exist. Below, we take care to offer several suggestions for our findings so that future research may untangle our results, especially research that adopts the I-E-O model and attempts to measure peer effects.

Evaluation of Hypotheses and Recommendations

The positive effects of DE among less selective institutions is consistent with research that shows DE is beneficial for individual graduation rates, especially at less selective institutions. Yet the negative effects of DE at more selective institutions are not consistent with individual-level research. For the positive effects, our hypothesis to be evaluated is whether these effects are above and beyond what we would expect from the cumulative effects of individual rates of graduation. That is, are peer effects potentially operating?

Research suggests that about 16% of incoming freshmen enter 4-year institutions with DE credit (Thomas et al., 2013). Further, An (2013b) provided the most recent national estimate of the effect of DE on graduation rates where students who entered 4-year institutions with DE credits had graduation rates that were 7 percentage points higher than those without such credits. Thus, at the institutional level, we would expect institutions who have a policy of accepting DE credits to realize an average 6-year graduation rate of 1.1 percentage points higher (7.0 x 0.16) that represents the cumulative contributions of individual-level effects. This is an average figure across all institutional types as there are no available data disaggregated across the six institutional selectivity levels.

Our positive effect sizes for 6-year graduation rates was 4.9 percentage points for noncompetitive institutions and 3.5 percentage points for less competitive institutions. These figures all exceed the expected 1.1 percentage point gain calculated above, suggesting that institutions do indeed benefit in the aggregate by having a policy of accepting DE credits. Whether peer effects are the driving force can only be answered by future research that contains data on multiple levels (individuals and institutions) collected with multiple methods (qualitative and quantitative). Yet our findings do suggest that they may be a contributing mechanism and that research needs to more fully assess this possibility. These future studies could also adopt the I-E-O approach to inform the selection of study variables.

The negative effects of DE at the more selective institutions is unanticipated as no prior research suggests such a finding. Future research will need to focus on several issues to advance our study and better inform institutions about the effects of their DE policies. First, not all DE experiences and credits are academically equal, which may influence whether DE

is an accurate proxy for the Inputs of college readiness and preparation and how it should be used as part of admissions policies. Indeed, research will need to consider two intersecting characteristics of DE programs: some high-school students take DE courses at a high school and not on a college campus, and some take DE courses with career and technical/vocational foci and not an academic focus. Research finds that the positive effects of DE credits on individual-level college outcomes is, on average, superior when high-school students take DE courses on college campuses and when these courses are academic focused (Speroni, 2011a, 2011b). Thus, future research will need to measure the DE profile of incoming freshman cohorts across institutional selectivity categories to tease out our findings. It is possible that institutions have incoming cohorts that differ in the type of and place where their DE credits were earned while in high school (i.e., Inputs), and that these differences in a cohort's DE profile across institutions could be a contributing explanation for our disparate findings.

Second, it will be important to consider the fit between the other academic and background characteristics of an incoming freshman cohort with DE credits and an institution's peer, academic, and structural characteristics (i.e., Inputs and Environments). For example, An (2015) showed that students who enter college with DE credits had lower ACT scores, came from households with lower parental education levels, and were more likely to be an underrepresented. Cowan and Goldhaber (2015) found that students who participated in DE programs in Washington State high schools are more likely to attend 2-year institutions compared to 4-year institutions and complete high school with a GED compared to their peers who did not take DE courses. While not directly measured, these findings suggest that students who earned DE credits may be less prepared academically for more selective institutions and partly explain our results. Now that more students are entering college with DE credits, institutions may be well served if they conduct institutional research assessing the academic trajectories of these students.

In this study, we attempted to account and correct for data and statistical issues that may have influenced our results. We did so by using a strong set of institutional covariates that have been shown to predict most of the differences in graduation rates between institutions, including measures of selectivity, expenditures, and proxies for stratifying student characteristics such as the percent of students receiving federal aid. We also employed regression and statistical techniques and adjustments that handled influential observations and correlated outcomes as well as produced conservative statistical tests of significance. Still, our research must be interpreted within the limitations of IPEDS.

First, it is possible that the data do not contain institutional characteristics that could further account for the effects of DE on graduation rates across institutional selectivity levels. Second, as per the SRK, we followed an institutional cohort over six years, where this cohort was comprised of first-time and full-time freshmen who stayed at the same institution. This definition covers about 25–30% of all college students in 4-year institutions, depending on the institution's characteristics, and does not take into account the academic outcomes after transferring out of the initial institution (Hess et al., 2009). Thus, our results are generalizable only to these types of students and do not capture the experiences of the other diverse set of students in higher education institutions in the United States. Lastly, future research will need to expand our binary measure of DE policy to include further information about an institution's policy, how the institution counts and applies DE credits, and the percentage of students who enter with DE credits. Saying this, our research does provide the first baseline statistical assessment on the relationships between an institution's admissions practice regarding college credit earned through DE programs and their graduation rates. This assessment provides a firm foundation for the development of future assessment research on the effects of institutional practices regarding DE credits among incoming students. Our research should also motivate institutions to analyze their DE practices and policies and to assess the academic inputs and outcomes of those students who enter college with DE credits.

This assessment provides a firm foundation for the development of future assessment research on the effects of institutional practices regarding DE credits among incoming students. Our research should also motivate institutions to analyze their DE practices and policies and to assess the academic inputs and outcomes of those students who enter college with DE credits.

References

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. London: Sage.

Allen, D. (2010). *Dual enrollment: A comprehensive literature review & bibliography*. New York, NY: CUNY.

An, B.P. (2013a). The influence of dual enrollment on academic performance and college readiness: Differences by socioeconomic status. *Research in Higher Education, 54*, 407–432.

An, B.P. (2013b). The impact of dual enrollment on college degree attainment: Do low-SES students benefit? *Educational Evaluation and Policy Analysis, 35*, 57–75.

An, B.P. (2015). The role of academic motivation and engagement on the relationship between dual enrollment and academic performance. *The Journal of Higher Education, 86*, 98–126.

Astin, A.W. (1991). *Assessment for excellence: The philosophy and practice of assessment and evaluation in higher education*. New York, NY: Macmillan.

Astin, A.W., & antonio, a.l. (2012). *Assessment for excellence: The philosophy and practice of assessment and evaluation in higher education* (2nd ed.). Lanham, MD: Rowman and Littlefield.

Baum, C.F. (2008). Modeling proportions. *Stata Journal, 8*, 299–303.

Blackboard Institute. (2010). *Dual enrollment: A strategy for educational advancement of all students*. Washington, DC: Blackboard Institute.

Booij, A.S., Leuven, E., & Oosterbeek, H. (2015). *Ability peer effects in university: Evidence from a randomized experiment*. IZA Discussion Paper No. 8769. Bonn, Germany: Institute for the Study of Labor.

Bound, J., Lovenheim, M. F., & Turner, S. (2012). Increasing time to baccalaureate degree in the United States. *Education Finance and Policy, 7*(4), 375–424.

Coleman, J.S., Campbell, E.Q., Hobson, C.J., McPartland, J., Mood, A.M., Weinfeld, F.D., & York, R.L. (1966). *Equality of educational opportunity*. Washington, DC: U.S. Government Printing Office.

Conley, D.T. (2005). *College knowledge: What it really takes for students to succeed and what we can do to get them ready*. San Francisco: Jossey-Bass.

Conley, T., Mehta, N., Stinebrickner, R., & Stinebrickner, T. (2015). *Social interactions, mechanisms, and equilibrium: Evidence from a model of study time and academic achievement*. NBER Working Paper 21418. Cambridge, MA: National Bureau of Economic Research.

Cowan, J., & Goldhaber, D. (2015). How much of a "running start" do dual enrollment programs provide students? *The Review of Higher Education, 38*(3), 425–460.

Estell, D.B., & Perdue, N.H. (2013). Social support and behavioral and affective school engagement: The effects of peers, parents, and teachers. *Psychology in the Schools, 50*, 325–339.

Griffith, A.L., & Rask, K.N. (2016). The effect of institutional expenditures on employment outcomes and earnings. *Economic Inquiry, 54*, 1931–35.

Hardin, J.W., & Hilbe, J.M. (2012). Generalized estimating equations. Boca Raton, FL: CRC Press.

Heck, R.H., Lam, W.S., & Thomas, S.L. (2014). State political culture, higher education spending indicators, and undergraduate graduation outcomes. *Educational Policy, 28*, 3–39.

Hess, F. M., Schneider, M., Carey, K., & Kelly, A.P. (2009). *Diplomas and dropouts: Which colleges actually graduate their students (and which don't)*. Washington, DC: American Enterprise Institute.

Jackson, C.K. (2010). *The effects of an incentive-based high-school intervention of college outcomes*. NBER Working Paper 15722. Cambridge, MA: National Bureau of Economic Research.

Karp, M. M. (2012). 'I don't know, I've never been to college!' Dual enrollment as a college readiness strategy. *New Directions in Higher Education, 158*, 21–28.

Lerner, J. B., & Brand, B. (2006). *The college ladder: Linking secondary and postsecondary education for success for all students*. Washington, DC: American Youth Policy Forum.

Mitchell, M., Palacios, V., & Leachman, M. (2014). *States are still funding higher education below pre-recession levels*. Washington, DC: Center of Budget and Policy Priorities.

National Center for Education Statistics. (2016). *Digest of education statistics: 2015 (NCES 2016-014)*. Washington, DC: Institute of Education Sciences.

Nechyba, T.J. (2006). Income and peer quality sorting in public and private school. *Handbook of the Economics of Education, 2*, 1327–1368.

Newell, M.A. (2014). What's a degree got to do with it? The civic engagement of associate's and bachelor's degree holders. *Journal of Higher Education Outreach and Engagement, 18*(2), 67–89.

Niu, S.X., & Tienda, M. (2008). Choosing colleges: Identifying and modeling choice sets. *Social Science Research, 37*, 416–433.

Rabovsky, T.M. (2014). Using data to manage for performance at public universities. *Public Administration Review, 64*, 1540–6210.

Russell, A. (2011). *A guide to major U.S. college completion initiatives (A higher education policy brief)*. Washington, DC: American Association of State Colleges and Universities.

Shapiro, D., Dundar, A., Yuan, X., Harrell, A., & Wakhungu, P.K. (2014). *Completing college: A national view of student attainment rates – Fall 2008 cohort*. Herndon, VA: NSCRC.

Shin, J.C. (2010). Impacts of performance-based accountability on institutional performance in the U.S. *Journal of Higher Education, 60*(1), 47–68.

Smith, M.A., Place, A.W., Biddle, J.R., Raisch, C.D, Johnson, S.L., & Wildenhaus, C. (2007). The Ohio Postsecondary Enrollment Opportunities (PSEO) Program: Understanding its under-utilization. *Journal of Educational Research & Policy Studies, 7*, 80–114.

Speroni, C. (2011a). *Determinants of students' success: The role of advanced placement and dual enrollment programs*. New York, NY: National Center for Postsecondary Research.

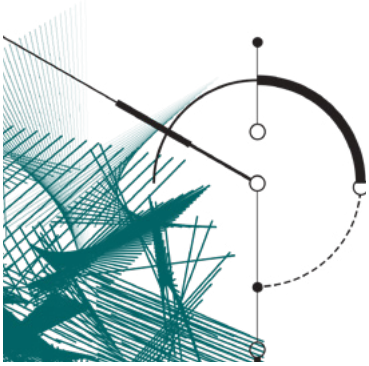
Speroni, C. (2011b). *High school dual enrollment programs: Are we fast-tracking students too fast?* New York, NY: National Center for Postsecondary Research.

Struhl, B., & Vargas, J. (2012). *Taking college courses in high school: A strategy for college readiness*. Washington, DC: Jobs for the Future.

Thomas, N., Marken, S., Gray, L., Lewis, L., & Ralph, J. (2013). *Dual credit and exam-based courses in U.S. public high schools: 2010-11*. Washington, DC: National Center for Education Statistics.

Tinto, V. (2012). *Completing college: Rethinking institutional action*. Chicago, IL: University of Chicago Press.

Webber, D.A., & Ehrenberg, R.G. (2010). Do expenditures other than instructional expenditures affect graduation and persistence rates in American higher education? *Economics of Education Review, 29*, 947–958.



Abstract

Student affairs, like all arms of academe, has taken up the mantle of assessing college student learning and development in their unique programs and experiences. Yet, cultures of assessment in student affairs organizations are rarely examined empirically. This study provides results from an exploratory factor analysis of data gathered using the *Student Affairs Survey of Assessment Culture*. The resulting factor model explained 58% of the variance and included four factors of hypothesized cultures of assessment in student affairs: a) Clear Commitment to Assessment, b) Assessment Communication, c) Connection to Change, and d) Fear of Assessment. Internal consistency estimates (Cronbach's α) were sufficient for each factor, exceeding .78, at minimum. Discussion about new means of theorizing about cultures of assessment in student affairs and pragmatic advice on leading student affairs assessment efforts are offered.

AUTHORS

Matther B. Fuller, Ph.D.
Sam Houston State University

Forrest C. Lane, Ph.D.
Sam Houston State University

An Empirical Model of Culture of Assessment in Student Affairs

Student affairs, like all arms of academe, has taken up the mantle of assessing college student learning and development in their unique programs and experiences. Best practices in assessing college student learning and development within student affairs contexts have emerged from this literature (Bingham & Bureau, 2015; Henning & Roberts, 2016; Schuh, Biddix, Dean, & Kinzie, 2016). One of those best practices includes the development of a culture of assessment in both institutional (Baas, Rhoads, & Thomas, 2016; Douchy, Segers, Gijbels, & Struyven, 2007; Haviland, 2014; Kuh, et al., 2015; Suskie, 2014) and student affairs contexts (University of Pittsburgh, 2012; Schuh, 2013). The concept of a "culture of assessment" has not only become commonplace parlance for presidents, provosts, and faculty; it is a term of considerable attention for vice presidents of student affairs, deans of students, and directors of student affairs departments as well.

Despite this attention, cultures of assessment in student affairs organizations are rarely examined empirically. This gap is considerably problematic since institutions purport to value the use of evidence to inform decision making. Without a functional, synthetic, data-driven foundation from which to theorize about cultures of assessment in higher education advancements in the practice of student affairs assessment will remain conjectural and relegated to the applications of current, trending best practices. As soon as a new assessment process comes into prominence, the community of student affairs practitioners will face decisions in redirecting and redefining the culture of their division. In contrast, divisions of student affairs practicing evidence-based approaches are purported to have sustainable, transformative, long-term cultures of assessment guiding them through many organizational challenges (Henning & Roberts, 2016; Schuh, 2013).

This study seeks to provide an empirical foundation for further research in student affairs organizations' cultures of assessment. The present analyses call upon empirical evidence to illustrate the foundations of the concept of cultures of assessment in student affairs contexts. Using the *Student Affairs Survey of Assessment Culture*, the researchers

CORRESPONDENCE

Email
mfuller@shsu.edu

examined the underlying factor structure inherent in the survey data. The *Student Affairs Survey* is an adaptation of the *Administrators Survey*, augmented for administration to mid-level student affairs leaders. The researchers explored the underlying structure using exploratory factor analysis (EFA) methods to determine if the *Student Affairs Survey* accurately measured hypothesized cultures of assessment. The results of this analysis may offer new abilities to theorize about cultures of assessment and offer practitioners opportunities to refine leadership of student affairs assessment. Discussion and theorization about future research and practice are offered after a comprehensive review of student affairs assessment literature, methods, and results.

Review of Relevant Literature on Cultures of Assessment in Student Affairs

Literature pertaining to assessment in student affairs is currently enjoying considerable attention in scholarly discourse. This growth in prominence is led by efforts of scholar-practitioners actively engaged in research and the conscious efforts of professional organizations such as the National Association of Student Personnel Administrators (NASPA), the American College Personnel Association (ACPA), and the Association for the Assessment of Learning in Higher Education (AALHE), among others. Moreover, the growth of staff members and departments whose sole purpose is the coordination or leadership of student affairs assessment efforts is also noteworthy (Roper, 2015).

Though assessment is now commonplace throughout many student affairs divisions and departments much remains to be done to examine how assessment becomes a foundational element of a student affairs division's culture.

Though assessment is now commonplace throughout many student affairs divisions and departments much remains to be done to examine how assessment becomes a foundational element of a student affairs division's culture. Long (2012) argued the necessity of a few unique characteristics for student affairs to be called a profession within higher education. Paramount in these defining characteristics is the presence of a number of graduate programs in student affairs and evaluation and assessment systems aimed at improving program effectiveness. Therefore, examining assessment's contribution to division-wide cultures of assessment is of critical importance and connects to larger discourses of the importance of student affairs in academe (Long, 2012).

Scholarship on assessment practices and their use in student affairs is a new phenomenon. As early as the 1980s and 1990s, scholars (Barr, 1993; Kuh & Banta, 2000) were recognizing that assessment methods most often employed in classrooms and academic programs held possibilities for assessment learning and development in co-curricular environments and programs. However, the developments throughout the 1990s and 2000s focused on enhancing the integration of academic and co-curricular efforts further heightened the importance of assessment in student affairs (Banta & Associates, 2002). Moreover, discourses critical of the importance of student affairs in modern academe have also contributed to the sense that student affairs must prove its worth and assessment has stood as the primary means through which this worth is proved (Kirschner, 2016).

Recent calls for additional literature have seen a shift in discourses of student affairs assessment from a scholarship of assessment practice to scholarship on cultures of assessment. Whereas prior literature (Bingham & Bureau, 2015; Bresciani, Zelna, & Anderson, 2004; University of Pittsburgh, 2012) has outlined best practices in assessment of student learning and development, many practitioners and scholars (Bresciani et al., 2004; Douchy et al., 2007; Haviland, 2014; Baas et al., 2016) recognize the need to begin studying assessment as a unique facet of the student affairs profession. Calls for this enhanced scholarship on cultures of assessment include the need to examine how divisions of student affairs' cultures support or hinder the use of evidence in decision-making (Schuh, 2013). According to Schuh (2013), such examinations are the next frontier of scholarship in student affairs assessment.

Scholarship on student affairs cultures of assessment is limited, in part, due to a dearth of empirical evidence on cultures of assessment in student affairs. This lack of evidence and a synthetic theory of assessment culture has been noted in scholarship of assessment in academic settings (Long, 2012). To date, no literature calling for the empirical examination of cultures of assessment in student affairs has been published. However, many scholars (Bingham & Bureau, 2015; Bresciani et al., 2004;) have argued that the development of student

affairs cultures of assessment may be beneficial to practice and the advancement of student affairs as a profession. Therefore, the present study sought to fill this void by offering an initial examination of cultures of assessment through the perspective of mid-manager and higher-level staff in student affairs.

Method

Sample

The sample was drawn from volunteers willing to submit a listing of student affairs staff at the mid-manager and higher level of employment within their college or university for participation in this study. In the summer 2016 semester a nation-wide call for participation in the study was sent to 4,129 chief student affairs officers (CSAO). The Higher Education Directory, a nationwide directory of higher education leaders' contact information, was used to gather email addresses for CSAOs. These contacts were then invited to participate in the study by providing the lead researcher with the e-mail addresses for student affairs practitioners the CSAO deemed to be at the mid-manager level or higher. Most CSAOs were able to easily identify a list of mid-managers for inclusion in the study. Only e-mail addresses were submitted to the lead researcher using a contact file template. This allowed for the e-mail addresses to be entered into an online surveying system without an overt intrusion on individuals' privacy and identity.

The modified version of the Administrators Survey emerged in 2013 as part of an effort to focus on the unique contexts of student affairs assessment. Rather than focusing on institutional cultures of assessment as the Administrators and Faculty Surveys do, wording was augmented to focus on division-wide cultures of assessment.

Instrument

The *Student Affairs Survey* was used to measure student affairs administrator attitudes toward institutional assessment culture. Assessment culture is defined in the *Student Affairs Survey* as the overarching institutional ethos that is both an artifact of the way in which assessment is conducted and, simultaneously, a factor influencing and augmenting assessment practice (Fuller, 2011). The *Student Affairs Survey* parallels other *Surveys of Assessment Culture*, namely the *Administrators Survey* and the *Faculty Survey*. The *Administrators Survey* contains 48 items measured on a six-point Likert scale (1 = *Strongly Disagree*; 6 = *Strongly Agree*) and was first piloted in 2011 to a nationwide stratified, random sample of institutional research and assessment directors. An exploratory factor analysis (EFA) of the data from this sample suggested a five-factor model of the data: (a) Faculty Perceptions, (b) Use of Data, (c) Sharing, (d) Compliance or Fear Motivators, and (e) Normative Purposes for Assessment (Fuller, Skidmore, Bustamante, & Holzweiss, 2016). Reliability coefficients for each factor measured are reported to range between .792–.922 (Fuller & Skidmore, 2014; Fuller et al., 2016).

The modified version of the *Administrators Survey* emerged in 2013 as part of an effort to focus on the unique contexts of student affairs assessment. Rather than focusing on institutional cultures of assessment as the Administrators and Faculty Surveys do, wording was augmented to focus on division-wide cultures of assessment. This modified instrument was piloted in 2014 to an advisory panel of 12 experts drawn from student affairs units across the United States. Additional revisions were made, though most revisions could be categorized as slight wording revisions. The resulting instrument, the *Student Affairs Survey of Assessment Culture*, was administered in the present study to examine cultures of assessment in student affairs organizations.

Procedures

An anonymously recorded, electronic version of the *Student Affairs Survey* was sent to identified participants during the summer 2016 term. A total of 2,234 mid-manager or higher-level leaders were invited to participate from 59 institutions[1] across the United States.

[1]Study included 9 community college systems, which are accredited as a single institution at the system-level. These systems, if broken down into their sub-institutions, would increase the total number of institutions to 141 institutions. However, most of these institutions only volunteered 3 or 4 staff members to the study, making a system-wide comparison more appropriate.

Institutions volunteering to participate in the study were found in all six regional accreditation regions in the United States and from a variety of institutional sizes. The smallest participating institution reported only two mid-managers or above which constituted the entire professional staff at this institution. In contrast, several large, research-intensive universities opted to participate in the study, with the largest offering 309 staff as participants in the study.

Though limitations exist in the dispersion of institutions across the nation and institutional types, as well as the voluntary nature of participation in the study, the researchers were satisfied that a respectable number and mixture of institutions were represented in the study to warrant an exploration of this nature. Of the total 1,624 student affairs practitioners invited to participate in the study, 771 responded to the survey, offering a response rate of 47.5 percent.

Data Analysis

Although the *Student Affairs Survey of Assessment Culture* was designed with the intent of paralleling other Surveys of Assessment Culture, the specific survey items and wording of these items varied slightly across the surveys and there was no empirical evidence to support any common factor structure. As such, data from student affairs administrators were examined through an exploratory factor analysis (EFA). Exploratory factor analysis is a less restricted approach grounded in the same common factor model and allows for greater flexibility in the rotational strategies used for factor extraction (Flora & Flake, 2017).

Although there is some debate in the literature regarding the most appropriate method of extraction (Henson & Roberts, 2006), principal axis factoring was used as the extraction method given that the purpose of this study was to identify latent constructs. Factors were obliquely rotated using Promax criteria and a delta of zero given the relationship between factors reported in Fuller and Skidmore (2014). Because the Kaiser-Guttman rule (i.e., Eigenvalues > 1) and scree test can result in the over extraction of factors (Zwick & Velicer, 1986), parallel analysis (O'Conner, 2000) was also used in determining the number of factors to retain. Both the factor pattern matrix and structure matrix were considered in the interpretation of factors (Henson & Roberts, 2006).

Results

Nine factors were initially extracted using the Kaiser-Guttman rule (Eigenvalue >1) and these factors explained approximately 62% of the variance in the items. Examination of the scree plot suggested three or four possible factors within the data. Parallel analysis (O'Connor, 2000) indicated that five factors should be extracted using the 95th percentile of randomly generated eigenvalue means. Because prior research identified five factors among a sample of university administrators and faculty under a different version of this instrument, and because parallel analysis tends to be more accurate than both the EV >1 rule and the scree plot (Henson & Roberts, 2006; Keiffer, 1999; O'Connor, 2000; Zwick & Velicer, 1986), a second iteration of the analysis was performed specifying only five factors be extracted from the data.

Although the Student Affairs Survey of Assessment Culture was designed with the intent of paralleling other Surveys of Assessment Culture, the specific survey items and wording of these items varied slightly across the surveys and there was no empirical evidence to support any common factor structure.

The five-factor model explained 52% of the variance in the items but there were several concerns with this model. Three items (2B, 5R, 33) resulted in pattern, structure, and communality coefficients that were considered to be low based on guidance from the literature (Fabrigar, Wegener, MacCallum, & Strahan, 1999). Moderate levels of communality (e.g., .40–.70) are typically necessary to produce accurate estimates of the population parameters for sample sizes such as those used in this study (Fabrigar et al., 1999). When these items were removed from the analysis, one of the five factors contained only three items with factor pattern and structure coefficients ≥ .40. This was considered too few items to represent the factor (Fabrigar et al., 1999). It was decided that the remaining 45 items were best examined using a four-factor model.

The four-factor model with 45 items explained approximately 51% of variance in the items but several of these items (U3, 22, 29, 30R, 50R 50R) were identified as having both low pattern and structure coefficients (< .40). The variance explained by the factors improved to 55% when these items were removed. Internal consistency was then examined for each of the four factors using alpha coefficient. Alpha was not acceptable for two of the factors

($\alpha < .80$). Reliability analysis suggested that alpha could be improved for these two factors if items were removed from the data. Because one of these items had a low communality ($h^2 = .299$) and the other had the lowest pattern coefficient on factor 4 (-.313) both items were removed from the analysis.

The final model consisted of 38 items and four factors that explained 56% of the variance within items (Table 1). Factor 1 contained 15 items (3, 6, 8, 9R, 12, 13, 18, 21, 23, 25, 26R, 31, 36, 49, U2) and was labeled Clear Commitment to Assessment. Sample questions from this component included "Assessment is expected as a part of my division's continuous improvement process," or "Upper Student Affairs Administrators have made clear their expectations regarding assessment." Factor 2 consisted of 11 items (48, 49, 51R, 52, 53, 54, 55, 58, 66, 4H, U5) and was labeled Assessment Communication. Sample questions in this component included "Communication of assessment results has been effective," or "Assessment results are regularly shared throughout my division." Factor 3 contained 11 items (7R, 8, 13, 56R, 58, 60, 61, 66, 67, 3J, U2) and was labeled Connection to Change. Items indicative of this component included "Change occurs more readily when supported by assessment results," or "Assessment results are used for improvement." Factor 4 consisted of 8 items (4R, 7R, 10R, 11R, 57R, 62R, 64R, 65R) and was labeled Fear of Assessment. Items in this component included, but were not limited to, "Assessment results are used to scare student affairs staff into compliance with what the administration wants," or "Administrators use assessment to punish student affairs staff members." The Pearson r correlation coefficients between factors are provided in Table 2.

The model developed in the present study may offer new insights into cultures of assessment among student affairs professionals. Previously, this scholarship has remained largely conjectural; scholars and practitioners hypothesize about the importance or nature of an organizational culture of assessment yet they operated from a dearth of empirical evidence on the topic.

A total of 14 of the 52 items were removed from the analyses due to having communalities less than 0.40, through comparison of factor pattern weights for each factor, or to improve factor reliability. Table 3 provides a listing of items removed from the analyses and the reasoning behind their removal. Though none of the items removed represent a significant number of items so as to constitute additional factors the researchers did engage in iterative rounds of analyses to reduce the model to its current parsimonious form.

As such, future analyses with similar or different populations may reveal different results and these items could be suggestive of directions for future research or interpretation of results. In particular, similarly worded items which were removed could be suggestive of additional, latent constructs for future consideration or higher-order factors. For example, three of the items (Q5, Q33R, Q4R) logically relate to the purpose of assessment. Such a construct has been noted in studies focusing on faculty and administrative populations (Fuller et al., 2016; Fuller & Skidmore, 2014). Conceivably, factor 1 [Clear Commitment to Assessment] offers similar concepts as a purpose of assessment factor in that one should have a clear understanding of the purpose of assessment in order to be committed to it. Similarly, Questions Q53, Q50, Q54, and QS3L could conceivably relate to a factor pertaining to the use of assessment data. Fuller et al. (2016) and Fuller (2016) noted the importance of the use of data in creating and sustaining an institutional culture of assessment. These removed items could relate to the third factor in the current study, Connection to Change, in that the use of data could be the vehicle through which data are used for change purposes. Finally, a number of removed items (Q2B, Q22, Q33R, QU3, Q23) relate to clarifying who is responsible for assessment within the student affairs division. Fuller et al. (2016) argued that officially delegating the responsibility for assessment to a specific person, office, or collection of offices is an important leadership tactic for supporting a culture of assessment. These items, though removed, may relate to other factors pertaining to responsibility for assessment or support structures for assessment. While it is important to note that these items do appear to offer some logical similarities these items were removed from the present analyses through analytical iterations and with sound justification for doing so. Their inclusion in future studies may be beneficial to the scholarship on culture of assessment and could generate unique results.

Table 1

Item Means, Standard Deviations, Factor Pattern Coefficients (P), Structure Coefficients (r_s), and Communalitites (h^2)

			Factor 1		Factor 2		Factor 3		Factor 4	
Item	*M*	*SD*	P	r_s	P	r_s	P	r_s	P	r_s
Q18	4.48	1.51	**.84**	.85	-.02	.65	-.20	.55	.03	-.35
Q21	4.28	1.36	**.83**	.80	.11	.62	-.18	.56	.05	-.30
Q36	3.97	1.41	**.77**	.78	.12	.57	.00	.40	.02	-.28
Q25	4.63	1.19	**.71**	.75	.09	.58	.09	.56	.07	-.34
Q3	5.20	0.93	**.67**	.73	-.17	.54	.17	.58	-.01	-.28
Q31	4.12	1.28	**.63**	.70	-.01	.67	.22	.51	.07	-.28
Q6	4.20	1.26	**.63**	.68	.07	.43	.12	.30	.00	-.25
Q26R	3.20	1.53	**-.58**	.66	-.18	.40	.13	.49	.08	-.30
Q12	4.37	1.26	**.49**	-.66	.01	-.53	.28	-.38	.02	.34
Q8	4.43	1.11	**.45**	.66	-.16	.51	**.40**	.58	.08	-.30
Q49	3.92	1.37	**.45**	-.63	**.36**	-.50	.03	-.52	.04	.47
Q23	4.05	1.64	**.41**	.56	.27	.38	-.18	.56	-.04	-.21
Q9R	3.16	1.36	**-.40**	.50	-.07	.46	-.14	.26	.23	-.23
Q53	3.22	1.32	.03	.64	**.82**	.83	-.06	.54	.03	-.33
Q51R	3.44	1.40	.04	.64	**-.75**	.82	.02	.55	.14	-.31
Q4H	3.41	1.31	.09	.56	**.74**	.80	.01	.46	-.05	-.22
Q52	3.76	1.39	.10	.57	**.71**	.76	-.03	.46	.03	-.23
Q48	3.64	1.46	.12	-.54	**.71**	-.76	.03	-.47	-.02	.36
QU5	3.52	1.46	-.03	.57	**.59**	.74	.22	.60	-.11	-.37
Q55	3.38	1.27	.17	.51	**.56**	.64	-.05	.39	.03	-.20
Q66	4.14	1.23	-.14	.41	**.49**	.58	**.36**	.58	.19	-.03
Q54	4.27	1.14	.08	.34	**.39**	.55	.09	.49	-.12	-.04
Q67	4.57	1.14	-.27	.47	.15	.54	**.74**	.43	-.13	-.32
Q61	4.53	1.09	.13	.61	.00	.54	**.69**	.81	-.10	-.42
Q3J	4.83	1.13	-.08	.60	-.09	.55	**.68**	.79	-.14	-.37
Q56R	4.33	1.16	.13	.62	.05	.58	**.66**	.72	-.05	-.19
Q60	4.05	1.27	.28	.36	.11	.46	**.54**	.72	.17	-.34
Q58	3.88	1.26	-.08	.70	**.44**	.61	**.45**	.71	.24	-.32
QU2	4.10	1.29	**.37**	.65	.10	.50	**.43**	.65	.04	-.42
Q13	4.49	1.15	**.38**	.35	-.04	.32	**.39**	.63	-.12	-.33
Q57R	2.30	1.16	-.07	-.33	.00	-.22	.08	-.23	**.72**	.72
Q62R	1.84	0.97	-.08	-.28	.13	-.14	-.02	-.23	**.64**	.64
Q10R	2.94	1.49	.25	-.40	-.12	-.27	-.03	-.25	**.57**	.61
Q65R	2.73	1.39	-.23	-.56	.01	-.43	.10	-.49	**.55**	.59
Q7R	2.41	1.22	.26	-.42	.06	-.38	**-.36**	-.32	**.48**	.56
Q64R	3.23	1.42	-.13	-.10	-.16	-.14	.04	-.16	**.46**	.51
Q4R	3.18	1.38	.07	-.14	.00	-.14	-.14	-.35	**.46**	.49
Q11R	2.95	1.43	-.29	-.21	-.01	-.18	-.15	-.27	**.40**	.48
Initial Eigenvalues			14.78		2.73		1.94		1.68	
Trace			12.26		11.13		10.00		5.47	
% Variance Explained			32.26		29.29		26.32		14.39	

[a]The total variance explained reflects the initial eigenvalues. Trace values cannot be added to obtain total variance explained after rotation because factors were correlated.

Note. Factor pattern coefficients greater than |.30| are bolded, underlined, and were retained for that component. Percentage variance is post-rotation; percentage of variance is trace divided by 38 (# of items) times 100. The eigenvalue of the fifth, non-retained factor was 1.17. h^2 = communality coefficient.

Table 2

Factor Correlation Matrix

Factor	*M*	*SD*	1	2	3	4
1	4.16	0.70	--			
2	3.69	0.80	.70	--		
3	4.16	0.74	.63	.61	--	
4	2.68	0.84	-.43	-.32	-.38	--
α	.84	--	.82	.83	.85	.79

Table 3

Removed Items

Item	Question	Reason for Removal
Q5R	The purpose of assessment depends largely on who is asking for assessment results.	Factor pattern weight less than 0.40. One of 2 lowest loading items on Factor 4
Q33R	Assessment for accreditation purposes is prioritized above other assessment efforts.	Lower communality then Q5R. One of 2 lowest loading items on Factor 4
Q4R	Assessment is an exercise primarily for compliance purposes.	Factor pattern weight less than 0.40.
Q2B	Faculty are in charge of assessment at my institution.	Factor pattern weight less than 0.40.
Q22	I can name the office at my institution that leads assessment efforts for accreditation purposes.	Factor pattern weight less than 0.40.
Q30R	Assessment is primarily the responsibility of student affairs staff.	Factor pattern weight less than 0.40.
Q53L	Student affairs staff consistently receive assessment data from administrators.	Factor pattern weight less than 0.40.
Q50R	Assessment results are NOT intended for distribution.	Factor pattern weight less than 0.40.
Q54	Assessment results are available from administrators by request.	Factor pattern weight less than 0.40.
Q29	Assessment is primarily the responsibility of faculty members.	Factor pattern weight less than 0.40.
QU3	Assessment is primarily the responsibility of upper student affairs administrators.	Factor pattern weight less than 0.40.
Q23	I can name the office at my institution that leads assessment efforts for student learning.	Factor pattern weight less than 0.40.
QU4	Upper student affairs administrators are supportive of making changes.	Factor pattern weight less than 0.40.
QS3L	Assessment results have no impact on resource allocations.	Improve α for Factor 4

Discussion

The model developed in the present study may offer new insights into cultures of assessment among student affairs professionals. Previously, this scholarship has remained largely conjectural; scholars and practitioners hypothesize about the importance or nature of an organizational culture of assessment yet they operated from a dearth of empirical evidence on the topic. The aforementioned model is suggestive of factors of a division-wide culture of assessment in student affairs. These factors offer opportunities to consider cultures of assessment in the student affairs context anew. For example, the factors pertaining to the clarity of assessment's purpose and communication about assessment offer opportunities for student affairs leaders to reflect upon the regularity and clarity with which they talk about assessment with student affairs staff. Offering clear comments on assessment's purpose, providing regular "success stories" as exemplars, or sharing assessment results with staff in a public manner are just a few practices that advance or sustain an organizational culture of assessment in student affairs. Participants were asked to respond to an open-ended, qualitative question in the *Student Affairs Survey* that asked how they prefer to receive communication about assessment results. Though further analyses of these data are needed an overwhelming majority of respondents indicated they preferred to receive e-mail notifications about assessment results from the CSAO or the CSAO's assessment designee.

The results from this study suggest that student affairs staff may approach assessment with far greater nuance than administrators and faculty—yet also with some notable similarities between the groups.

The results from this study suggest that student affairs staff may approach assessment with far greater nuance than administrators and faculty—yet also with some notable similarities between the groups. For example, in studies of the factor structure inherent in the *Administrators Survey* (Fuller & Skidmore, 2014), the factors listed included a) Clear Commitment, b) Connection to Change, and c) Vital to Institution. Factors such as Clear Commitment to Assessment and Connection to Change closely align to corresponding factors

found in Fuller and Skidmore's (2014) study of administrators' perspectives. However, items related to a sense of vitality to their institution's future did not coalesce into a similar factor in the present study. Similarly, in examining data collected from the *Faculty Survey* (Fuller et al., 2016), a) Faculty Perceptions, b) Use of Data, c) Sharing, d) Compliance or Fear Motivators, and e) Normative Purposes for Assessment were found to form the underlying factors of faculty perceptions of institutional cultures of assessment. Here, a notable similarity between student affairs and faculty populations includes a factor related to fear of assessment. Indeed, student affairs staff may approach division assessment efforts with some trepidation or skepticism. In the present study participants were asked to agree or disagree with the statement "The majority of student affairs staff in my division are afraid of assessment (69R)." Nearly half—49.7%—of respondents indicated that to some extent they agreed with this statement. For many student affairs leaders assessment has remained a fearful endeavor—a regulatory mechanism that significantly reduces their time on core functions or a punishment received at the whim of an institutional leader.

Though student affairs staff approach assessment differently from other administrators and faculty on campus, fear may be a tremendous unifying force in a struggle against assessment—in solidarity with their faculty colleagues. Chief student affairs officers and assessment leaders may find it useful to redefine discourses of fear of assessment by engaging key student affairs staff in dialogue about their assessment fears, hopes for their units, and fundamental perspectives on student learning. Assessment, as a process aimed at transformation, is often fearful for many higher-education leaders (Bresciani et al., 2010). Student affairs assessment leaders can do much to support their colleagues through such transformations. Useful leadership tactics in this support phase include, among others, listening to staff members' needs and concerns, managing or staggering tasks due to avoid a sense of overwhelm, and initiating discussions about assessment that are contextualized by staff members' fundamental perspectives on student learning (i.e., not using assessment to tell staff their fundamental perspectives on student learning are flawed but instead using it as a means to talk about student learning in general; Fuller & Skidmore, 2014). Student affairs assessment leaders may find it useful to heighten or reconceptualize their division's fundamental discourses about student learning (Henning & Roberts, 2016). Assessment is often viewed by student affairs staff as a construct that supports accountability or other externally motivated discourses (Henning & Roberts, 2016; Suskie, 2014). Instead, it could serve as an evidence-based means of reflecting upon the nature and purpose of student learning and development. The present study offers student affairs leaders opportunities to reflect upon their practice and develop new ways of talking about and engaging in assessment in their division such that assessment is a framework focused on supporting or advancing student learning and development.

Though student affairs staff approach assessment differently from other administrators and faculty on campus, fear may be a tremendous unifying force in a struggle against assessment—in solidarity with their faculty colleagues.

The four-factor solution emerging in the present study, along with the refinement away from a nine-factor solution associated with other versions of the survey, suggest the need for continued refinement and revision of the *Student Affairs Survey of Assessment Culture*. After consideration by a panel of student affairs experts, the Council of Scholars, the current instrument was developed by making modifications to the *Administrators Survey of Assessment Culture*. However, differences in the conceptualization of assessment culture between student affairs staff, administrators, and faculty are best answered through other statistical approaches (e.g., chi-square test of fit, RMSEA, etc.). Empirically testing those differences was beyond the scope of this study but would be required to understand the specific ways in which assessment culture varies between these groups. Additional studies will focus on the comparison of conceptualizations of cultures of assessment across administrative, faculty, and student affairs groups. The present analyses, however, offer a foundation for such future studies by providing the psychometric properties of the *Student Affairs Survey of Assessment Culture*. Moreover, the instrument appears to offer a sound, refined approach to empirically examining cultures of assessment in student affairs contexts. As such, additional revisions to the instrument are not expected in the immediate future and ongoing studies will be conducted to further explore this complex topic.

Lastly, though student affairs assessment literature was reviewed in the development of this instrument it was not the only scholarship reviewed, offering opportunities to focus on student affairs contexts in future revisions. Moreover, additional scholarship (Bingham & Bureau, 2015; Henning & Roberts, 2016; Schuh et al., 2016)—including many works focusing on student affairs assessment—have been authored since the *Administrators Survey* was crafted and even some have been published following the summer 2016 administration of the *Student Affairs Survey of Assessment Culture*. These additional works may highlight nuanced approaches to student affairs assessment worth exploring through future studies.

Conclusion

Student affairs leaders have been asked to operate and lead their units with a dearth of empirical evidence about cultures of assessment in student affairs. This gap in the literature is made all the more problematic by the fact that assessment, as a function of modern academe, is a process aimed at the inclusion and use of evidence in decision-making processes. The present study calls upon data from the Student Affairs Survey to examine fundamental concepts undergirding how student affairs practitioners conceptualized their division's culture of assessment. The model offered through the exploratory factor analysis provides an initial conceptualization of assessment cultures in student affairs contexts. Further theorization and analyses will reveal new considerations and augment practice through evidence-based scholarship.

References

Baas, L., Rhoads, J. C., & Thomas, D. (2016). Are quests for a "culture of assessment" mired in a "culture war" over assessment? A Q-methodological inquiry. Sage Open, 1–17. doi:10.1177/2158244015623591

Banta, T. (2002). *Building a scholarship of assessment* . San Francisco: Jossey-Bass.

Barr, M. J. (1993). *The handbook of student affairs administration*. The Jossey-Bass higher and adult education series. Jossey-Bass Publishers, San Francisco, CA.

Bingham, R. P., & Bureau, D. (2015). Tenet one: Understand the "why" of assessment. In R. P. Bingham, D. Bureau, & A. G. Duncan (Eds.), *Leading assessment for student success: Ten tenets that change culture and practice in student affairs* (pp. 9–21). Sterling, Virginia: Stylus Publishing, LLC.

Bresciani, M. J., Zelna, C. L., & Anderson, J. A. (2004). *Techniques for assessing student learning and development: A handbook for practitioners*. Washington, D.C.: National Association of Student Personel Administrators.

Bresciani, M. J., Gardner, M. M., Hickmott, J. (2010). *Demonstrating student success: A practical guide to outcomes-based assessment of learning and development in student affairs*. Sterling, VA: Stylus Publishing.

Douchy, F., Segers, M., Gijbels, D., & Struyven, K. (2007). Assessment engineering: breaking down barriers between teaching and learning, and assessment. In D. Boud, & N. Falchicov (Eds.), *Rethinking assessment in higher education: Learning for the longer term* (pp. 87–100). New York, NY: Routledge.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*(3), 272.

Flora, D. B., & Flake, J. K. (2017). The purpose and practice of exploratory and confirmatory factor analysis in psychological research: Decisions for scale development and validation. *Canadian Journal of Behavioral Science, 49*(2), 78–88.

Fuller, M. B. (2011). *Conceptual framework for the Survey of Assessment Culture*. Montgomery, TX: Fuller Educational Consulting.

Fuller, M. B. (2016). *Nation-wide results from the Faculty Survey of Assessment Culture*. Huntsville: Sam Houston State University. Retrieved from www.shsu.edu/assessmentculture

Fuller, M. B., & Skidmore, S. (2014). An exploration of factors influencing institutional cultures of assessment. *International Journal of Educational Research, 65*(1), 9–21.

Fuller, M., Skidmore, S., Bustamante, R., & Holzweiss, P. (2016). Empirically exploring higher education cultures of assessment. *The Review of Higher Education, 39*(3), 395–429. doi:10.1353/rhe.2016.0022

Haviland, D. (2014). Beyond compliance: How organizational theory can help leaders unleash the potential of assessment. *Community College Journal of Research and Practice, 38*(9), 755–765. doi:0.1080/10668926.2012.711144

Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research. Educational and *Psychological Measurement, 66*, 393–416.

Henning, G., & Roberts, D. (2016). *Student affairs assessment: Theory to practice*. Sterling, VA: Stylus Publishing.

Kieffer, K. M. (1999). An introductory primer on the appropriate use of exploratory and confirmatory factor analysis. *Research in the Schools, 6*(2), 75–92.

Kirschner, J. (2016, Jan. 25). *Proving our worth in student affairs: Identifying and addressing critical needs in higher education*. Retrieved from NASPA Blog: https://www.naspa.org/about/blog/proving-our-worth-in-student-affairs

Kuh, G. D., & Banta, T. W. (2000). Faculty-Student Affairs Collaboration on Assessment-Lessons from the Field. About Campus, 4(6), 4–11.

Kuh, G. D., Ikenberry, S. O., Jankowski , N. A., Cain, T. R., Ewell, P. T., Hutchings, P. T., & Kinzie, J. (2015). *Using evidence of student learning to improve higher education*. San Francisco, CA: Jossey-Bass.

Long, D. (2012). *The foundations of student affairs assessment: A guide to the profession*. In L. J. Hinchliffe, & M. A. Wong (Eds.), Environments for student groth and development: Librarians and student affairs in collaboration (pp. 1–39). Chicago, IL: Assocation of College and Research Libraries.

O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instrumentation, and Computers, 32*, 396–402.

Roper, L. D. (2015). Student affairs assessment: Observations of the journey, hope for the future. *Journal of Student Affairs Inquiry*, 1–15.

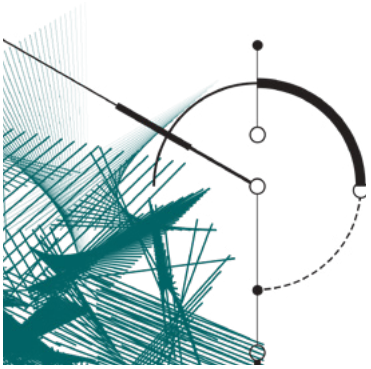
Schuh, J. H. (2013). *Developing a culture of assessment in student affairs*. In J. H. Schuh, New Directions for Student Services (Vol. 142, pp. 89–98). San Francisco, CA.

Schuh, J. H., Biddix, J. P., Dean, L. A., & Kinzie, J. (2016). *Assessment in Student Affairs*. San Francisco: Jossey Bass Pub. Co.

Suskie, L. C. (2014). *Five dimensions of quality: A common sense guide to accreditation and accountability*. San Francisco, CA: Jossey-Bass.

University of Pittsburgh. (2012). *Developing a culture of assessment: Student affairs model*. Retrieved from Culture of assessment: http://www.middlestates.pitt.edu/node/625

Zwick, W. R., & Velicer, W. F. (1986). Factors influencing five rules for determining the number of components to retain. *Psychological Bulletin, 99*, 432–442



Abstract

Effective assessment practice requires clearly defining and operationalizing terminology. We illustrate the importance of this practice by focusing on academic "undermatching"—when students enroll in colleges that are less academically selective than those for which they are academically prepared. Undermatching has been viewed as a potential obstacle in the United States' goal of increasing degree attainment but operationalizing undermatching is difficult. Using ELS: 2002, a national dataset from the U.S. Department of Education National Center for Education Statistics (NCES, 2014), we developed eight operationalizations of undermatching by altering three commonly used variables. We then compared the number and demographics of students who were identified as undermatched. Differences in operationalizations resulted in significant differences in undermatching by gender, race, parental education, and socioeconomic status. Results of this study illustrate the importance of the need to operationalize terminology used in assessment carefully and consistently.

AUTHORS

Ann M. Gansemer-Topf
Iowa State University

Jillian Downey
Truman State University

Ulrike Genschel
Iowa State University

Definitions Matter: Investigating and Comparing Different Operalionalizations of Academic Undermatching

Effective assessment practice requires clearly defining and operationalizing terminology, but assessment professionals often need to create their own definitions of student populations. For example, research investigating science, technology, engineering, and math (STEM) fields varies in who is included as a STEM major, with some studies including social science majors such as psychology; others limit the definition to hard sciences such as biology, chemistry, or engineering. First-generation students may be defined as those who have no college experience, those who have at least one parent without a college degree, or those who have no parents with a college degree. When these varying definitions are the subject of research studies the results may vary. Toutkoushian and Stollberg (2015) found that varying the definition of first generation altered the number of students who were identified as such—subsequently affecting policies and practices aimed at improving student success for this population of students.

Therefore, research that investigates how the operationalizations of variables may influence assessment results and implications is critical. This study focuses on a specific population—academically undermatched students—to highlight an often overlooked but essential assessment practice: clearly defining the terminology and methods. Academic "undermatching"—when students enroll in colleges that are less academically selective than those for which they are academically prepared—has been viewed as an impediment to degree attainment (Bowen, Chingos, & McPherson, 2009; Executive Office of the President, 2014). Researchers have operationalized undermatching in a variety of ways using a variety of datasets (e.g., Belasco & Trivette, 2015; Bowen et al., 2009; Heil, Reisel, & Attewell, 2014; Rodriquez, 2013; Smith, Pender, & Howell, 2013). Results of these studies have varied: Roderick, Coca, and Nagaoka (2011) found that approximately 62% of college-going students were likely to undermatch; Bowen et al. (2009) and Smith et al. (2013) concluded that 40% were likely to undermatch; and Belasco and Trivette calculated that about 28% were likely to undermatch. Each study was based on a different population of students.

CORRESPONDENCE

Email
anngt@iastate.edu

Rodriquez (2015) utilized one population of students and compared three approaches (acceptance rate, enrollment rate, and predicted rate) to undermatch and found that the percentages and characteristics of students defined as undermatched varied among the three approaches. Our study narrowed this variability further. We wanted to investigate if there were differences in undermatch when we used the same population and same approach but altered the variables within this approach. Our intent was to examine how small differences in operationalizations may change who is identified as undermatched. Given the importance for assessment professionals to clearly define their populations, our goal was to undertake research on a topic of national importance (i.e., academic undermatching) as a way to illustrate how variations in operationalizations might affect our assessment results and implications of these results. Using one national dataset, and operationalizing undermatching in eight different ways by altering similar variables, we sought to answer two research questions:

1. How consistent were different operationalizations in their ability to define students as undermatched?
2. In comparing different operationalizations of undermatching, were there differences in the demographic characteristics of students (gender, race/ethnicity, parental education, and income) for those classified as undermatched?

Assessment and Undermatch Research

Assessment is a valuable process that can guide institutional change, improvement, and strategic planning (Bresciani, Gardner, & Hickmott, 2012; Middaugh, 2011; Schuh, Biddix, Dean, & Kinzie, 2016) but this process requires developing goals that are clear, measurable, and meaningful (Banta, Jones, & Black, 2009; Bresciani et al., 2012; Suskie, 2010). Although developing clear and measurable goals is a consistent theme throughout the literature, less emphasis is placed on the importance of clarifying the terms and definitions within these outcomes or identifying the population that is being assessed. This lack of clarification in defining data and populations can lead to inconsistent data collection processes, measures, and interpretations (McLaughlin & Howard, 2004) that can undermine institutional efforts to effectively improve, change, or plan. Our study wanted to illustrate this point by focusing on academic undermatching.

In *Crossing the Finish Line*, Bowen et al.'s (2009) national study highlighted the negative relationship between undermatching and degree attainment. Holding all academic and demographic variables constant, students attending higher selective institutions were more likely to graduate than students at less selective institutions. Undermatched students are students who attend less selective schools; therefore, they are less likely to graduate.

Given the importance for assessment professionals to clearly define their populations, our goal was to undertake research on a topic of national importance (i.e., academic undermatching) as a way to illustrate how variations in operationalizations might affect our assessment results and implications of these results.

With the national push to raise completion rates for all students (ACE, 2013) and undermatching being viewed as an obstacle for degree completion (Executive Office of the President, 2014), a significant amount of attention has been focused on minimizing undermatch (Bastedo & Jaquette, 2011; Bowen et al., 2009; Hoxby & Avery, 2012; Roderick et al., 2008). Research on this topic has investigated if certain subpopulations are more likely to undermatch than others; results have been mixed. Rodriquez (2013) found that Latino students were more likely to undermatch than their White peers, and that low-income, first-generation students were also more likely to undermatch than students from middle- or high-income families with parents who had more than a high-school education. Bowen et al. (2009) found that African-American students were more likely to undermatch; Belasco and Trivette (2015) found that Latino and African-American students were less likely to undermatch. Belasco and Trivette also noted that females were more likely to undermatch than males, contradicting Smith et al.'s (2013) findings. This study attempted to narrow the variability among past studies by comparing how seemingly minor changes to the operationalizations alter who is classified as undermatched. We examined academic undermatching because, despite the significant amount of national attention focused on implications of undermatch for degree completion, the term undermatch remains difficult to define. Therefore, we determined that this topic would be an excellent example of how definitions matter. By engaging in this process, we hoped to reiterate the need for assessment professionals to engage in definitional rigor and clarity.

Methods

This quantitative study utilized the U.S. Department of Education's National Center for Education Statistics Education Longitudinal Study of 2002 (ELS: 2002; NCES, 2014) and the Barron's Selectivity Ratings (Barron's Educational Series, 2009). The ELS: 2002 captures students' demographics, high-school academic data, financial aid and college choice information (schools applied to and accepted at), the higher-education institution where the students enrolled, and the NCES selectivity classification of that institution. This dataset has been used in past studies of undermatching; (i.e., Belasco & Trivette, 2015; Rodriquez, 2013; Smith et al., 2013). Applying contrasting operationalizations to a dataset that had been used for undermatching provided us the opportunity to view if these differences changed who was defined as undermatched. We captured Barron's selectivity rating by merging that dataset with ELS.

Sample

The ELS data contained a nationally representative sample (N=11,840[1]). We used the panel sampling weights provided by NCES. Students with missing data were deleted, resulting in a sample of 8,020 students. Subsequent analysis demonstrated that this sample was not significantly different from the larger sample, and the sample size was sufficient enough to complete our analysis.

Undermatch Operationalization

In developing our operationalizations to answer Research Question 1, we reviewed previous literature that had statistically defined undermatching. After examining these multiple approaches, we chose to utilize the "eligibility frontiers" with the variables used by Bowen et al. (2009) and Belasco and Trivette (2015) to determine a student's access level. Both studies classified a student as undermatched if the selectivity level of school the student attended was less than the selectivity level of school for which the student had access. Bowen et al. (2009) and Belasco and Trivette (2015) created eligibility frontiers that utilized categorized information on high-school GPA and standardized test scores (i.e., SAT or ACT). To create an eligibility frontier, we considered only those students who applied to schools in the highest selectivity level. For each GPA and SAT score combination, a proportion was calculated indicating how likely it was for students that fall into that particular combination to be accepted to schools at the highest selectivity level. If this proportion was larger than a preselected threshold value, all students that fell into that particular category were deemed to have access to the highest-selectivity-level school. If the proportion was less than the threshold, the procedure was repeated for the next highest level of selectivity and continued until the highest level of access was determined for each GPA and SAT score combination. We chose to use the eligibility frontier approach to determine access because it allowed us to easily examine how changing minor pieces of the definition may influence undermatching. Additionally, this approach utilized GPA and SAT scores, two widely reported student characteristics.

Including more student-level variables may more accurately predict undermatch, but for our purpose we wanted to use a more parsimonious definition in order to examine how small changes in these few variables may change whether a student is defined as undermatched.

Other definitions of undermatching have used other high-school variables such as Advanced Placement credits, number of high-school credits (Rodriquez, 2013; Smith et al., 2013), or high school location (Hoxby & Avery, 2012). Including more student-level variables may more accurately predict undermatch, but for our purpose we wanted to use a more parsimonious definition in order to examine how small changes in these few variables may change whether a student is defined as undermatched.

Data Analysis

We converted all standardized test scores (i.e., ACT or SAT) into SAT scores. We then modified the following three factors (school selectivity classification, GPA and SAT categorization, and calculation of access probability) to examine if students were consistently identified as undermatched.

[1]rounded to the nearest 10s by publication requirement of IES

Selectivity classification. This study included two measures of institutional selectivity: Barron's classification (Barron's Educational Series, 2009) and the selectivity variable found in NCES datasets. Barron's selectivity levels range from 0–6 and are based on high-school GPA, high-school rank, ACT/SAT scores, and acceptance rates. The NCES variable is used in national datasets. Ratings are 0–5 and based on admission policy (i.e., open or not), the number of applicants, number of students admitted, and the 25th and 75th percentiles of ACT/SAT scores.

High-school GPA categorization. One GPA categorization began at 2.0 with 0.3 point increases and was chosen because this was the categorization used by Belasco and Trivette (2015). The second GPA categorization began at 1.0 with 0.5 point increases and was chosen because these were the cutoffs used in other NCES datasets. We categorized SAT scores similar to Belasco and Trivette and Bowen et al. (2009) but did not want to significantly increase the number of operationalizations to compare; thus, we did not vary the SAT categorization cutoffs.

To answer Research Question 2, we limited our sample to only those classified as undermatched and then examined if gender, race/ethnicity, parental education, and socioeconomic status were affected similarly across operationalizations for those defined as undermatched.

Calculation of access probability. We calculated the access probabilities in two ways. The first calculation used all applications. For example, suppose we are considering the highest level of school selectivity. For a given GPA and SAT combination, the access probability was calculated by dividing the total number of acceptances by the total number of applications for all students in the GPA and SAT combination of interest. The second calculation of access probability aggregated over students (Belasco & Trivette, 2015): for students in a given GPA and SAT combination, the access probability was calculated by taking the total number of students that were accepted to at least one highest-selectivity-level school divided by the total number of students that applied to at least one highest-selectivity-level school. It is important to note that the first calculation of access probability used all applications but did not take into account the dependence of multiple observations from one student. In contrast, the second calculation of access probability aggregated all applications and acceptances over a student, thus not taking into account the total number of applications and acceptances for each student.

Regardless of the selectivity-level classification, high-school GPA categorization, or method used to calculate access probability, if the access probability for a given GPA and SAT combination and selectivity level was greater than or equal to 90% based on 10 or more observations, a student in that GPA and SAT combination was deemed to have access to that particular school selectivity level. If there were fewer than 10 observations, no conclusions were reached for the particular school selectivity level.

We obtained eight different operationalizations (O1, O2…O8) as a result of two levels for each of the three factors (see Table 1). For all eight operationalizations, an eligibility frontier was created that we used to categorize the level of school a student had access to.

Table 1

Description of Eight Operationalizations of Undermatching

Operationalization	Classification of School Selectivity	GPA Categorization	Access Probability Calculation
1	NCES	Start at 2.0, increase by 0.3	All Applications
2	NCES	Start at 2.0, increase by 0.3	Student Aggregate
3	Barron's	Start at 2.0, increase by 0.3	All Applications
4	Barron's	Start at 2.0, increase by 0.3	Student Aggregate
5	NCES	Start at 1.0, increase by 0.5	All Applications
6	NCES	Start at 1.0, increase by 0.5	Student Aggregate
7	Barron's	Start at 1.0, increase by 0.5	All Applications
8	Barron's	Start at 1.0, increase by 0.5	Student Aggregate

We then compared this level of access to the level of school the student first attended. If the level of school the student attended was less than the level of school to which they had access the student was classified as undermatched. To answer Research Question 1, we classified students as undermatched for all eight operationalizations and identified how often these different operationalizations agreed for each student.

Next, we examined gender, race/ethnicity, parental education, and socioeconomic status across all eight operationalizations by calculating the percentage identified as undermatch for all categories in each demographic variable. For example, if there were 4,230 females in the sample population, we examined what percentage of females were classified as undermatched using each operationalization. We then conducted a Pearson's chi-square test of independence (Agresti, 2012) to determine if there exists an association between each demographic variable and undermatching. A Pearson's chi-square test is used to establish if the outcomes of one variable are related to the outcomes of a second variable. For example, we conducted a Pearson's chi-square test for gender and operationalization 1, which told us if gender and being undermatched using operationalization 1 were associated. Comparing the outcomes of the chi-square tests across all eight operationalizations allowed us to determine if being undermatched was related to gender for all definitions or just a select few. To account for the multiple comparisons, we implemented a Bonferroni adjustment (Oehlert, 2000) at the individual variable level resulting in a level of significance of $\alpha/n = 0.05/8 = 0.00625$.

All operationalizations showed a statistically significant relationship between being undermatched and race/ethnicity as well as socio-economic status, meaning there was inconsistency across operationalizations.

To answer Research Question 2, we limited our sample to only those classified as undermatched and then examined if gender, race/ethnicity, parental education, and socioeconomic status were affected similarly across operationalizations for those defined as undermatched. We again calculated sample proportions to investigate if the demographic characteristics were similar across operationalizations.

Results

Tables 2 and 3 show comparison of the eligibility frontiers based on O1–O4 and O5–O8. For readability purposes, we chose only to illustrate four operationalizations per table. One cell represents a given GPA and SAT categorization. Each cell is split into four quadrants. The numbers in each quadrant represent the operationalization (i.e., O1 is in the upper left quadrant in Table 2). The colors of each quadrant represent the school selectivity level a student had access to with darker colors corresponding to higher selectivity level schools. For example, students with a GPA between 2.3 and 2.6 and an SAT score between 1200 and 1290 using O1 had access to at most a Level 2 selectivity school. Using O2, students had access to a Level 3 school; O3 students had access to a Level 5 school and O4 students had access to a Level 4 school. One might assume that as SAT and GPA increase, the level of access also should increase. However, this lack of monotonicity (Johnson & Wichern, 2007) was present in all eight eligibility frontiers under consideration. For O1–O4, aside from Level 1 selectivity schools (indicated by blank cells), there is only one GPA and SAT combination (GPA from 3.2 to 3.5 and SAT between 1100 and 1190) for which all four operationalizations resulted in the same level of selectivity access. In comparing O5–O8, other than Level 1 selectivity schools, there are no GPA and SAT combinations that resulted in the same level of access (see Table 3).

We then examined how consistently students were identified as undermatched for the eight definitions (see Table 4). Of the 8,020 students, the proportions of classified students varied by definitions between 5.1% classified by one out of eight definitions, 8.6% by two out of eight, and 8.7% classified by all eight definitions. In the sample, 4,360 (54.3%) were classified as not undermatched by all eight definitions; 3,660 (45.7%) were classified as undermatched by at least one definition; 1,650 (20.6%) students were classified as undermatched by at least five definitions. Of the students classified as undermatched by at least one definition (n=3,660), 700 (19.1%) were consistently classified as undermatched using all eight definitions.

Likewise, we examined the sample proportions of undermatched students within categories of each demographic variable (e.g., male, females; see Table 5). Using O1, 15.2% of females would be considered undermatched compared to 39.4% if using O4. A similar pattern was found for men, with 15.1% of males classified as undermatched using O1 and 37.8% using O4. Operationalization 4—which used Barron's Selectivity Rating, calculated

Table 2

Comparison of Eligibility Frontiers for Operationalizations 1 through 4

SAT

GPA	<800	800-890	900-990	1000-1090	1100-1190	1200-1290	1300-1390	≥1400
<2.0	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3 4
(2.0,2.3)	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3 4
(2.3,2.6)	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3 4
(2.6,2.9)	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3 4
(2.9,3.2)	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3 4
(3.2,3.5)	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3 4
(3.5,3.8)	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3 4
≥3.8	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3 4

Note: Number indicates operationalization. Colors indicate the level of selectivity to which a student has access. Darker colors indicate that the student has access to more highly selective school.

KEY

	Access to level 1
	Access to level 2
	Access to level 3
	Access to level 4
	Access to level 5

access probabilities by aggregating over students, and started with GPA at 2.0 and increased by .3—resulted in the highest sample proportion of students being classified as undermatched in all categories. Operationalization 5 had the lowest proportion of students defined as undermatched for females (14.9%), males (15.1%), African American (5.7%), Asian (14.2%), Biracial (15%), parents with some college (17.3%), parents with a college degree (12.8%), and socioeconomic status in the low- (16.3%) and middle-high income (10.7%). Operationalization 1 had the lowest proportion of students defined as undermatched for White (16.6%), Hispanic (16.6%), and parents with no college (16.8%).

A higher proportion of females were classified as undermatched as compared to males except when O5 was used. Results for race/ethnicity were mixed. Whites had the highest proportion of students defined as undermatched except for O1 and O5 when Hispanics had the highest proportion. Students identified as African American had the lowest proportion identified as undermatched and Pacific Islander the second lowest. When using O2, O5, and O6 a higher percentage of students whose parents had no college were classified as undermatched. For O1, O3, O4, O7, and O8 a higher percentage of students with parents who had some college were classified as undermatched. In comparing socioeconomic status, O1 and O5 had the highest proportion of low-income students whereas the other operationalizations had the highest proportion of middle-low income students. college were classified as undermatched. In comparing socioeconomic status, O1 and O5 had the highest proportion of low-income students whereas the other operationalizations had the highest proportion of middle-low income students.

For all definitions, between 43–50% of undermatched students had parents with a bachelor's degree or higher and less than 20% had parents with no college degree. Of those identified as undermatched over 80% were in the low-or middle-low income category, regardless of operationalization.

Table 3

Comparison of Eligibility Frontiers for Operationalizations 5 through 8

SAT

GPA	<800	800-890	900-990	1000-1090	1100-1190	1200-1290	1300-1390	≥1400
<1.0	5 6 7 8	5 6 7 8	5 6 7 8	5 6 7 8	5 6 7 8	5 6 7 8	5 6 7 8	5 6 7 8
(1.0,1.5)	5 6 7 8	5 6 7 8	5 6 7 8	5 6 7 8	5 6 7 8	5 6 7 8	5 6 7 8	5 6 7 8
(1.5,2.0)	5 6 7 8	5 6 7 8	5 6 7 8	5 6 7 8	5 6 7 8	5 6 7 8	5 6 7 8	5 6 7 8
(2.0,2.5)	5 6 7 8	5 6 7 8	5 6 7 8	5 6 7 8	5 6 7 8	5 6 7 8	5 6 7 8	5 6 7 8
(2.5,3.0)	5 6 7 8	5 6 7 8	5 6 7 8	5 6 7 8	5 6 7 8	5 6 7 8	5 6 7 8	5 6 7 8
(3.0,3.5)	5 6 7 8	5 6 7 8	5 6 7 8	5 6 7 8	5 6 7 8	5 6 7 8	5 6 7 8	5 6 7 8
≥3.5	5 6 7 8	5 6 7 8	5 6 7 8	5 6 7 8	5 6 7 8	5 6 7 8	5 6 7 8	5 6 7 8

Note: Number indicates operationalization. Colors indicate the level of selectivity to which a student has access. Darker colors indicate that the student has access to more highly selective school.

KEY

	Access to level 1
	Access to level 2
	Access to level 3
	Access to level 4
	Access to level 5

Table 4

Agreement of Eight Operationalizations (N=8,020)*

Students Classified As Undermatched	N*	%
Classified as undermatched by all 8	700	8.7
Classified as undermatched by 7/8	230	2.8
Classified as undermatched by 6/8	570	7.0
Classified as undermatched by 5/8	170	2.1
Classified as undermatched by 4/8	670	8.2
Classified as undermatched by 3/8	260	3.2
Classified as undermatched by 2/8	690	8.6
Classified as undermatched by 1/8	410	5.1
Not classified as undermatched by any	4360	54.3

* rounded to the nearest 10s by publication requirement of IES

NCES and Barron's classification systems produced significantly different results. When using NCES classifications, lower percentages of students were classified as undermatched compared to using Barron's. Operationalizations 4 and 8 were relatively similar suggesting that when using Barron's and calculating access probabilities by aggregating over student, only adjusting GPA, similar proportions were obtained.

Table 5

Proportion of Each Demographic Characteristics Defined as Undermatched Based on Eight Operationalizations (N=8,020)

		Operationalization							
	n	1	2	3	4	5	6	7	8
% Defined as undermatched		15.0	30.9	22.9	38.4	14.9	27.9	20.3	36.3
Gender									
Female	4,230	15.17	32.59	24.67	39.39	14.89	29.02	22.12	36.79
Male	3,510	15.13	29.32	20.97	37.78	15.10	27.18	18.43	36.21
Race/Ethnicity									
White	5,000	16.56	35.28	27.03	44.38	16.56	32.43	24.10	41.94
African American	850	6.15	14.32	8.28	16.45	5.68	11.48	6.86	14.91
Asian	780	15.38	24.87	18.59	31.92	14.23	22.69	17.18	30.26
Hispanic	290	16.61	31.83	19.03	32.53	17.65	27.68	16.96	31.49
Biracial	770	15.25	28.81	19.30	35.20	14.99	25.03	16.43	33.51
American Indian	30	12.9	22.58	16.13	29.03	12.9	19.35	16.13	25.81
Pacific Islander	20	8.70	30.43	8.70	34.78	8.70	26.09	8.70	34.78
Parental Education									
No College	1,310	16.77	35.15	23.51	37.83	17.61	30.63	20.75	34.53
Some College,	2,380	18.12	33.98	25.23	41.76	17.33	29.98	22.04	38.86
Bachelor's Degree	4,050	12.88	28.13	21.49	37.11	12.76	26.35	19.39	35.80
Socioeconomic									
Low (< $50,000)	3,230	16.59	32.53	23.22	37.83	16.34	28.93	20.47	34.82
Middle Low (50 – 100)	3,140	15.56	32.81	25.70	41.93	15.69	29.91	22.83	39.89
Middle High (100 – 200)	1,240	11.50	25.99	17.89	34.74	10.69	23.48	16.60	34.25
High (> $200,000)	400	9.45	18.16	14.18	28.36	9.45	17.41	10.95	27.86

Operationalizations 1 and 5 produced lower proportions of students identified as undermatched. These definitions both used NCES classification and calculated access probabilities using all applications and differed only by GPA.

We conducted chi-square tests for independence between each demographic variable and each operationalization (see Table 6). Statistically significant results were found for each demographic variable although the number of statistically significant results varied across demographic variables. For gender, three operationalizations (2, 3, 7) were statistically significant, suggesting a relationship between gender and being classified as undermatched when using these three operationalizations.

Fewer than half of the students were consistently defined as undermatched for all eight operationalizations, thus illustrating the importance of clearly and formally defining variables.

For parental education, O1–O6 were statistically significant while O7 and O8 were not. Thus, for operationalizations one through six parental education is associated with being undermatched, but this association is not present for definitions seven and eight. All operationalizations showed a statistically significant relationship between being undermatched and race/ethnicity as well as socio-economic status, meaning there was inconsistency across operationalizations.

We then limited our analysis to only those students who were classified as undermatched and examined the proportion of students in each demographic category (Table 7). Of those classified as undermatched a higher percentage of students were female for all definitions. The difference between females and males was greatest for O7 (57.5% vs. 39.7%) and least for O5 (52.9% vs. 44.5%). For race/ethnicity, White students were the highest proportion identified as undermatched. Approximately 7–10% of those undermatched identified as Asian or Biracial, 3–4% identified as African American or Hispanic and less than .5% were American Indian or Pacific Islander across all eight operationalizations. For all definitions, between 43–50% of undermatched students had parents with a bachelor's degree or higher and less than 20% had parents with no college degree. Of those identified as undermatched over 80% were in the low- or middle-low income category, regardless of operationalization.

Table 6

Chi-square Test Statistic and p-value for Testing Independence between Characteristic and Undermatch for Eight Operationalizations using a Bonferroni Adjustment Significance Level of $\alpha = 0.05/8 = 0.00625$

	Op1	Op2	Op3	Op4	Op5	Op6	Op7	Op8
Gender (M & F Only) (n = 7740)	0.00 (0.985)	9.42 (0.002)*	14.63 (0.000)*	2.04 (0.154)	0.05 (0.818)	3.11 (0.078)	15.78 (0.000)*	0.25 (0.614)
Race/Ethnicity (White, African Amer, Asian, Hisp, & Birace) (n = 7680)	61.36 (0.000)*	167.63 (0.000)*	165.97 (0.000)*	268.13 (0.000)*	68.97 (0.000)*	176.33 (0.000)*	151.71 (0.000)*	252.83 (0.000)*
Parental Education (n = 7740)	35.29 (0.000)*	35.88 (0.000)*	12.10 (0.002)*	14.12 (0.001)*	33.11 (0.000)*	14.38 (0.001)*	6.53 (0.032)	8.73 (0.013)
Socio-economic (n = 8000)	28.67 (0.000)*	53.86 (0.000)*	48.94 (0.000)*	41.07 (0.000)*	33.56 (0.000)*	41.97 (0.000)*	44.65 (0.000)*	35.12 (0.000)*

*p < 0.00625

Table 7

Comparing Proportions of Students Defined as Undermatched in each Individual Operationalization with Students Who Were Identified as Undermatched in All Operationalizations

	Operationalization								
	1	2	3	4	5	6	7	8	All
n	1210	2480	1840	3080	1190	2230	1630	2910	700
Gender									
Female	53.28	55.72	56.89	54.09	52.85	54.97	57.49	53.45	57.68
Male	44.07	41.58	40.11	43.02	44.46	42.70	39.74	43.63	40.03
Race/Ethnicity									
White	68.63	71.19	73.57	71.93	69.38	72.52	73.96	71.92	71.74
African American	4.32	4.89	3.81	4.51	4.03	4.34	3.56	4.33	3.30
Asian	9.96	7.84	7.90	8.08	9.31	7.92	8.23	8.10	10.19
Hispanic	3.98	3.72	3.00	3.05	4.28	3.58	3.01	3.12	3.30
Biracial	9.71	8.93	8.07	8.76	9.65	8.59	7.74	8.82	8.46
Parental Education									
No College	18.17	18.55	16.73	16.03	19.30	17.91	16.65	15.48	18.51
Some College	35.77	32.65	32.70	32.22	34.56	31.92	32.19	31.72	33.72
Bachelor's Degree	43.32	46.06	47.47	48.80	43.37	47.81	48.28	49.81	45.34
Socioeconomic									
Low (< $50,000)	44.40	42.38	40.82	39.58	44.21	41.76	40.54	38.55	45.34
Middle Low (50 – 100)	40.50	41.58	43.92	42.67	41.28	41.99	43.98	42.95	41.18
Middle High (100 – 200)	11.78	12.97	12.04	13.92	11.07	12.98	12.59	14.52	10.90
High (> $200,000)	3.15	2.95	3.11	3.70	3.19	3.13	2.70	3.84	2.44

Discussion and Implications for Research and Practice

Fewer than half of the students were consistently defined as undermatched for all eight operationalizations, thus illustrating the importance of clearly and formally defining variables. In this section we will highlight our key findings and discuss how these findings can inform and improve assessment work.

Methods Influence Definitions

In past studies the percentage of students identified as undermatched varied from 28% to 62% and their demographic breakdowns differed (Belasco & Trivette, 2015; Bowen et al., 2009; Rodriquez, 2013; Smith et al., 2013). These variations are likely the result of studying different populations of students and applying different techniques. Our study

illustrates that even using the same dataset and techniques but slightly changing variables can result in significant differences in the percentage and characteristics of students identified as undermatched. In other words, methods matter. This finding has implications for student learning assessment work. Competency in a certain discipline can be measured through various methods: standardized tests, comprehensive exams, or portfolios. The percentage of students who pass and the measure of student learning can vary based on assessment given. Therefore, when determining which student learning assessment to administer, it is important to consider the consequences of each of the methods.

Definitions Influence Subpopulations and Interpretations

Different operationalizations can tell different stories about subpopulations of students. For each category of students, the range of who is defined as undermatched varies significantly. Students whose parents have no college are defined as undermatched at the highest proportions for O2, O5, O6; students whose parents have some college have the highest proportions for O1, O3, O4, O7, and O8. The proportion of African Americans identified as undermatched ranges from 5.7% to 16.5%: three times as many African Americans were identified as undermatched using O4 compared to O1.

The study also illustrates the importance of recognizing limitations of the variables used in operationalizations and the consequences of these limitations on results and implications.

Our study also illustrates how the population of students can influence interpretations of results. In examining undermatch, the results and subsequent conclusions differ when comparing the demographics of undermatched students based on the total student population (Table 5) to demographics of undermatched students based on only those defined as undermatched (Table 7). For example, when examining O3 using the total student population (Table 5) similar proportions of students with parents with no college (23.5%) are as likely to be undermatched as students whose parents have some college (25.2%) and a Bachelor's degree (21.5%). Using the same operationalization but examining those students who are undermatched, almost half of the population (47.5%) of the students have parents with Bachelor's degrees versus 16.7% whose parents have no college (Table 7). The former results could be interpreted that parental education level is not related to undermatching whereas the latter may suggest that undermatching is more common for students whose parents have a college degree.

These variations in populations and subpopulations are similar to challenges faced in monitoring and reporting STEM results. Some definitions of STEM include majors such as psychology, which significantly increases the number of individuals in STEM, as well as the percentage of women and underrepresented students in STEM. Women are considered underrepresented in STEM but in some majors (e.g., biology) they may be equally or overrepresented. Additionally, whereas Asian Americans may be considered an underrepresented minority group within the college student population, they are not considered an underrepresented minority group population within STEM (NACME, n.d.). It is therefore critical that assessment professionals determine and delineate the populations for which they are reporting.

Recognize Limitations

The study also illustrates the importance of recognizing limitations of the variables used in operationalizations and the consequences of these limitations on results and implications. Because our study calculated undermatch based only on those students who had standardized test scores any student lacking this information was not included—potentially eliminating a significant number of undermatched students. For example, a student who is not considering college, or considering a college that does not require standardized test scores, may choose not to take a standardized exam. This student may be undermatched but because appropriate data to determine this undermatching was unavailable this student was not included in this study.

This too mirrors assessment practice. Institutions provide retention and graduation rates but these are often based on a cohort of full-time, direct-from-high-school students who begin in the fall. This restriction omits transfer students or students who begin part

time. Most surveys provide students two choices for gender, overlooking students who identify as transgender. Low-income students may be defined as Pell-eligible while ignoring those students whose families may make less than $50,000 but did not receive a Pell Grant. Good assessment practice requires examining who may or may not be included and the implications of these decisions.

There is No "Perfect": Strive for Clarity and Consistency

Effective assessment practice requires clear and consistent definitions but many times assessment professionals examine student populations and outcomes that lack this clarity and consistency.

Changes in operationalizations produced varied results and also illustrate that no one definition is perfect. Our results make it difficult to identify the "best," "most valid," or "most reliable" operationalization. Nevertheless, the results provide insights into consequences of different decisions. Using all applications (versus student aggregate information) results in lower proportions of students classified as undermatched because the access probability was always smaller than when calculated aggregating over a student. A higher percentage of students were classified as undermatched when Barron's selectivity classification was used. Using NCES classification of data found within NCES-sponsored restricted datasets may be easier to use but because there are fewer selectivity categories it may also decrease the proportion of students identified as undermatched.

For researchers interested in a broad definition of undermatch, using Barron's classification and calculating the student aggregate provides the greatest likelihood of being defined as undermatched. Researchers wanting to be most consistent may include only those students who were defined as undermatched for each operationalization, recognizing that this approach also minimizes the sample size. Statistically speaking, the "most valid" may be O5 because it has the highest degree of monotonicity (i.e., as GPA and SAT scores increased so did the likelihood of being admitted into a more highly selective institution). There is not one approach but many. Decisions on which operationalization to use must be made within the context of the research study, its purpose, research questions, and potential implications.

Choosing definitions and providing rationale for these decisions is needed in assessment practice. Institutions differently define categories of students "at-risk" or "under represented" and then assess their success. The definitions of the student population (i.e., at-risk) and the definition of success (e.g., retention, GPA, graduation) can lead to different results, so it is necessary that the definitions be clearly articulated and used consistently.

With so many potential choices and approaches it is important to heed the advice from Schuh and Upcraft (2001) who remind us that no assessment is perfect but one can strive for "good enough." There may not exist a universal definition for many of the topics we want to assess and we may not achieve complete accuracy (Suskie, 2009). However, we can work toward a good enough definition—one for which there is a strong rationale, one that can most effectively address the assessment questions, and most critically, one that can assist us in achieving our higher-education missions and goals.

Conclusion

Effective assessment practice requires clear and consistent definitions but many times assessment professionals examine student populations and outcomes that lack this clarity and consistency. Assessment professionals create definitions for the concepts they are examining but the consequences of these definitions may often be overlooked or not understood. Using academic undermatching as an example, we created eight unique operationalizations of undermatch that subsequently led to different results and conclusions. This study contributes to effective assessment practice by reinforcing the importance and implications of clearly defining student populations, terms, and variables.

References

Agresti, A. (2012). *Categorical data analysis* (3rd. ed.). Hoboken, NJ: Wiley.

American Council on Education (ACE). (2013). *National Commission on Higher Education Attainment*. Retrieved from http://www.acenet.edu/news-room/Pages/National-Commission-on-Higher-Education-Attainment.aspx

Banta, T.W, Jones, E.A, & Black, K. (2009) *Designing effective assessment: Principles and profiles of good practice*. San Francisco, CA: Jossey-Bass.

Barron's Educational Series. (2009). *Barron's Profiles of American Colleges of 2009*. Hauppauge, NY: Barron's Educational Series.

Bastedo, M.M., & Flaster, A. (2014). Conceptual and methodological problems in research on college undermatch. *Educational Researcher, 43*(2), 93–99. doi:0013189X1452303

Bastedo, M. N., & Jaquette, O. (2011). Running in place: Low-income students and the dynamics of higher education stratification. *Educational Evaluation and Policy Analysis, 33*(3), 318–339.

Belasco, A. S., & Trivette, M. J. (2015). Aiming low: Estimating the scope and predictors of postsecondary undermatch. *The Journal of Higher Education, 86*(2), 233–263.

Bowen, W. G., Chingos, M. M., & McPherson, M. S. (2009). *Crossing the finish line: Completing college at America's public universities*. Princeton, NJ: Princeton University Press.

Bresciani, M. J., Gardner, M. M., & Hickmott, J. (2012). *Demonstrating student success: A practical guide to outcomes-based assessment of learning and development in student affairs*. Sterling, VA: Stylus Publishing.

Executive Office of the President. (2014). *Increasing college opportunity for low-income students: Promising models and a call to action*. Retrieved from https://www.google.com/#q=increasing+opportunity+for+low+income+students+US

Fosnacht, K. (2014). *Selectivity and the college experience: How undermatching shape the college experience among high-achieving students*. Paper presented at the American Educational Research Association, Philadelphia, PA.

Fosnacht, K. (2015). *Undermatching and the first-year experience: Examining effect heterogeneity*. Paper presented at the Association for the Student of Higher Education, Denver, CO.

Heil, S., Reisel, L., & Attewell, P. (2014). College selectivity and degree completion. *American Educational Research Journal, 51*(4) 1–23. doi: 10.3102/0002831214544298

Hoxby, C. M., & Avery, C. (2012). *The missing" one-offs": The hidden supply of high-achieving, low income students* (No. w18586). National Bureau of Economic Research.

Johnson, R. A., & Wichern, D. W. (2007). *Applied multivariate statistical analysis* (6th ed.). Englewood Cliffs, NJ: Prentice Hall.

McLauglin, G. W., & Howard, R. D. (2004). *People, processes, and managing data* (2nd ed.). Tallahassee, FL: Association of Institutional Research.

Middaugh, M. F. (2011). *Planning and assessment in higher education: Demonstrating institutional effectiveness*. San Francisco, CA: Jossey-Bass.

National Action Council for Minorities in Engineering (n.d). *Underrepresented minorities in STEM*. Retrieved from http://www.nacme.org/underrepresented-minorities

National Science Foundation (2017). *Women, minorities, and persons with disabilities in science and engineering*. Retrieved from https://www.nsf.gov/statistics/2017/nsf17310/data.cfm

Oehlert, G. W. (2000). *A first course in design and analysis of experiments*. New York, NY: W.H. Freeman.

Roderick, M., Nagaoka, J., Coca, V., Moeller, E., Roddie, K., Gilliam, J., & Patton, D. (2008). *From high school to the future: Potholes on the road to College*. Retrieved from http://ccsr.uchicago.edu/publications/high-school-future-potholes-road-college

Roderick, M., Coca, V., & Nagaoka, J. (2011). Potholes on the road to college high school effects in shaping urban students' participation in college application, four-year college enrollment, and college match. *Sociology of Education, 84*(3), 178–211.

Rodriguez, A. (2013). *Unpacking the black box: Estimating the high school-level effects of undermatching among under-represented students* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses database. (No. AAI3594848)

Rodriguez, A. (2015). Tradeoffs and limitations: Understanding the estimation of college undermatch. *Research in Higher Education, 56*(6), 566–594.

Schuh, J. H., Biddix, J. P., Dean, L. A., & Kinzie, J. (2016). *Assessment in student affairs*. San Francisco, CA: Jossey-Bass.

Schuh, J. H., & Upcraft, M.L. (2001). *Assessment practice in student affairs: An applications manual*. San Francisco, CA: Jossey-Bass.

Smith, J.I., Pender, M., & Howell, J. (2013). The full extent of academic undermatch. *Economics of Education Review, 32*, 247–261.

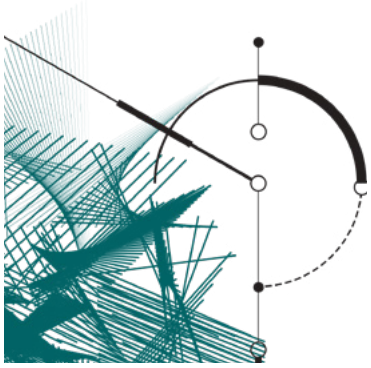
Suskie, L. (2010). *Assessing student learning: A common sense guide*. San Francisco, CA: Jossey-Bass.

Toutkoushian, R., & Stollberg, R. (2015). *Talking 'bout my generation: Defining 'First-Generation Students" in higher education research*. Paper presented at Association of Institutional Research, Denver, CO.

U.S. Department of Education, National Center for Education Statistics (NCES). (2014). *Education Longitudinal Study of 2002* (ELS: 2002). (Data file). Restricted data license.

Abstract

Externally imposed assessment requirements in higher education call for documented attention to using assessment results for program improvement. Although this systematic process promises to lead to better learning outcomes it has also been challenged as ineffective and even harmful. What can make assessment truly meaningful and move beyond the accountability mandate? Our goal in the work described here has been to advance institutional capacity for a sustained, internally valued system of learning outcomes assessment. Our approach deems faculty engagement to be essential to drive the process and improve educational results. We propose a developmental perspective on assessment capacity, describe our effort to measure and promote a supportive climate for it in our own institution, and draw conclusions about what contributes the most to its advancement. Our results point to central roles for faculty peer attitudes and collaborative institutional leadership.



AUTHORS

John F. Stevenson, Ph.D.
University of Rhode Island

Elaine Finan, M.S.
University of Rhode Island

Michael Martel, M.A.
University of Rhode Island

Measuring Assessment Climate: A Developmental Perspective

Externally mandated requirements for assessment of learning outcomes in higher education have been in place for many years, increasingly emphasizing the use of assessment results for program improvement (Banta & Blaich, 2010; Fontenot, 2012; Kezar, 2013; Kuh & Ewell, 2010; Peterson & Augustine, 2000). What is the most effective path for getting there? In this article we draw on related literatures from the field of program evaluation dealing with evaluation capacity building (ECB) and evaluation utilization to highlight one path for moving faculty from doing assessment to using assessment results. These literatures have conceptually and empirically informed our local effort to measure and promote organizational readiness for a mature assessment system. We turn assessment toward the assessment process itself—with the same aspiration to promote internally directed, data-driven improvements. This article seeks to provide a conceptual context and rationale for our approach, show how we measured "assessment climate," describe major findings, examine how we used the findings as a catalyst, and sketch some of the organizational changes we have promoted. We present our model and its application to support our claims for "what works" to build assessment capacity. We intend this to be useful for others who are working to build assessment capacity in their own institutions.

Using the Literature on Evaluation Capacity Building to Improve Assessment

As we will document below, evaluators across a wide range of settings have studied and attempted to promote "evaluation capacity," the organizational features and individual competencies associated with successful evaluation. We view assessment as a specialized form of outcome evaluation, and research on ECB provides valuable insights into the issues faced by those who engage in assessment in higher-education settings.

Challenges shared by evaluation and assessment are readily apparent in Preskill's (2014) summary of the hard work that remains to be done to clarify means for solidifying ECB in practice. Her list of challenges included: (1) moving line staff (i.e., faculty) toward

CORRESPONDENCE

Email
jstevenson@uri.edu

using data in decision making in a "culture of inquiry"; (2) building the capacity of senior leaders (i.e., top administrators) to shape and sustain a learning culture; (3) transferring newly acquired skills to long-term, sustainable practice; and (4) evaluating the success of ECB interventions themselves (i.e., enhanced faculty competencies, effective reports, curricular improvements in response to data, and sustained assessment practice).

This article seeks to provide a conceptual context and rationale for our approach, show how we measured "assessment climate," describe major findings, examine how we used the findings as a catalyst, and sketch some of the organizational changes we have promoted.

The field of evaluation has also focused extensively on how the evaluation process can influence program improvement, with clear applicability to the assessment context (Jonson, Guetterman, & Thompson, 2014). Patton's (2008) focus on the special role of "process use" is particularly relevant for the assessment context. He defined this type of use in terms of how programs are improved by the process of doing evaluation, long before any outcome data are used to guide alterations in the program. "Evaluative thinking" is beneficial as it challenges stakeholders in the program to ask critical questions about what the intended effects of the program really are, how they could be measured, and what causal connection they have to elements of the program.

Most evaluation theories emphasize the importance of stakeholder involvement to build evaluation capacity—with accumulating empirical evidence documenting the benefits of doing so, particularly for those most directly involved in delivering the program being evaluated. Clinton (2014) showed the importance of stakeholder engagement by demonstrating its mediating effect on the impact of evaluation. Brandon & Fukunaga (2014) provided more details on the empirical support for stakeholder engagement in a systematic review of the literature, noting some problems (e.g., the importance of adequate resources for building the evaluation capacity of stakeholders) along with clear indications of the pattern of positive effects on evaluation use and influence. Botcheva, White, and Hufman (2002) incorporated the notion of "learning cultures" as an aspect of ECB. Taylor-Ritzler, Suarez-Balcazar, Garcia-Iriarte, Henry, and Balcazar (2013) tested an empirical model for personal and organizational factors affecting evaluation capacity outcomes (use of evaluation findings and incorporation of evaluation into established work processes). Taylor-Ritzler et al.'s (2013) structural equation model results suggest that favorable organizational learning capacity conditions (leadership, learning climate, resources) directly influence capacity outcomes and mediate the role of individual factors (knowledge, skills, and attitudes). In fact, in their findings there was no direct influence of individual factors (which were most likely to be affected by training and technical assistance) on manifest capacity.

Assessment Culture: Moving from Accountability to Learning

The higher-education setting evinces the same crucial role for a culture supporting faculty engagement in the assessment process and use of the results. However, becoming a "learning community" is not easy, even for institutions devoted to learning (Angelo, 1999; Axelson & Flick, 2009; Driscoll & Wood, 2007; Kezar, 2013; Kuh & Ewell, 2010; Ndoye, 2013). As in many other ECB contexts (e.g. Botcheva, White, & Huffman, 2002; Owczarzak, Broaddus, & Pinkerton, 2016; Preskill & Boyle, 2008), evaluators in higher education have struggled to move from the initial external accountability impetus for learning outcomes assessment to an internal, intrinsically motivated learning role for assessment. Fuller and Skidmore (2014) referred to a "culture of assessment" vs. a "culture of compliance" (p. 10). Jonson, et al. (2014) used the labels "improvement paradigm" vs. "accountability paradigm." Walser (2015) advocated "meeting in the middle" between the competing purposes for assessment; however, in the broader evaluation context the genuine possibility of compromise has been questioned (Patton, 2008). Leviton (2014) made this one of her challenges to ECB researchers, noting that accountability associated with external funding can distort what programs think evaluation is for, affecting the way it is viewed, valued, and conducted. Faculty are just as skeptical as staff in many other kinds of organizations about the real intent of this data collection activity as well as outraged by its effects on their already overburdened workloads (Axelson & Flick, 2009; Banta & Blaich, 2010; Blaich & Wise, 2011; Buller, 2013; Cain & Hutchings, 2015; Jonson et al., 2014; Kezar, 2013).

The factors within institutions that promote meaningful assessment have been widely discussed. Terminology for these concepts can be used in various and overlapping ways but the themes are clear. Chief among these themes is the role of a supportive culture, which

we will review in detail below. Additional factors identified as beneficial are leadership by both administrators and faculty; organizational policies and structures; mutual trust among stakeholders; and a shared vision for the goals of assessment, reflected in shared language (Angelo, 1999; Banta, Lund, Black, & Oblander, 1996; Cain & Hutchings, 2015; Kezar, 2013). Banta et al. (1996) elaborated the role of leadership, with elements including administrative commitment, represented by administrative structure and reward structure; adequate resources, including clerical support, summer faculty support, mini-grants, and technical support; and faculty and staff development opportunities.

As noted above, efforts to describe and measure aspects of the institutional and departmental environment for assessment have frequently been linked to conceptions of "culture" (Fontenot, 2012; Fuller & Skidmore, 2014; Grunwald & Peterson, 2003; Kezar, 2013; Peterson & Augustine, 2000). A focus on "assessment culture" has evolved as evaluators in the assessment context try to understand factors beyond the design of training and technical assistance (over which they usually have some control) to broader contextual forces that may facilitate or impede the desired end goal of a sustained, routinized process for improving higher education results. Fuller and Skidmore (2014) have provided a useful introduction to the concepts usually embedded in definitions and measures of culture, noting that in the United States the phrase "culture of assessment" typically refers to "the deeply embedded values and beliefs collectively held by members of an institution influencing assessment practices at their institution" (p. 10). Walser's (2015) definition aimed at an end state "… when assessment work and use is an integrated part of the college or university routine" and calls for "…faculty, staff, students, and administrators to work together" (p. 59). Sometimes the term "culture" has a broader meaning, referring to institutional precursors that are hospitable to assessment (or not), such as campus leaders' demonstrated valuing of learning from evidence; campus-wide valuing of quality of teaching, setting improvement of educational performance as a primary goal; an institutional norm embracing transparency in the service of improvement on shared goals; and valuing community, collaboration, and participation (Banta et al., 1996; Cain & Hutchings, 2015). While bemoaning the frequent vagueness of definitions of "culture" in research on assessment, Kezar (2013) generally gravitated to the broader norms-beliefs-values perspective. Her review is very helpful for demonstrating the variety of hypotheses and varied roles attributed to culture in research on assessment. She reported that organizational culture is generally found to be more important than practical, policy, and technical support for assessment in determining successful adoption. Relevant for the present study, Cain and Hutchings (2015) contrasted "culture" and "climate." They defined culture as "the long-standing way a group understands itself and its shared values," characterized as "deeply embedded and resistant to change," consistent with Kezar (2013). On the other hand, they described climate as "more immediate and changeable," involving "feelings and understandings about organizational life" (Cain & Hutchings, 2015, p. 101).

A focus on "assessment culture" has evolved as evaluators in the assessment context try to understand factors beyond the design of training and technical assistance (over which they usually have some control) to broader contextual forces that may facilitate or impede the desired end goal of a sustained, routinized process for improving higher education results.

The content of a measure of assessment culture provides more definitional specificity regarding the concepts involved. Fuller and Skidmore (2014) presented a 34-item scale (agreement on 5-point Likert scales) based on the work of Maki (2010) on principles of inclusive commitment to assessment. Their exploratory factor analysis (PCA) yielded three factors labeled Clear Commitment, Connection to Change, and Vital to Institution. High-loading items for Clear Commitment included "adequately staffed assessment office" and "clear definition of assessment." For Connection to Change the strongest items were "administrators want to know about student learning" and "assessment results are used in campus publications/ speeches." The high-loading items for Vital to the Institution included "assessment is vital to the institution's future" and "assessment and teaching (sic)."

A separate, closely related line of research has focused on faculty involvement and satisfaction with assessment as dependent variables, with a number of posited predictors. Building on the work of Grunwald and Peterson (2003), Fontenot (2012) examined attitudes, concerns, and involvement of community college faculty with assessment. Her factor analysis of Attitudes yielded two factors: Benefits (e.g., "improved the quality of education at this institution") and Faculty Reluctance (e.g., "limits time," "a distraction," "fear of results").

Promoting Meaningful Assessment with a Climate Survey

Next we turn to the development of our own measure and plans for its use.

Central to both the developmental stages and the climate scales is the conviction that formative use of assessment to improve educational outcomes calls for a major shift in perception of the role of assessment for both faculty and administrators.

Developmental Framework. To guide our work, we applied a five-stage developmental model for institutional assessment capacity (see Table 1) developed by the first author with several associates (Stevenson, 2011; Stevenson & Monteiro, 2013; Stevenson, Treml, & Paradis, 2009). The original conceptualization of the stages (Stevenson et al., 2009) was based on the literature dealing with characteristics of colleges and universities associated with good assessment practices (e.g., Angelo, 1999; Axelson & Flick, 2009; Banta et al., 1996) and more specific designations of possible stages in the development of these practices (Allen, 2004; Bresciani, Zelna, & Anderson, 2004; Wehlburg, 1999).

Although our model is specific to the assessment context, it draws on a long tradition. The literature on learning organizations (e.g., Argyris & Schon, 1978; Cousins, Goh, Clark, & Lee, 2004; Preskill & Torres, 1999) implicates the value of having a model for how improved internal processes evolve. Demonstrating the utility of this kind of approach, Rogers (2003) proposed a five-stage developmental scheme in his well-known work on diffusion of innovations in organizations. The three latter stages during implementation are most relevant for ECB: Redefining/Restructuring, during which the necessary infrastructure is developed and the innovation is adapted to fit the organization's context; Clarification, in which the internal diffusion process builds understanding of how integration can work and leads to gradual embedding across the organization; and Routinization, in which the innovation becomes an accepted, sustainable aspect of functioning. Preskill and Boyle (2008) noted the general utility of stage models, including Rogers', for understanding organizational change as an aspect of ECB.

Two particular advantages of the developmental approach are that (1) success can be defined by movement from one stage to the next, rather than only by achieving a final outcome, and (2) the strategies useful for making each step may be examined separately so that the most effective means for forward movement can be determined stage-by-stage. Classic work on individual processes of change (Norcross, Krebs, & Prochaska, 2010) has long shown the value of these two contributions. Kreiner and Herr-Zaya (2005) demonstrated the value for understanding organizational change in the ECB context, suggesting that each step may require different internal capacities and may respond to capacity building influences differently.

Planning for Use. The first author originally conceived the *Assessment Climate Scale* as a means to probe and prompt institutional movement from one developmental stage to the next (Stevenson et al., 2009). Hence the more long-term connotations of "assessment culture" seemed less appropriate than the malleable conception of "climate." Central to both the developmental stages and the climate scales is the conviction that formative use of assessment to improve educational outcomes calls for a major shift in perception of the role of assessment for both faculty and administrators. This conception calls for a move from the initial external-accountability impetus present on many campuses, with its threat of summative use and potential for superficial measures, to internal recognition of pedagogical relevance by faculty—a "culture of evidence" in Kuh and Ewell's (2010) terms. The scale drew on the pool of knowledge regarding faculty attitudes and beliefs that might inhibit or promote change toward the kind of idealized assessment culture described by Walser (2015), and anticipated Kezar's (2013) conclusion that norms, beliefs, and values will prove more important than structural progress in moving toward that goal. Our scale is not intended to measure broad cultural precursors of successful assessment, nor institutional evaluation capacity, nor is it a needs assessment. Its premise is more like that of action research (Fals Borda, 2001), aiming to speak faculty's perceived truths to those with power—power to communicate genuine belief in the value of an ideal assessment culture and support forward movement with policies, recognition, and resources.

Table 1

Building a Culture of Assessment: Developmental Stages

Stage	Description
Stage 1: Denial	"No one really cares about this and we all have more important things to do; it's a passing fad."
Stage 2: External Demand	Administration: "We have to!"
	Faculty: "*You* have to!" (denial still rampant for faculty)
	Fear/defensiveness
	Top-down pressure reduces sense of intrinsic value, "buy-in"
	Few resources of any kind devoted to assessment (workload recognition, faculty time, direct funding, staff time, technology (portfolio, web, IR, etc.), training in skills, supportive administrative structures)
	Faculty concern about trivialization of learning (reductionist, privileges surface learning, factory model, consumer model)—both genuine and defensive
	Administrators starting to send faculty to conferences, consider needs, build capacity
Stage 3: Tentative Commitment	Early adopters on board (administrators and faculty)
	Strong leadership at the administrative level (key person)
	Initial internal structures (faculty advisory committee, staff resource)
	First round public statement of learning objectives by programs is initiated
	A few faculty accepting responsibility, working with administrators
	Accredited programs ready to go
	Capacity-building (e.g., conferences, workshops) starting to pay off; more awareness of non-trivializing approaches to assessment
Stage 4: Full-scale Effort	Clear expectations and incentives at the program level—uniform, visible, insistent
	Regular monitoring of assessment progress by program, department, college, university
	Positive rewards for "completing the loop," recognizing needed improvements and acting on that recognition
	Critical mass of faculty and chairs accept necessity
	Growing recognition of potential pedagogical value of the process (intrinsic motivation)
	Formalization of support structures and decision-making structures with necessary resources
	Models available, peer support and mentoring built in
	Attention to ways of incorporating into strategic planning, aligning with overall mission and vision of the institution, connecting to college deans' concerns
	Web visibility at department, college, and university levels
Stage 5: Maintenance and Refinement	Late adopters and resisters targeted
	Mature resources and structures allow longitudinal tracking of outcomes
	Pioneers ready for more sophisticated efforts at alignment, taking risks in questioning the premises in their learning outcomes
	Leadership at every level sees the genuine value and is committed to providing the resources on a stable basis

Method

Sample

We chose department chairpersons as respondents. At our institution, chairs function as a kind of "bridge" between faculty and administrators. The administration (college deans and provost) holds them directly accountable for producing assessment reports from their departments. The new pressure on faculty workload for assessment-related activities has rapidly grown, including a number of time and competency demands: convening with colleagues to define learning outcomes for their degree programs; developing a curriculum map linking their courses and other degree requirements to those outcomes; developing ways to quantify student learning (e.g., grading rubrics); administering, scoring, and reporting on department-generated means for evaluating student work in their courses; meeting to discuss the results with colleagues and determine recommendations for future action; following up with implementation of pedagogical and curricular changes; and re-assessment. As these expectations were promulgated from the provost's level via a newly created assessment office and a joint faculty-administration committee, chairs were expected to convey the demands

and their rationale to their colleagues. Thus we saw the chairs' perspective as a particularly informative one to track the development of a mature assessment system over time, and to prompt consideration of needed changes in policies and practices for assessment.

We invited all department chairs (and the directors of department-equivalent academic programs) to participate in this survey in Fall 2009, Fall 2012, and again in Fall 2015. In 2009, 30 of 51 responded (58.8%); in Fall 2012, it was 36 of 61 (59.0%); and in 2015 it was 28 of 49 (57.1%). In order to preserve anonymity in the data set, we did not include respondent descriptors (e.g., college, gender, rank) in the survey. In 2015, 18% of the chairs indicated that they remembered taking one or both of the prior surveys, suggesting a high degree of turnover.

Survey Design

Content of the survey is organized into six major domains: (1) chairs' personal attitudes toward assessment, (2) institution-wide faculty norms regarding the value of assessment, (3) leadership commitment, (4) infrastructure support for assessment, (5) department-level implementation, and (6) university-wide implementation. Response choices range from 1=*strongly disagree* to 5=*strongly agree*. A final structured item addresses chairs' perception of how far the institution has come in the development of a useful, sustainable assessment system, using the five-stage model described in Table 1: (1) denial ("It's a passing fad"), (2) external demand ("The administration says we must; give us the time and resources or do it yourselves"), (3) tentative commitment ("Leaders are committed and some of us are too"), (4) full-scale effort ("Most of us accept the necessity and there are policies and resources available to help"), and (5) maintenance and refinement ("We see the value and regularly use the results at all organizational levels"). The original 2009 survey consisted of 37 items; we added seven items for the 2012 version for a total of 44 items; and in 2015 still further revisions were made, leading to a total of 51 items. The added items addressed changing facts on the ground at our institution. We provided an open-ended space for qualitative comments in all three years (see Table A in the Appendix for the current version of the instrument).

Procedure

Content of the survey is organized into six major domains: (1) chairs' personal attitudes toward assessment, (2) institution-wide faculty norms regarding the value of assessment, (3) leadership commitment, (4) infrastructure support for assessment, (5) department-level implementation, and (6) university-wide implementation.

We administered the survey online via Survey Monkey, with an invitation to participate and IRB assurances accompanied by an e-mailed link, followed by a brief introduction at the beginning of the survey explaining its purpose and defining key terms. We chose mid-October as a promising time in the annual calendar of chairs' duties, and the survey was thus administered during that time-frame for each of the three iterations. Chairs were given three weeks to respond, with two reminders sent during that period.

Survey Results

Item-level Responses

We tested significance of changes over time at the item level with one-way analyses of variance (see Table A). These provide evidence that the chairs perceived forward progress on some important issues. Chairs responding in 2015 were less likely to agree that faculty resist assessment for fear of negative consequences (item #9). Chairs in 2015 were more likely to agree that faculty value transparency (item #10), that the university tracks assessment evidence and results (item #19), and that the university is defining, measuring, and reporting university-wide learning outcome objectives on a regular basis (item #47).

Other item-level results indicate perceived movement in a negative direction regarding the value of assessment. In 2015 there was significantly lower agreement that college deans recognize and support assessment (item #14) and that programs that do not comply with assessment reporting requirements will receive negative consequences (item #22).

The last item on the survey (#51) measured what the chairs thought about the university's current stage in the establishment of program-level assessment. Figure 1 graphically displays the modal response, Stage 2, "External Demand," indicating that administrative leaders require faculty compliance to meet assessment demands without added support for faculty. This was selected by 50.0% of the respondents. The second highest choice was for

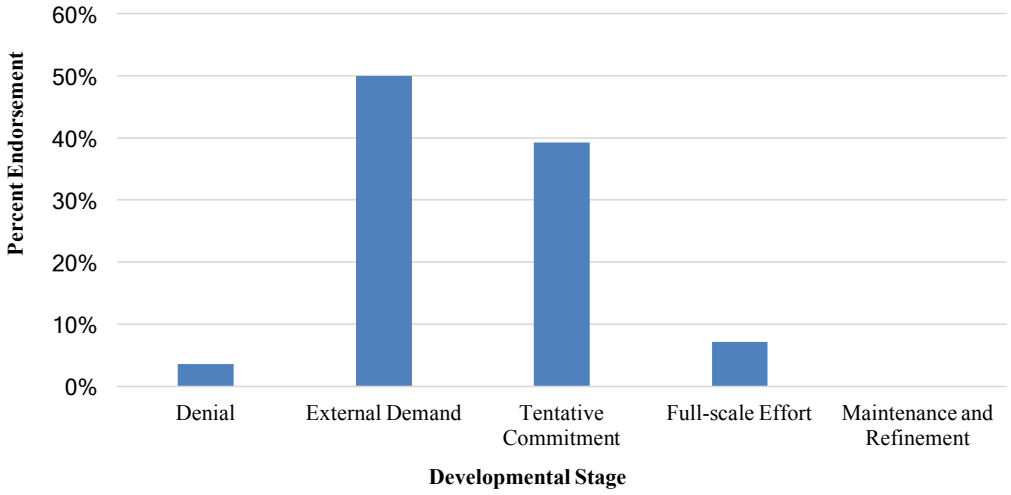


Figure 1. Assessment Climate Survey (2015): Responses to Question 51, "In which stage in the development of learning outcomes assessment would you judge this institution to be?"

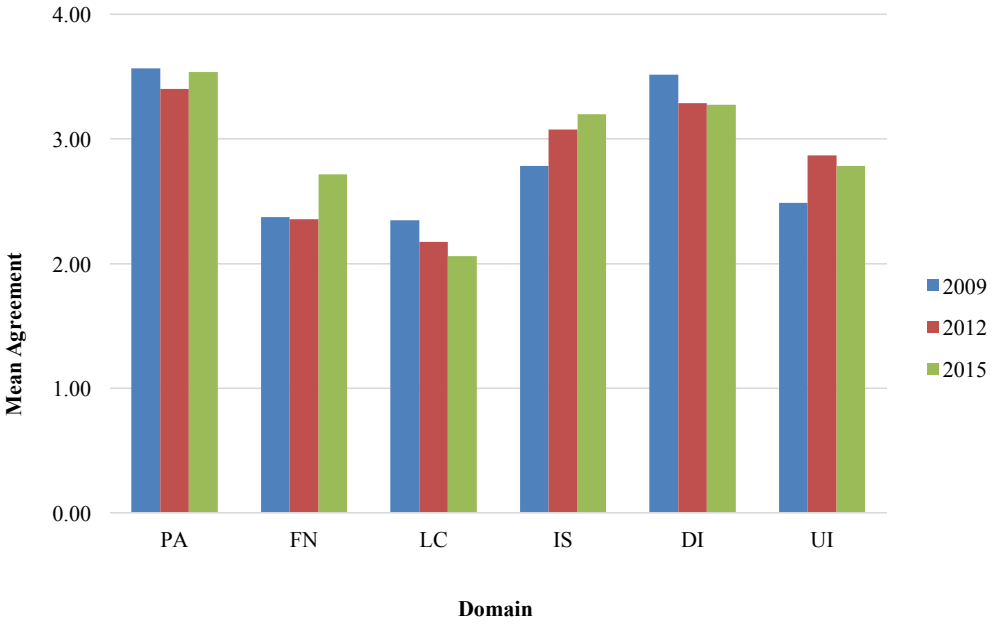


Figure 2. Assessment Climate Survey Domain Scale Averages: 2009, 2012, 2015. Domain scales are Personal Attitude toward Assessment (PA); Faculty Norms (FN); Leadership Commitment (LC); Infrastructure Support for Assessment (IS); Department-level Implementation (DI); and University-wide Implementation (UI).

Stage 3, "Tentative commitment," indicating chairs' sense that faculty are starting to join with campus leadership in institutionalizing assessment, selected by 39.3% of the respondents. No one endorsed Stage 5, "Maintenance and refinement."

Domain Scale Patterns

Figure 2 presents results for the six domain scales, with means calculated on the basis of relevant items available for all three time points (averaging agreement with positively worded items and disagreement with negatively worded items, which are denoted "R" in Table A). Table 2 provides some statistical information about the domain scales based on the 2015

Table 2

Assessment Climate Domain Scale Properties and Correleations for 2015 Sample

Domain Scales	No. of Items	Mean	S.D.	Alpha	Inter-scale Correlations FN	LC	IS	DI	UI
Personal Attitude toward Assessment (PA)	6	3.34	.674	.694	.590**	.353	.523**	.538**	.555**
Faculty Norms (FN)	6	2.71	.561	.677	-	.231	.432*	.354	.277
Leadership Commitment (LC)	10	2.21	.555	.747		-	.532**	.097	.529**
Infrastructure Support for Assessment (IS)	10	3.07	.554	.814			-	.204	.505**
Department-level Implementation (DI)	9	3.29	.726	.785				-	.144
University-wide Implementation (UI)	8	2.84	.442	.613					-

Note. N = 28.
*p < .05, **p < .01

responses, including Cronbach's Alpha reliabilities and inter-scale correlations. The scales have Alphas ranging from .61–.81, suggesting some degree of internal consistency, although they were lower than we would have liked for University-wide Implementation, Personal Attitude, and Faculty Norms. Personal Attitudes correlated positively with all other scales except Leadership Commitment. Leadership Commitment was strongly correlated with Infrastructure Support and University-wide Implementation (p<.01). Infrastructure Support was positively correlated with all of the other scales except Department-level Implementation. Intriguingly, Department-level Implementation was not significantly correlated with University-wide Implementation.

The patterns over time clearly indicate that chairs consistently viewed the value of assessment for their own departments as relatively high and believed infrastructure support for assessment was steadily rising.

Table 3 reports analyses of domain-level patterns of change over time. Two of the scales achieved statistical significance in one-way analyses of variance. The chairs' perceptions of faculty norms supportive of assessment went up significantly in 2015 and perceptions of University-wide Implementation increased significantly between 2009 and 2012 and remained at that level in 2015. The patterns over time clearly indicate that chairs consistently viewed the value of assessment for their own departments as relatively high and believed infrastructure support for assessment was steadily rising. Significant item-level changes reported above are consistent with those trends, and several item-level analyses in the Infrastructure Support domain also approached significance in the positive direction. On the other hand, Leadership Commitment remained the lowest domain score and continued a downward trend from past administrations. The significant item-level changes within that domain reflected the negative trend.

Table 3

Significance of Domain Scale Change by Mean Agreement over Time

Scale	Mean Agreement* 2009	2012	2015	F	df	p<
Personal Attitude toward Assessment	3.57	3.40	3.54	.501	91	n.s.
Faculty Norms	2.37	2.36	2.71	3.94	91	.023*
Leadership Commitment	2.35	2.17	2.06	1.98	91	n.s.
Infrastructure Support for Assessment	2.78	3.08	3.20	2.22	91	n.s.
Department-level Implementation	3.51	3.29	3.27	.891	91	n.s.
University-wide Implementation	2.49	2.87	2.78	3.98	91	.022*

Note. Mean agreement calculated for items included at all 3 time points.
*p < .05

We also examined the relationship between the six domain-based scales and the chairs' perceived stage of institution-wide assessment (item #51) for the 2015 responses, using data for the four stages with responses. A stepwise discriminant function analysis (DFA) indicated that Leadership Commitment was clearly playing the dominant role in determining judgment of stage. A single function solution with an Eigenvalue of .736 located Stage 1 and Stage 2 very close together and spread Stage 3 and Stage 4 further along the single dimension (Wilks' Lambda = .576; X^2 = 13.52; p<.004; 50.0% of the cases classified correctly). With a more liberal F-to-enter, the first function (Eigenvalue 1.152; canonical correlation of .732, explaining 83.3% of the variance) again featured Leadership Commitment with a loading of .855, followed by Faculty Norms (loading .627). Once more the first two stages were literally on top of each other with stages 3 and 4 spread out along the first dimension (Wilks' Lambda = .375, X^2 = 23.05; p<.006; 60.7% of the cases correctly classified).

One theme was very persistent: the workload burden remained a severe impediment, even for those who saw value in the work.

Qualitative Responses

We analyzed qualitative responses to the final open-ended item of the survey, inductively developing themes, and found some shifts over time in those responses. After 2009 there was less concern about technical support, and by 2015 there was more recognition of assessment's value. One theme was very persistent: the workload burden remained a severe impediment, even for those who saw value in the work. The chairs' sense that the burden was compounded by a sense of the task's futility did diminish over time. It also appeared that there was some positive anticipation of the potential value of assessment: in 2009, it was recognized as an expectation for new programs (an accountability motivation); by 2015 there was more grasp of the potential for internal use and consistency with faculty values, although those were offset by the frustration with lack of workload relief, recognition, or reward.

Discussion and Action Steps Taken

Using the Results to Prompt Action

We began our work with an "action research" conception of the survey as a means for promoting reflection and change within our institution, and we discuss our results in that context. Our survey design was improved by an early and ongoing relationship with the campus assessment office, which also actively promoted attention to the findings. The survey process itself was an intervention, influencing chairs' views regarding assessment by highlighting the availability of resources and portraying potential for internal utility. Turning to our use of the survey results, after each administration we presented the findings to various decision-making groups in a "good news–bad news" framework, drawing on prescriptions from the literature. The rationale for the survey was clearly stated in our internal reports:

> "As an organization developing the capacity to conduct and learn from program-level assessment of student learning outcomes, our institution is investing resources and implementing policies for assessment. The survey gives us something with which to benchmark our progress over time and identify strengths and weaknesses in our overall progress. The findings can inform policy and resource allocation decisions as we go forward."

Limitations

The limitations of our methodology were acknowledged at the outset of our internal reports, anticipating possible resistance to the findings by some decision makers. These limitations include: (1) the sample size is small, reflecting our choice of chairs as the population of interest, making statistical significance more difficult to achieve; (2) the response rate is not as high as we would have liked, although it is not out of line with other similar survey contexts; and (3) the overlap between samples over time presents a statistical issue, and the effort to preserve anonymity in order to increase trustworthiness of responses, as well as the high turnover rate, make it impossible to consider a "repeated measures" approach to analyses of change over time. Thus it is best to consider each year's quantitative results as a cross-sectional snapshot of what a majority of chairs thought at that time, with the qualitative

comments as a "triangulating" set of evidence. Moving beyond the internal perspective on limitations we note that the generalizability of our scale and its findings to other academic settings remains uncertain, particularly for institutions of varying sizes and purposes. Our own setting is a mid-sized public research university. We believe that locally tailored variations will make the approach we describe here maximally effective. The scale's dimensions and the developmental stage model guiding it are more generalizable, as we drew them from many published sources cited above.

We concluded that supportive infrastructure enables but does not motivate. The demand was increasingly clear to the chairs but the leadership's genuine commitment to properly support the work and use the findings was not.

Good News–Bad News

To convey the significance of our findings, the "good news" we presented in our most recent internal report included the high level of chairs' own reported valuing of assessment, which remained the highest domain scale score across all three time points, with department-level implementation remaining second highest. Infrastructure support, including things like faculty training, models for what is expected in reports, clear policies for reporting, an office providing many forms of assistance, and a useful website, was the third highest domain and shows a steady positive trend over time. We concluded that we appeared to be on the right track for providing what is needed to make assessment both feasible and useful.

Chairs' view that faculty norms were supportive of assessment made a significant upward jump in the 2015 results. More chairs agreed that faculty value transparency, including open discussion of learning outcomes; fewer agreed that their colleagues believe assessment is unrelated to a concern for student learning or that faculty resist assessment due to fear of negative findings. Agreement that the institution's faculty is committed to the goal of having every student graduate with abilities and values consistent with the mission and strategic plan went up fifteen points between 2009 and 2015. This suggests that chairs saw their own colleagues moving toward more acceptance of the necessity of engaging in these activities, and more recognition of the value of doing so. Our presentation of those positive conclusions treated them as confirmation of meaningful progress, consistent with recommendations in the literature cited above.

We followed with some "bad news," also based on the comparison of our findings with recommendations in the literature. Leadership Commitment remained the lowest domain score and continued a downward trend from past administrations. Significant downward item-level changes (in support from deans and a lack of negative consequences for noncompliance) provided more concrete substantiation of that concern. Increased administrative tracking (#19) may not be seen as a positive thing if it is just considered "bean-counting" (as one qualitative comment suggested).

Most dramatic from our standpoint was where chairs believed the university was in terms of developmental stage of growth in assessment capacity. Stage 2, "External Demand," with administrative leaders requiring faculty compliance to meet that demand, was not what we expected to be the modal response. In prior administrations we had not included that final item, believing that we could derive conclusions about stage from the domain scales. Clearly we were wrong, as we had previously judged the university to be between Stage 3 (tentative commitment) and Stage 4 (full-scale effort) based on the chairs' own positive attitudes, their perceived level of implementation within their own programs, and their perceptions of the improving infrastructure. The Discriminant Function Analysis helps with understanding what was going on: leadership commitment was the most powerful indicator for chairs of whether the institution was really moving toward an assessment system that is internally valued at all levels. The qualitative responses, although from a small subset of the respondents, amplified the level of frustration with administrative leadership. We concluded that supportive infrastructure enables but does not motivate. The demand was increasingly clear to the chairs but the leadership's genuine commitment to properly support the work and use the findings was not.

We presented all of those findings in a series of decision-making contexts: first within the university's assessment office, where data analysis took place and some thoughts about possible recommendations were generated; then to the university-wide assessment committee with representation from both administration and faculty; later as one part of an agenda for

a series of meetings arranged by assessment office staff with each college dean; and lastly to the "Deans' Council," which is chaired by our provost. Formats for presentation varied. We engaged the university assessment committee in an active discussion with graphic presentation of major quantitative results, the qualitative comments, and skeletal recommendations used to stimulate ideas for new policies and practices. The deans and provost got an "elevator talk" executive summary and a few recommendations in an attempt to generate ideas for next steps. Following those presentations, we conveyed a final complete report with more detailed recommendations to the chairs themselves.

Actions Taken

Most of the tangible changes we can point to were generated by the university assessment committee. Their deliberations in response to the results led to (1) an annual recognition event honoring assessment reports that meet specified peer-review criteria, (2) agreement on the need to offer peer models showing how assessment can be both meaningful (internally useful) and manageable (feasible with limited resources), and (3) clearer emphasis on assessment reporting and use in the cyclic academic program review process, which provides an opportunity for departments to negotiate for resources and demonstrate their accomplishments. In one large college the dean's recognition of the survey's implications led to creation of a new college-level committee to focus on supporting and tracking departmental assessment activities.

Two complements to the survey release process bolstered its impact. One was a change in assessment policy to reduce the reporting burden for degree programs with their own external accreditation reporting requirements. The other was the developing plan for assessing a new general education program, launched in the fall of 2016, which imposed university-wide learning outcome requirements. The assessment needs for that new program are driving a new set of resources and training activities, new technical advances in data management for assessment, and rapidly expanding faculty awareness of how assessment "works." It remains to be seen, however, whether the leadership for this transformation will be able to emphasize "learning culture" over "accountability culture."

Conceptual Implications of the Findings: Stage Progression

Leaders of campus assessment, both faculty and administrative, put an intensive amount of effort into developing assessment policies, necessary governance structure, a variety of training opportunities and on-line resources, and various types of incentives (e.g., mini-grants, off-campus conference opportunities). It is not surprising that they would expect "infrastructure support" accomplishments to give chairs a sense of the remarkable progress the university is making. However, our results confirm and elaborate what others have found before us: leadership and campus culture provide the impetus for integrating assessment into a meaningful process of program improvement. Taylor-Ritzler et al. (2013) contrasted individual capacity building with institutional leadership and organizational culture, showing that in their data individual factors only had influence via the mediating role of those organizational factors. In the higher-education context, Kezar's (2013) review found "organizational culture" and "leadership" to be consistently recognized as primary sources of constraint and facilitation, followed by "organizational policies, practices, and structures." Her discussion of campus culture posited "clarity and commitment of leadership" as a force for transforming culture. Based on her analysis, leaders appear to have pervasive means to influence the assessment process.

However, our results confirm and elaborate what others have found before us: leadership and campus culture provide the impetus for integrating assessment into a meaningful process of program improvement.

As previously noted, one of the helpful aspects of a stage perspective is that it allows for identifying differing capacity-building strategies as most effective in different stages. The university studied in this case example seems to be "stuck" in some ways despite notable progress on faculty attitudes and infrastructure to support assessment. It may help to consider whether differing emphases might help it to move forward developmentally. We have local evidence from several years of peer-reviewed assessment reports showing that most degree programs are now compliant with requirements and doing a reasonable job of meeting them (Finan, Stevenson, Monteiro, & Martel, 2015). However, the *Climate Survey* adds some

key stakeholder perspective on how the process is perceived, the extent of true integration into decision making, and the perceived barriers. The qualitative comments are especially telling for the chairs' frustration with a mandate for activity without academic value. And yet the value seems obvious to evaluators: programs are routinely learning from their students about what is working well (and can be celebrated) and what is not (and calls for some experimenting with altered pedagogy and/or curriculum). Evaluative thinking in the form of "curriculum maps" that link program requirements to intended learning outcomes can drive the assessment process. Perhaps the early emphasis on infrastructure development, policies, and training have moved the accountability mandate forward (to Stage 2/3) at the expense of a recognition that the purpose is truly aligned with what faculty themselves value. As in other evaluation contexts, evaluators may see "empowerment" where those who are doing the work see "exploitation" (Stevenson, Mitchell, & Florin, 1996).

What can move our institution past that developmental impasse, to Stages 4 and 5? Cain and Hutchings (2015, p. 96) advocated paying close attention to "how assessment is talked about" and linked to faculty values and expertise. Fuller and Skidmore's (2014) "Connection to Change" factor seems especially relevant for our predicament, and Angelo's (1999) prescription identified shared motivation and shared language as essential pillars for the transformation process. Owczarzak et al. (2016) and Jonson et al. (2014) warned of the dangers of leadership focus on accountability, and Leviton (2014) questioned whether leaders always share evaluators' rosy view of the value of "evaluative thinking." Owczarzak et al. (2016) also offered some helpful suggestions for progress that can have relevance for the higher-education context, including the use of peer-nominated experts to provide ongoing consulting, and accessible qualitative narratives documenting how assessment can work for departments. An important point made by several authors including Kezar (2013) is that faculty leaders are as important as administrative leaders. Respected peers can influence the perception of norms, and provide models for positive use. We recognize now that our survey should have done more to explore that aspect of leadership and will do so in the future.

For some challenges it is difficult to find a prescription. Workload burden reduction and staff turnover (especially in key roles like chair) remain difficult to address.

Conclusions

From the perspective of the chairs in our study it was not faculty acceptance nor even the enabling infrastructure that was most important for determining how close we were to a fully realized assessment culture. The most important domain in our climate framework was the communicated support from administrative leaders and their commitment to motivate assessment as an internally useful process. Those were the keys to a sustained quality-improvement system. We conclude that interventions to improve infrastructure and assessment competencies are needed on a continuing basis but they will not lead to the desired goal without clear messages and incentives from leaders. Heed Leviton's (2014) advice: understand what top managers believe about the value of assessment, and watch out for the distorting effects of an accountability culture. Getting from grudging compliance to enlightened conversation takes leadership that believes in transparency, learning from evidence, and collaboration.

We conclude that interventions to improve infrastructure and assessment competencies are needed on a continuing basis but they will not lead to the desired goal without clear messages and incentives from leaders.

We view our measure as a means to the end of moving the developmental process along, and attempted to leverage the results of our periodic surveys via the policy-making channels of the institution. Campus assessment policies are now evolving from efforts to clarify expectations, provide training and consultation, and establish peer review feedback, toward greater recognition for success, models for good practice, and integrated academic program review policy. The latter has resource implications for departments and aligns departmental objectives with the college and university mission. This marks it as a particularly hopeful sign. We continue to aspire to promote collegial conversations informed by data as well as academic values, leading to creative insights regarding pedagogy and curriculum. This enterprise may best be served by the continuing recruitment of highly respected faculty leaders. Advancement of genuine enthusiasm for the effort involved will also take a broader initiative to enhance transparency, trust, and confidence that contributions to assessment will be recognized, rewarded, and respected as time-consuming professional achievements.

References

Allen, M.J. (2004). *Assessing Academic Programs in Higher Education*. Bolton, MA: Anker.

Angelo, T.A. (1999). Doing assessment as if learning matters most. *AAHE Bulletin, 51*(9), 3–6.

Argyris, C., & Schon, D. (1978). *Organizational learning: A theory of action perspective*. Reading MA: Addison Wesley.

Axelson, R., & Flick, A. (2009). Sustaining assessment: A post-epidemiological approach using the program evaluation standards. *Assessment Update, 21*(3), 5–7.

Banta, T.W., & Blaich, C. (2010). Closing the assessment loop. *Change: The Magazine of Higher Learning, 43*(1), 22–27.

Banta, T.W., Lund, J.P., Black, K.E., & Oblander, F.W. (1996). *Assessment in practice: Putting principles to work on college campuses*. San Francisco: Jossey-Bass.

Blaich, C., & Wise, K. (2011). *From gathering to using assessment results: Lessons from the Wabash national study*. National Institute for Learning Outcomes Assessment, Occasional Paper #8.

Botcheva, L., White, C.R., & Huffman, L.C. (2002). Learning cultures and outcomes measurement practices in community agencies. *American Journal of Evaluation, 23*(4), 421–434.

Brandon, P.R., & Fukunaga, L.L. (2014). The state of the empirical research literature on stakeholder involvement in program evaluation. *American Journal of Evaluation, 35* (1), 26–44.

Bresciani, M.J., Zelna, C.L., & Anderson, J.A. (2004). *Assessing Student Learning and Development: A Handbook for Practitioners*. Washington, DC: National Association of Student Personnel Administrators

Buller, J.L. (2013). Academic leadership 2.0. *Academe, 99*(3), 28–33.

Cain, T. R., & Hutchings, P. (2015). Assessment at the intersection of teaching and learning. In G.D. Kuh, S. O. Ikenberry, N. A. Jankowski, T. R. Cain, P. T. Ewell, P. Hutchings, & J. Kinzie (Eds.). *Using evidence of student learning to improve higher education*. (pp. 95–116). San Francisco, CA: Jossey-Bass.

Clinton, J. (2014). The true impact of evaluation: Motivation for ECB. *American Journal of Evaluation, 35*(1), 120–127.

Cousins, J.B., Goh, S.C., Clark, S., & Lee, L.E. (2004). Integrating evaluative inquiry into the organizational culture: A review and synthesis of the knowledge base. *Canadian Journal of Program Evaluation, 19*(2), 99–141.

Driscoll, A., & Wood, S. (2007). *Developing outcomes-based assessment for learner-centered education*. Sterling, VA: Stylus.

Fals Borda, J. (2001). Participatory (action) research in social theory: Origins and challenges. In P. Reason & H. Bradbury (Eds) *Handbook of Action Research* (pp. 27–37), London: Sage.

Finan, E., Stevenson, J.F., Monteiro, K., & Martel, M. (2015, October). *Using peer review feedback to guide assessment capacity-building*. Paper presented at the annual meeting of the American Evaluation Association, Chicago, IL.

Fontenot, J.S. (2012). *Community college faculty attitudes and concerns about student learning outcomes assessment*. Doctoral dissertation, University of Illinois at Urbana–Champaign, Urbana, IL. Retrieved from http://hdl.handle.net/2142/34321

Fuller, M.B., & Skidmore, S.T. (2014). An exploration of factors influencing institutional cultures of assessment. *International Journal of Educational Research, 65*, 9–21.

Grunwald, H., & Peterson, M.W. (2003, April). Factors that promote faculty involvement in and satisfaction with institutional and classroom student assessment. *Research in Higher Education, 44*(2), 173–204.

Jonson, J.L., Guetterman, T., & Thompson, R.J. (2014). An integrated model of influence: Use of assessment data in higher education. *Research and Practice in Assessment, 9*, 18–30.

Kezar, A. (2013). Institutionalizing student outcomes assessment: The need for better research to inform practice. *Innovative Higher Education, 38*(3), 189–206.

Kreiner, P.W., & Herr-Zaya, K. (2005). *Diffusion of outcome evaluation practices in substance abuse prevention programs: Organizational and interorganizational predictors*. Working paper, Schneider Institute for Health Policy, Brandeis University, Waltham, MA.

Kuh, G.D., & Ewell, P.T. (2010). The state of learning outcomes assessment in the United States. *Higher Education Policy and Management, 22*(1), 9–28.

Leviton, L.C. (2014). Some underexamined aspects of evaluation capacity building. *American Journal of Evaluation, 35*(1), 90–94.

Maki, P. (2010). *Assessing for learning: Building a sustainable commitment across the institution* (2nd ed). Sterling, VA: Stylus.

Ndoye, A. (2013). Promoting learning outcomes assessment in higher education: Factors of success. *Journal of Assessment and Institutional Effectiveness, 3*(2), 157–175.

Norcross, J.C., Krebs, P.M., & Prochaska, J.O. (2010). Stages of change. *Journal of Clinical Psychology, 67*(2), 143–154.

Owczarzak, J., Broaddus, M., & Pinkerton, S. (2016). Audit culture: Unintended consequences of accountability practices in evidence-based programs. *American Journal of Evaluation, 37*(3), 326–343.

Patton, M.Q. (2008). *Utilization-focused evaluation* (4th ed.). Thousand Oaks, CA: Sage.

Peterson, M.W., & Augustine, C. (2000). External and internal influences on institutional approaches to student assessment: Accountability or improvement? *Research in Higher Education, 41*(4), 443–479.

Preskill, H. (2014). Now for the hard stuff: Next steps in ECB research and practice. *American Journal of Evaluation, 35*(1), 116–119.

Preskill, H., & Boyle, S. (2008). A multidisciplinary model of evaluation capacity building. *American Journal of Evaluation, 29*(4), 443–459.

Preskill, H., & Torres, R.T. (1999). *Evaluative inquiry for learning in organizations*. Thousand Oaks, CA: Sage.

Rogers, E.M. (2003). Diffusion of Innovations. New York: Free Press.

Stevenson, J.F. (2011, November). *Nurturing readiness for a "culture of learning" for general education*. Paper presented at the annual meeting of the American Evaluation Association, Anaheim, CA.

Stevenson, J.F., Mitchell, R.E., and Florin, P. (1996). Evaluation and self-direction in community prevention coalitions. In D.M. Fetterman, S.J. Kaftarian, and A. Wandersman (eds.) *Empowerment evaluation: Knowledge and tools for self-assessment and accountability*. Newbury Park, CA: Sage.

Stevenson, J.F., & Monteiro, K.A. (2013, October). *Benchmarking assessment climate: What does it take to promote a higher education "learning organization?"*. Paper presented at the annual meeting of the American Evaluation Association, Washington, DC.

Stevenson, J.F., Treml, M., & Paradis, T. (2009, November). *Assessing readiness for a "culture of learning."* Paper presented at the annual meeting of the American Evaluation Association, Orlando FL.

Taylor-Ritzler, T., Suarez-Balcazar, Y., Garcia-Iriarte, E., Henry, D.B., & Balcazar, F.E. (2013). Understanding and measuring evaluation capacity: A model and instrument validation study. *American Journal of Evaluation, 34*(2), 190–206.

Walser, T.M. (2015). Evaluability assessment in higher education: Supporting continuous improvement, accountability, and a culture of assessment. *Journal of Assessment and Institutional Effectiveness, 5*(1), 58–77.

Wandersman, A. (2014). Moving forward with the science and practice of evaluation capacity building (ECB): The why, how, what, and outcomes of ECB. *American Journal of Evaluation, 35*(1), 116–119.

Wehlburg, C. (1999). How to get the ball rolling: Beginning an assessment program on your campus. *AAHE Bulletin, 51*(9), 7–9.

Appendix

Table A

Assessment Climate Survey Items and Results

Instructions: Please answer each question by clicking on the appropriate response. Where you are unsure of an answer please provide your own impression. In this survey the term "assessment" is used to refer to the series of steps in defining and measureing students' learning outcomes in order to draw useful conclusions about the effectivenes of educational programs (e.g. majors) in achieving their intended outcomes and to act on those conclusions. In this context these "learning outcomes" would be defined at the program level and be measured in ways that reflect the program faculty's intentions.

Items in Domains	Mean Agreement[1]		
	2009	2012	2015
Sample size (N=)	30	35	28
I. Personal attitude toward assessment			
1. Assessment of learning outcomes for our majors is very important.	3.90	3.69	3.46
2. Assessment of learning outcomes does not yield useful results. (R)[2]			2.64
3. General education outcome objectives are complementary to our objectives for the major.	3.20	3.58	3.43
4. Assessment should be the job of the administration, not the faculty. (R)	2.77	2.60	2.32
5. Assessment of student learning outcomes is here to stay.	3.80	3.40	3.75
6. We faculty need to keep checking ourselves to improve the chances that our students graduate with the skills and attitudes we believe they need.	4.17	3.89	4.36
II. Institution-wide faculty norms			
7. Most departments here are now taking assessment seriously.	2.93	3.19	3.04
8. Most faculty on this campus believe assessment is unrelated to genuine concern for student learning. (R)	3.52	3.69	3.29

[1] Ratings are from 1 (=strongly disagree) to 5 (=strongly agree). Superscript letters (a, b, c) are used to indicate significant differences ($p<.05$ 2-tailed) between means across years.
[2] Reverse-keyed items for scoring the domain scales.

9. Many faculty resist assessment because they fear negative assessment findings that could damage individuals or programs. (R)	3.67 [a]	3.53 [a]	2.64 [b]
10. At this institution, faculty highly value transparency, including open disclosure of our students' learning outcomes.	2.70 [a]	2.69 [a]	3.29 [b]
11. The faculty at this institution are committed to the goal of having every student at the university graduate with abilities and values consistent with our university's mission and strategic plan.	3.40	3.37	3.61
12. At this institution, assessment of student learning outcomes has become a highly valued, consistently practiced, aspect of our culture.	2.33	2.17	2.29
III. Leadership commitment			
13. The administration supports assessment, from the Provost on down.	3.17	3.03	2.96
14. Our college dean/associate dean recognizes and supports the value of assessment.	4.07 [a]	3.72	3.36 [b]
15. Our college dean/associate dean discusses our departmental assessment reports with us.			2.46
16. There are no rewards or incentives for chairs or program directors participating in assessment. (R)	4.07	4.42	4.32
17. There are no incentives for faculty to participate in assessment (e.g. annual review recognition). (R)	4.00	4.50	4.07
18. There are few administration-provided resources for assessment. (R)	4.00	4.17	3.75
19. The administration keeps track of programs' assessment activities and results.	2.07 [a]	1.92 [a]	2.79 [b]
20. Adequate time is provided for those who are asked to do the work of assessment.	2.97	3.43	2.43
21. Programs that excel at assessment are formally recognized at the institution-wide level.			3.14
22. Departments that choose not to assess their programs will experience negative consequences.	3.62 [a]	3.44 [a]	2.07 [b]
IV. Infrastructure support			
23. Faculty and chairs have easily accessible opportunities to learn about how to conduct useful assessment.	2.73	2.89	3.00
24. Expectations for what is to be done and reported for program assessment are clear.	2.33	2.47	2.61
25. A clear policy for a 2-year cycle of assessment reporting is now in place.		3.17	3.36

26. There is adequate training provided for those who are asked to do the work of assessment.	2.17	2.53	2.79
27. There are models for what is expected in an assessment report.	2.79	2.86	3.29
28. The two-year reporting cycle works well for my department.		2.75	2.32
29. Departments receive useful feedback on our assessment reports.		2.94	2.61
30. There is an office on campus that provides assistance of many kinds for assessment.	3.40	3.92	3.86
31. There is a helpful website on campus addressing assessment progress and expectations.	2.93	3.25	3.50
32. There is a policy-setting committee to guide assessment on this campus.	3.10	3.58	3.36
V. Department-level implementation			
33. My department has workable assessment plan(s) for our undergraduate program(s).	4.04 [a]	3.08 [c]	3.71 [b]
34. My department has workable assessment plan(s) for our graduate degree program(s). (Please skip if not applicable for your department.)		2.54 [a]	3.57 [b]
35. Our majors are aware of our department's learning objectives.	3.33	3.09	2.71
36. My department has conducted and reported one or more rounds of assessing learning outcomes for our undergraduate major(s).	4.00	4.37	4.11
37. My department has conducted and reported one or more rounds of assessing learning outcomes for our graduate major(s). (Please skip if not applicable for your department.)			3.43
38. My department uses assessment results in strategic planning.	3.40	3.06	2.86
39. Faculty in my department have discussions about our students and our hopes for them in the context of assessment.	3.27	3.17	3.29
40. My department has changed our curriculum design (requirements, courses, course content, etc.) in response to assessment results.	3.57	3.00	2.96
41. My department has made changes in how courses are taught (pedagogy) and what is covered in them on the basis of assessment results.			3.04
VI. University-wide implementation			
42. A majority of undergraduate majors across the campus have now gone through at least one cycle of assessment to reporting to program revision (sometimes termed "closing the loop").	3.03 [a]	3.56 [b]	3.44

43. A majority of graduate majors across the campus have now gone through at least one cycle of assessment – reporting - program revision.			3.28
44. Departments share ideas with other departments/programs for meaningful, manageable assessment.			2.18
45. Strategic planning at the university level uses assessment results.	2.36	2.77	2.50
46. Learning outcomes for degree programs are aligned with the broader missions of colleges and the institution.			3.07
47. University-wide objectives for students' learning outcomes are specified, measured, and reported on a regular basis.	2.10 [a]	2.51	2.71 [b]
48. Our general education program has clear, measurable outcome objectives.	2.41	2.51	2.50
49. General education addresses important learning goals at this institution.		3.59	3.11
50. My department is willing to contribute to the assessment of general education.		2.97	3.18

51. In which stage in the development of learning outcomes assessment would you judge that this institution is?

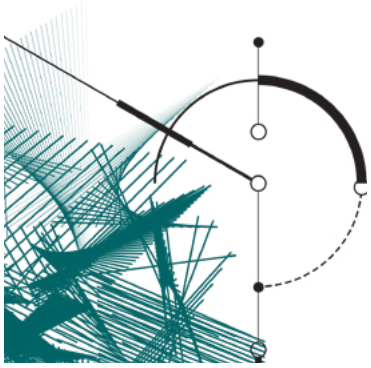
    Denial ("It's a passing fad"): 3.6%
    External Demand ("Administration says we must; we say give us time and resources or do it yourselves!"): 50.0%
    Tentative Commitment ("Leaders are committed; some of us are ready to follow"): 39.3%
    Full-scale Effort (A critical mass accept the necessity; policies and resources are in place to help): 7.1%
    Maintenance and Refinement ("We see the value and regularly use the results at all organizational levels"): 0.0%

52. This survey was previously administered to department chairs/directors in October 2009 and October 2012.
    Do you believe you took the survey at that time [either of those times]?

	Yes	Not Sure	No
2012	25.7%	17.1%	57.1%
2015	17.9%	35.7%	46.4%

Abstract

Written communication remains an important learning objective for colleges and universities as more and more students enter the workforce without necessary writing skills and experiences. This importance is increased for public colleges and universities within the state of Texas, as that state has adopted written communication as a core learning objective for its students. The efforts to assess student written communication at one four-year, public university in Texas are highlighted within this study. In particular, differences were examined in student writing performance based upon student race and gender. Using a one-way MANOVA it was determined that no statistically significant relationship existed between student writing performance and either gender or race. It is hoped that the assessment and analysis methodologies presented within this study may serve as models for other researchers seeking to evaluate written communication.



AUTHORS

Jeff Roberts, M.A.
Sam Houston State University

Carroll F. Nardone, Ph.D.
Sam Houston State University

Bill Bridges, Ph.D.
Sam Houston State University

Examining Differences in Student Writing Proficiency as a Function of Student Race and Gender

Written communication is widely considered to be an important skill for students to have mastered prior to graduating college and entering the workforce (Allan & Driscoll, 2014; Arum & Roska, 2011; Hart Research Associates, 2013, 2015b; Kelly-Riley, 2015). In a recent survey sponsored by the American Association of Colleges and Universities (AAC&U), 82% of employers indicated that it was important for graduating students to write effectively, and 81% of employers reported that they would be more likely to hire students who took multiple writing-intensive courses in college (Hart Research Associates, 2015b). These results were similar to those from a 2013 employer survey in which 80% of employers noted colleges and universities should place a greater emphasis upon written communication skills (Hart Research Associates, 2013). However, in the face of employer desire for students to have stronger written communication skills only 65% of surveyed students reported that they were well prepared with regard to written communication. Even more troubling, only 27% of the surveyed employers indicated they believed that recent graduates were entering the work force prepared to write effectively (Hart Research Associates, 2015b).

Results like these have led to some higher-education researchers holding negative perceptions regarding student writing proficiency (Arum & Roska, 2011; Secretary of Education's Commission on the Future of Higher Education, 2006). For example, the authors of the Spellings Commission report argued that graduating students lacked fundamental knowledge and skills as they graduated from college (Secretary of Education's Commission on the Future of Higher Education, 2006). These findings served as the basis for the influential and controversial book *Academically Adrift: Limited Learning on College Campuses* (Lederman, 2013; Arum & Roska, 2011). Arum and Roska (2011) further argued that colleges and universities were not doing an adequate job of preparing students with regard to several key skills, which included writing. Using data from the Collegiate Learning Assessment the authors determined that students, in general, made limited gains during

CORRESPONDENCE

Email
jeff.roberts@shsu.edu

their first two years of college (Arum & Roska, 2011). The scores of minority students lagged behind those of White students, with Black students showing virtually no gain in their scores.

Effective assessment of student writing represents an important tool colleges and universities can use to measure, and ultimately improve, student writing proficiency. However, using third-party, commercial instruments may not provide the meaningful answers institutional leaders are seeking. The measurement of written communication through the evaluation of authentic student artifacts, using locally developed processes, may instead provide institutions with a better perspective of their unique students' writing skills and proficiencies. In turn, these data can help give faculty, staff, and administrators the information they need to identify areas for improvement and to implement curricular and pedagogical changes necessary to increase the writing proficiency of students graduating from their institutions.

Literature Review

To help place this current study within a broader framework it may be of benefit to the reader to briefly examine of some of the existing literature on writing assessment and student written communication proficiencies. Anson (2010), Anson and Lyles (2011), and Behizadeh and Engelhard (2011) share many similarities, focusing on the development of "writing across the curriculum" programs throughout the recent history of higher education. The articles by Anson (2010) and Anson and Lyles (2011) were meta-analyses, examining studies pertaining to writing across the curriculum within 14 relevant journals. The authors of both studies used qualitative research techniques (e.g., citation analysis, content analysis, word count) to conduct further analysis of the articles identified from their searches (Anson, 2010; Anson & Lyles, 2011). Both Anson (2010) and Anson and Lyles (2011) examined roughly 20-year periods within their respective studies (1967 to 1986, Anson, 2010; 1986 to 2006, Anson & Lyles, 2011).

It is interesting to note that Anson and Lyles (2011) could only identify a limited number of articles focusing upon the assessment of student writing. The authors stated, "In the context of burgeoning interest in learning outcomes, assessment, and quality enhancement across all of higher education, the potential for further significant exploration of the uses of writing for assessment in other disciplines remains strong" (p. 15). This paucity of research was also recognized by Behizadeh and Engelhard (2011), who argued that the gap between writing theory and writing assessment was widening. Behizadeh and Engelhard (2011) did observe, though, that a new discipline focused on the assessment of student writing, which combined writing, composition, and measurement scholarship, seemed to be emerging within the literature.

The measurement of written communication through the evaluation of authentic student artifacts, using locally developed processes, may instead provide institutions with a better perspective of their unique students' writing skills and proficiencies.

Although the body of research on student writing is limited some studies do exist (Allan and Driscoll, 2014; Barnhisel, Stoddard, & Gorman, 2012; Cargill & Kalikoff, 2007; Desmet, Miller, Griffin, Balthazor, & Cummings, 2008; Faulkner, 2013; Good, Osborn, & Birchfield, 2012; Kelly-Riley, 2015). However, only a few of these studies provide significant discussion of assessment processes and student results (Allan & Driscoll, 2014; Desmet et al., 2008; Faulkner, 2013; Good et al., 2012; Kelly-Riley, 2015). An examination of these few studies show some of the interesting research being conducted around student writing.

Good et al. (2012) described how one university used both a locally developed writing rubric and a third-party, commercial assessment product, the Collegiate Assessment of Academic Proficiencies exam, to assess student writing. The use of multiple measures allowed the researchers to determine how well their locally developed instrument correlated with the Collegiate Assessment of Academic Proficiencies exam, to triangulate their assessment results, and to identify areas for improvement (Good et al., 2012).

Allan and Driscoll (2014) examined student written communication at Oakland University, a doctoral-research institution in Detroit, Michigan with roughly 16,000 undergraduate and 3,500 graduate students, scoring student writing artifacts from lower-level English courses with a rubric. They were then able to identify relative points of strength and weakness in student performance, gain perspectives regarding student perceptions of their own abilities, and provide faculty development opportunities (Allan & Driscoll, 2014). Finally,

Allan and Driscoll (2014) concluded that written reflections could be used alongside other in-class assignments to triangulate assessment and provide a better picture of student learning. Similarly, Desmet et al. (2008) examined students taking freshman composition courses at the University of Georgia. In particular, the authors looked to determine whether "revision improve(d) the quality of writing products" (p. 22). Their study showed that students did improve in a pre-to-post assessment.

Faulkner (2013) also conducted a university-level study of student writing, examining students at Cedarville University. Faulkner strongly advocated for both greater writing instruction and remediation across the curriculum and for implementing Writing in the Disciplines or Writing Across the Curriculum programs, arguing that one-semester remedial English programs cannot meaningfully improve student writing. Alarmingly, the results of Faulkner's study demonstrated that student writing scores actually went down from the freshman to senior years at that particular university (2013).

What is currently missing from this literature are studies examining student writing as a function of race and gender.

Finally, Kelly-Riley (2015) represents one of the more interesting studies examining student writing. Like the others, Kelly-Riley (2015) examined student writing; however, Kelly-Riley did so using a validation framework, attempting to validate the findings from a previous study of student writing (Hasswell, 2000). Kelly-Riley (2015) examined work from 30 students, from multiple points across their academic careers. Eight different domains related to student writing success were examined using a holistic rubric. The author determined that students made statistical gains across multiple domains and showed statistical improvement over time (Kelly-Riley, 2015).

What is currently missing from this literature are studies examining student writing as a function of race and gender. Race and gender can both represent at-risk factors in higher education (Gray, 2013). The influence of race (Aud, Fox, KewalRamani, 2010; Corona et al., 2017; Harper, 2012; Kim, 2011; Lucas & Paret, 2005; Strayhorn, 2010) and gender (Corona et al., 2017; Kim, 2011; Strayhorn, 2010; Voyer & Voyer, 2014) upon student success is prevalent within educational literature. However, these studies typically focus on general student success. Research examining student writing proficiency as a function of race or gender are almost nonexistent within the literature. In fact, higher-education institutions, in general, are not examining their data in this way. Acting on behalf of the AAC&U, Hart Research Associates conducted a survey with which they determined that 70% of institutional leaders reported tracking learning outcomes achievement data; however, only 16% of the responding institutions reported disaggregating data by race (Hart Research Associates, 2015a). Ultimately, student success in higher education is increasingly becoming a social justice issue (Gray, 2013); therefore, it is key for higher-education professionals to better understand how these factors may influence student writing performance.

Statement of the Problem

In the face of the challenges and concerns posed by government agencies, researchers (Arum & Roska, 2011; Secretary of Education's Commission on the Future of Higher Education, 2006), business leaders (Hart Research Associates, 2015b), and institutions must find ways to accurately assess, and help improve, student writing. These issues are particularly important for public colleges and universities in Texas, for the Texas Higher Education Coordinating Board has identified student written communication as one of the core learning objectives adopted for all public institutions within the state (Texas Higher Education Coordinating Board, 2015). However, all institutions seeking to assess student written communication, whether by state mandate or faculty choice, face similar challenges. The importance of assessing student written communication through the lens of race and gender is magnified given the importance of equity in higher education (Gray, 2013; Montenegro & Jankowski, 2017). However, there remains a significant gap in the literature in this area that needs to be addressed.

Purpose and Significance of the Study

Given the demonstrated importance of written communication for undergraduate students, and the criticisms of colleges and universities to adequately prepare students to write effectively, faculty and staff need to develop ways to assess student written communication.

The authors of this study seek to highlight the efforts of one four-year public university in southeast Texas to use a locally developed writing rubric to effectively assess the writing proficiency of students as they approached graduation. In particular, the authors attempted to determine what, if any, differences might exist in student writing scores as a function of student race and student gender. Not only does this study join the growing body of literature related to assessing student writing (Allan and Driscoll, 2014; Barnhisel et al., 2012; Cargill & Kalikoff, 2007; Desmet et al., 2008; Faulkner, 2013; Good et al., 2012; Kelly-Riley, 2015) this study also seeks to examine the important issue of equity in student achievement that many are raising in higher education (Gray, 2013; Hart Research Associates 2015a; Montenegro & Jankowski, 2017). Finally, the assessment methodologies and analysis techniques presented within this study may also serve as an example to other institutions seeking to evaluate student writing.

Research Questions

The following research questions were addressed in this study: (a) What was the difference in the student performance on an end-of-experience student writing assessment as a function of student race (i.e., White, Black, Hispanic, Other)?; and (b) What was the difference in the student performance on an end-of-experience student writing assessment as a function of student gender (i.e., male, female)?

Method

Participants

The authors of this study seek to highlight the efforts of one four-year public university in southeast Texas to use a locally developed writing rubric to effectively assess the writing proficiency of students as they approached graduation. In particular, the authors attempted to determine what, if any, differences might exist in student writing scores as a function of student race and student gender.

Student writing artifacts were selected from 4000-level, writing-enhanced courses at a four-year, public university in southeast Texas during the spring 2013 semester. A stratified, random sampling process was used in order to select authentic student writing artifacts. Several steps were taken to identify this sample pool and collect the writing artifacts. As the purpose of the original study was to examine the writing proficiencies of upper-division students, all students not classified as being juniors or seniors were excluded from the initial sample pool. This potential sample pool was then divided into separate stratum, by academic college. Students were randomly selected from within these stratum, with the total number of students selected being based upon the percentage of junior- and senior-level majors within each college for the spring 2013 semester. To identify the total number of artifacts selected from each course within the various stratum the total number of declared majors within the sample population for each college was divided by the total number of courses within that stratum. This methodology resulted in a sample pool that was representative of both the size and diversity of the studied university.

The instructors of record for each of the 203 writing-enhanced courses within the sample were then emailed requesting the selected student artifacts. All received artifacts were redacted of student and faculty identifying information in preparation for scoring, and were assigned a unique tracking code. Ultimately, 430 student artifacts from 153 writing-enhanced courses were received, of which 395 were chosen for scoring. A total of 27 submitted artifacts were unusable for the writing assessment (e.g., short-answer tests, papers written in a foreign language, illegible handwritten student work). Additionally, eight artifacts were used as anchor papers to norm faculty raters and were not included within the data for analysis.

Instrumentation

To obtain the writing scores used for data analysis in this research article the sampled student writing artifacts were scored using a locally developed writing rubric. Kuh et al. (2015) argued that "rubrics encourage the use of authentic student work for assessment" (p. 39). The rubric was separated into four different domains of student writing (i.e., Ideas/Critical Thinking/Synthesis, Style, Organization, Conventions). All artifacts were scored independently by two raters, with each rater scoring the artifact for each domain using a four-point scale. These individual domain scores were then averaged to provide an overall score for each student artifact.

Score Reliability

With any rubric-based assessments one important measure of reliability is the consistence of the scores (Banta & Palomba, 2015; Millett, Payne, Dwyer, Stickler, & Alexiou, 2008). Therefore, several steps were taken to ensure the consistency of the scoring process. An interdisciplinary group of faculty raters evaluated student artifacts over a two-day period using a locally developed rubric. At the beginning of the scoring session the group of raters were normed to the rubric using anchor papers. The entire group of raters scored identical papers and were then led through a discussion of their scores by a facilitator in order to bring everyone into agreement regarding how to appropriately apply the rubric. Twelve of these faculty members served as either a first or second rater for each artifact, with the first rater's score not being known by the second rater. When a discrepancy of two or more points was present between the average total scores for the first two raters one of two different faculty members served as a third rater. The score from the third rater was then used in place of the score that was furthest out of agreement.

Intraclass correlation coefficients (ICCs) were calculated to determine the level of interrater agreement for each of the four writing domains (i.e., Ideas/Critical Thinking/Synthesis, Style, Organization, and Conventions), the total overall score, and the overall average (Fleiss, 2003; Shrout & Fleiss, 1979). Because every rater did not evaluate every student writing artifact, a one-way random ICC was calculated. According to Cicchetti (1994), ICC agreement values below .40 demonstrate poor agreement, values from .40–.59 demonstrate fair agreement, values from .60–.74 demonstrate good agreement, and values above .75 demonstrate excellent agreement. The ICC agreement values for three of the four writing domains (i.e., Ideas/Critical Thinking/Synthesis, Style, Organization) were above a .60, indicating good agreement, while the ICC agreement value for conventions was .58, indicating fair agreement. The ICC agreement values for the total overall score and the overall average were both .80, indicating excellent agreement for the total scores (see Table 1 for a full breakdown of the ICC agreement values for this study).

The first purpose was to examine what differences might exist in student writing scores as a function of student race and student gender.

Table 1

Breakdown of ICC Agreement by Category Area

Domain Area	Intraclass Correlation for Average Measures
Ideas/Critical Thinking/Synthesis	.69
Style	.65
Organization	.64
Conventions	.58
Overall Artifact Average	.80

Results

Prior to conducting statistical procedures to address differences in student performance on an end-of-experience writing assessment as a function of student race and of student gender the normality of the dependent variables were first ascertained. The standardized skewness coefficients (i.e., the skewness value divided by its standard error) and the standardized kurtosis coefficients (i.e., the kurtosis value divided by its standard error) were all within the boundaries of normality, +/-3 (Onwuegbuzie & Daniel, 2001) for both research questions. However, the assumption for the Box's Test of Equality of Covariance was violated for both research questions. Finally, the Levene's Test of Equality of Error Variances revealed that the assumptions were met for both research questions. As the majority of the assumptions were met for both research questions, the use of a parametric, one-way Multivariate Analysis of Variance (MANOVA) was justified for this study (Field, 2009). The MANOVA procedures did not reveal a statistically significant difference

in student writing performance as a function of race (i.e., White, Black, Hispanic, Other), Wilks' Λ = .97, p = .56, or as a function of gender (i.e., male, female), Wilks' Λ = .99, p = .65 (see Table 2 for the descriptive statistics for these analyses).

Table 2

Descriptive Statistics for Student Writing Scores by Student Race and Gender

Student Demographic Characteristic	Ideas, Critical Thinking, Synthesis		Style		Organization		Conventions		Overall Student Average	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Race										
White (*n* = 259)	2.75	0.74	2.72	0.69	2.67	0.72	2.64	0.75	2.69	0.64
Black (*n* = 51)	2.61	0.81	2.56	0.72	2.55	0.66	2.41	0.73	2.53	0.65
Hispanic (*n* = 56)	2.57	0.66	2.62	0.60	2.53	0.64	2.46	0.70	2.54	0.56
Other (*n* = 28)	2.43	0.68	2.46	0.71	2.55	0.61	2.45	0.55	2.48	0.53
Gender										
Male (*n* = 143)	2.64	0.76	2.63	0.72	2.60	0.68	2.52	0.77	2.59	0.65
Female (*n* = 251)	2.71	0.72	2.69	0.67	2.64	0.71	2.60	0.71	2.66	0.61

Discussion

The purpose of this study was twofold. The first purpose was to examine what differences might exist in student writing scores as a function of student race and student gender. In doing so, this study joins a growing body of literature on the assessment of student writing (Allan and Driscoll, 2014; Barnhisel, et al., 2012; Cargill & Kalikoff, 2007; Desmet, et al., 2008; Faulkner, 2013; Good, et al., 2012; Kelly-Riley, 2015). Additionally, this study represents an example of how an institution is disaggregating student performance data (Hart Research Associates, 2015a), and is helping to answer questions regarding the equity of student learning (Montenegro & Jankowski, 2017). Finally, this study provides a model to other institutions for assessing student writing performance and analyzing those results. At first glance, the lack of statistically significant results within this study would seem disheartening. However, from an institutional perspective, these results are very important as they highlight the actual performance of that institution's students with regards to written communication.

At first glance, the lack of statistically significant results within this study would seem disheartening. However, from an institutional perspective, these results are very important as they highlight the actual performance of that institution's students with regards to written communication.

If a college or university is doing an adequate job of preparing its students to write effectively it would be natural to expect that all students, regardless of race or gender, would perform equitably upon an authentic writing assessment. Therefore, the lack of statistically significant results observed within this study could be interpreted by decision makers from that university to mean that they are preparing students equally well with regard to written communication. That said, equity does not necessarily mean quality. More information is needed to determine whether the level of student performance observed within this study was sufficient for end-of-experience students. A possible explanation for these results may also be that weaker students, regardless of race or gender, did not persist to the junior- or senior-year to be measured, thus limiting the differences observed by race or gender. It should be noted that while statistically significant differences in student scores by race and gender were not observed, White students scored higher than all other races and females scored higher than males across all four rubric domains. Further study is needed to better understand and interpret these results.

More work is also needed in order to determine whether the findings of this study are the result of some outside factors. The lack of statistical significance in the results of this study may not be representative of actual student performance, but instead may reflect error within the assessment process itself. For example, the locally developed rubric used within this study may not be sensitive enough to pick up the differences between the various student groups. There may also be flaws with the rubric itself which may be impacting the collected results. Finally, the sample size used for this study may also not have been have been sufficient to identify any differences by race or by gender.

Several steps can be taken in order to address these possible concerns. The first logical course of action would be to increase the size of the sample being used for analysis. This would allow the researchers to determine whether the results were the result of an insufficient sample size or were actually representative of student performance. It might also allow for separate statistical analysis on the racial groups included within the Other category (e.g., Asian or Pacific Islander, American Indian, International). Further replication of this study is needed to replicate and validate the results identified here (cf. Kelly-Riley, 2015).

Furthermore, as this initial study used a one-way MANOVA no attempt was made to examine the interactions between race and gender upon student written communication proficiencies. Follow-up studies are needed, with larger samples, to better understand how student performance can be affected by student membership within multiple groups. Additional variables, like socio-economic status and first-generation status, could also be included within such an analysis to better understand the nuances of student writing.

Therefore, the lack of statistically significant results observed within this study could be interpreted by decision makers from that university to mean that they are preparing students equally well with regard to written communication. That said, equity does not necessarily mean quality.

Efforts could also be made to help further validity of the rubric used to score student writing artifacts. For example, the same rubric could be used to also score writing artifacts from beginning students, the scores from which could be compared to those of end-of-experience students in order to determine whether the rubric was sensitive enough to pick up potential differences between the two groups. Also, cross-institutional scoring and comparison could offer opportunities for rubric validation. Already scored, redacted, and coded student artifacts could be traded between, and scored by, peer institutions in order to determine how student artifacts from one institution scored using the instrument from the other. Scores could then be compared using statistical analysis in order to determine how well the scores from the two rubrics correlated. This would both provide evidence for the validity of both institutions' assessment instruments, and would possibly give insight into how an institution's students were doing in comparison to peers.

As a parting warning, readers are cautioned to not overgeneralize the findings presented within this study. The examined population was limited to junior- and senior-level students attending one public, four-year Texas university, in 2013. The results from the analysis may therefore not be generalizable beyond the time, setting, and population involved within this study. Finally, although several steps were taken to try to ensure the validity and reliability of the methodologies used in this study, faculty, staff, and administrators may experience different results if they attempt to replicate the methodologies at their own institutions.

Conclusion

The data presented within this study represent only the first effort by one institution to evaluate the written communication proficiencies of its students, and the specific assessment methodologies highlighted here are not the only ways to evaluate student writing. Despite the promises of some groups to provide the magic bullet for evaluating written communication (e.g., the CLA+; Council for Aid to Education, 2015), it is impossible for any single test, measure, or rubric to provide all the information needed by institutions to improve student writing. Institutional improvement does not occur over night but instead takes the time and intentionality of faculty, staff, and administrators.

Student writing remains of great importance (Allan & Driscoll, 2014; Arum & Roska, 2011; Hart Research Associates, 2013, 2015b; Kelly-Riley, 2015), and those within higher education need to better prepare students to write effectively. In order to make the changes

that are necessary to improve student written communication, faculty, staff, and administrators must have the necessary data to make those changes. This study provides an overview of one institution's attempts to use authentic assessments to gather this needed data. In doing so, readers may be inspired to engage in their own local assessments of student writing.

References

Allan, E. G., & Driscoll, D. L. (2014). The three-fold benefit of reflective writing: Improving program assessment, student learning, and faculty professional development. *Assessing Writing, 21*, 37–55. doi:10.1016/jasw.2014.03.001

Aud, S., Fox, M., & KewalRamani, A. (2010). *Status and trends in the education of racial and ethnic groups* (NCES 2010–015). Washington, DC: U.S. Department of Education, National Center for Education Statistics, U.S. Government Printing Office. Retrieved from http://nces.ed.gov/pubs2010/2010015.pdf

Anson, C. M. (2010). The intradisciplinary influence of composition and WAC, 1967-1986. *The WAC Journal, 21*, 5–19.

Anson, C. M., & Lyles, K. (2011). The intradisciplinary influence of composition and WAC, part two: 1986-2006. *The WAC Journal, 22*, 7–19.

Arum, R., & Roska, J. (2011). *Academically adrift: Limited learning on college campuses*. Chicago, IL: University of Chicago Press.

Banta, T. W., & Palomba, C. A. (2015). *Assessment essentials: Planning, implementing, and improving assessment in higher education* (2nd ed.). San Francisco, CA: Jossey-Bass.

Barnhisel, G., Stoddard, E., & Gorman, J. (2012). *Incorporating process-based writing pedagogy into first-year learning communities: Strategies and outcomes*. The Journal of General Education, 61, 461–487. doi:10.1353/jge.2012.0041

Behizadeh, N., & Engelhard, G., Jr. (2011). Historical view of the influences of measurement and writing theories on the practices of writing assessment in the United States. *Writing Assessment, 16*, 190–211. doi:10.1353/jge.2012.0041

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago, IL: Rand McNally.

Cargill, K., & Kalikoff, B. (2007). Linked psychology and writing courses across the curriculum. *The Journal of General Education, 56*, 83–92. doi:10.1353/jge.2007.0017

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*, 284–290. doi:10.1037/1040-3590.6.4.284

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation. Design and analysis issues for field settings*. Chicago, IL: Rand McNally.

Corona, R., Rodriguez, V. M, McDonald, S. E., Velazquez, E., Rodriguez, A., & Fuentes, V. E. (2017). Associations between cultural stressors, cultural values, and Latina/o college students' mental health. *Journal of Youth & Adolescence, 46*(1), 63–77. doi:10.1007/s10964-016-0600-5

Council for Aid to Education. (2015). *CLA+ overview*. Retrieved from http://cae.org/participating-institutions/cla-overview/

Desmet, C., Miller, D. C., Griffin, J., Balthazor, R., & Cummings, R. E. (2008). Reflection, revision, and assessment in first-year composition eportfolios. *The Journal of General Education, 1*, 15–30.

Faulkner, M. (2013). Remediating remediation: From basic writing to writing across the curriculum. *The CEA Forum, 42*(2), 45–60.

Fleis, J. L. (2003). *Statistical methods for rates and proportions* (3rd ed.). New York, NY: Wiley.doi:10.1002/0471445428

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). Thousand Oaks, CA: Sage.

Gray, S. S. (2013). Framing "at risk" students: struggles at the boundaries of access to higher education. *Children & Youth Services Review, 35*, 1245–1251. doi:10.1016/j.childyouth.2013.04.011

Good, J. M., Osborne, K., & Birchfield, K. (2012). Placing data in the hands of discipline-specific decision makers: Campus-wide writing program assessment. Assessing Writing, 17, 140-149. doi:10.1016/j.asw.2012.02.003

Harper, S. R. (2012). *Black male student success in higher education: A report from the national Black male college achievement study*. Philadelphia, PA: University of Pennsylvania, Center for the Study of Race and Equity in Education.

Hart Research Associates. (2013). *It takes more than a major: Employer priorities for college learning and student success*. Washington, DC: Association of American Colleges and Universities. Retrieved from http://www.aacu.org/leap/presidentstrust/compact/2013SurveySummary

Hart Research Associates. (2015a). *Bringing equity and quality learning together: Institutional priorities for tracking and advancing underserved students' success. Key findings from a survey and in-depth interviews among administrators at AAC&U member institutions*. Washington, DC: Association of American Colleges and Universities. Retrieved from http://www.aacu.org/publications/bringing-equity-and-quality-learning-together

Hart Research Associates. (2015b). *Falling short? College learning and career success*. Washington, DC: Association of American Colleges and Universities. Retrieved from http://www.aacu.org/leap/public-opinion-research/2015-survey-falling-short

Hasswell, R. H. (2000). Documenting improvement in college writing: A longitudinal approach. *Written Communication, 17*, 307–352. doi:10.1177/0741088300017003001

Johnson, B., & Christensen, L. (2012). *Educational research: Quantitative, qualitative, and mixed approaches* (4th ed.). Los Angeles, CA: Sage.

Kelly-Riley, D. (2015). Toward a validational framework using student course papers from common undergraduate curricular requirements as viable outcomes evidence. *Assessing Writing, 23*, 60–74. doi: 10.1016/j.asw.2014.10.001

Kim, Y. M. (2011). *Minorities in higher education: Twenty-fourth Status Report* (2011 Supplement). Washington, DC: American Council on Education. Retrieved from http://diversity.ucsc.edu/resources/images/ace_report.pdf

Kuh, G. D., Ikenberry, S. O., Jankowski, N. A., Cain, T. R., Ewell, P. T., Hutchings, P., Kinzie, J. (2015). *Using evidence of student learning to improve higher education*. San Francisco, CA: Jossey-Bass.

Leaderman, D. (2013, May). Less academically adrift? *Inside Higher Education*. Retrieved from https://www.insidehighered.com/news/2013/05/20/studies-challenge-findings-academically-adrift

Lucas, S. R., & Paret, M. (2005). Law, race, and education in the United States. *The Annual Review of Law and Social Science, 1*, 203–231. doi:10.1146/annurev.ls.1.110105.100001

Millett, C. M., Payne, D. G., Dwyer, C. A., Stickler, L. M., & Alexiou, J. J. (2008). *A culture of evidence: An evidence-centered approach to accountability for student learning outcomes*. Princeton, NJ: Educational Testing Service. Retrieved from: https://www.ets.org/Media/Education_Topics/pdf/COEIII_report.pdf

Montenegro, E., & Jankowski, N. A. (2017). *Equity and assessment: Moving towards culturally responsive assessment*. (Occasional Paper #29). Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment (NILOA).

Onwuegbuzie, A. J. (2003). Expanding the framework of internal and external validity in quantitative research. *Research in Schools, 10*(1), 71–89.

Onwuegbuzie, A. J., & Daniel, L. G. (2001). Uses and misuses of the correlation coefficient. *Research in the Schools, 9*(1), 73–90.

Secretary of Education's Commission on the Future of Higher Education. (2006). *A test of leadership: Charting the future of U.S. higher education*. Washington, DC: U.S. Department of Education. Retrieved from http://www.ed.gov/

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychology Bulletin, 86*, 420–428. doi:10.1037/0033-2909.86.2.420

Smith, M. L., & Glass, G. V. (1987). *Research and evaluation in education and the social sciences*. Englewood Cliffs, NJ: Prentice Hall.

Strayhorn, T. L. (2010). When race and gender collide: Social and cultural capital's influence on the academic achievement of African American and Latino males. *The Review of Higher Education, 33*(3), 307–332. doi:10.1353/rhe.0.0147

Texas Higher Education Coordinating Board. (2015). *Elements of the Texas Core Curriculum*. Retrieved from http://www.thecb.state.tx.us/index.cfm?objectid=427FDE26-AF5D-F1A1-E6FDB62091E2A507

Voyer, D., & Voyer, S. D. (2014). Gender differences in scholastic achievement: A meta-analysis. *Psychological Bulletin, 114*, 1174–1204. doi:10.1037/a0036620

Book Review

Real-Time Student Assessment: Meeting the Imperative for Improved Time to Degree, Closing the Gap, and Assuring Student Competencies for the 21st Century Needs

Peggy L. Maki
Sterling, VA: Stylus Publishing, 2017,
214 pp. ISBN-10: 1620364883. Paperback $29.00

REVIEWED BY:
Abigail Lau, Ph.D.
Independent Assessment Consultant

In *Real-Time Student Assessment*, Peggy Maki challenges us to reframe our assessment commitment considering the needs of currently enrolled students. Maki's leadership and expertise in higher-education assessment is well known and her call for assessment practices that help current students attain their degrees is worthy of the attention of everyone involved in student-learning assessment. As George Kuh notes in the foreword for this book, her call to action is "just in time" to inform current discussions about how to gather more actionable assessment data and it is also just the beginning. The guidelines Maki provides in this book are valuable insights that will inevitably spur conversations about how student-learning assessment data can be used to bolster the achievement of currently enrolled students.

In this book, Maki introduces "real-time assessment" as a distinct approach to student outcomes assessment in which data is collected soon after matriculation and periodically during a degree program.

In this book, Maki introduces "real-time assessment" as a distinct approach to student outcomes assessment in which data is collected soon after matriculation and periodically during a degree program. The results are analyzed and interpreted immediately to enable timely actions and interventions to address the observed weaknesses of the students who were assessed. Maki contrasts real-time assessment with the more common "point-in-time" assessment practices in which student assessment of outcomes is conducted at or near program completion, observing the achievements of those who have persisted in their degree, and examining the data after those students have graduated, with the purpose of making programmatic changes to benefit future students.

Maki calls for an expanded commitment to real-time assessment, seeing the potential it has to address critical issues in higher education such as low levels of degree attainment and differential attainment rates across demographic groups. There is an urgency to her call; Maki's view is that responding *now* to demographic changes and increasing national demand for 21st century skills is imperative. Maki's call expresses a concern for equity and a desire to ensure that American higher education delivers on its promise for all admitted students. This urgent and aspirational tone is compelling, especially for readers who view Maki as an authority in the field and recognize the value of her perspective on future directions.

Maki calls for an expanded commitment to real-time assessment, seeing the potential it has to address critical issues in higher education such as low levels of degree attainment and differential attainment rates across demographic groups.

The call to action in this book is accompanied by Maki's sage advice for how to go about real-time assessment. Unlike Maki's book Assessing for Learning (2010), this book is not a step-by-step handbook with instructive guidance and resources. In contrast, this text is an expression of Maki's vision for real-time assessment. She sees that the necessary groundwork for real-time assessment of learning is in place with frameworks of learning outcomes and their associated measures already integrated at many institutions. Readers will be pleased that the hard work they have done to establish learning outcomes frameworks can serve as the foundation for real-time learning assessment. Maki suggests the Liberal Education and America's Promise (LEAP, www.aacu.org/leap) outcomes and the Degree Qualifications Profile (DQP, degreeprofile.org) as valuable resources that enable the early identification of specific achievement gaps that may prevent certain groups of students from achieving a college degree. Given these frameworks, Maki sets out guidelines for what will be needed to generate real-time student assessment data. She reviews technologies available to make real-time assessment work feasible, and suggests strategies for how institutions might begin shifting their assessment practices to benefit current students. Examples illustrating the concepts, technologies, and strategies are provided to give ideas and demonstrate possibilities.

Although the twin purposes of the book are intertwined in each chapter, the beginning chapters focus on providing evidence that supports Maki's call for a commitment to real-time assessment and the later chapters focus on providing guidance and identifying resources to enable and support real-time assessment. The first chapter sets the tone and provides the context for the rest of the book. Chapter 1 is a straight-forward presentation of data about how student demographics are changing and how degree attainment rates vary by demographic groups. Chapter 1 ends with a review of why college degrees are so important to individuals and society and reviews the national and economic demands for them. In Chapter 2, the need for real-time assessment becomes clearer as Maki presents facts showing the increasingly varied paths students take to their degrees, along with the questions this trend raises about the equitability of degree outcomes given the limitations of the credit system. She presents the outcomes-based initiatives in

higher education as movements that can validate the varied pathways students take and help ensure that each one leads to success. In Chapter 3, Maki explains the five learner-centered commitments that institutions make when setting out to adopt and integrate an outcomes framework. The commitments described here are consistent with those called for by accreditors and other advocates of current assessment practices. Readers will find these commitments familiar and be challenged to consider how these commitments, when fully integrated, form a necessary foundation for real-time assessment. Chapter 4 is where readers will find the six principles Maki has insightfully identified to support effective real-time assessment. This chapter also includes institutional examples to illustrate these principles and ends with a helpful chart comparing the who, when, and how of real-time and point-in-time assessment approaches. In Chapter 5, Maki overviews five types of academic technologies that support real-time assessment. I suspect that many readers will already be familiar with the types of technology reviewed here and will find the chapter valuable for how it clarifies the feasibility of the vision Maki has for real-time assessment. It is hard for me to imagine real-time assessment being possible without these technologies. Chapter 6 provides ideas for how to focus real-time assessment when comprehensive implementation on a large scale seems impractical. I found this last chapter most helpful because the ideas for how to narrow down assessment efforts could be helpful for anyone trying to figure out where to start in improving assessment practice at an institution.

What I love about this book is that it brings the conversation about learning assessment into context with the conversation about graduation and degree quality. This connection is much needed and drawing attention to it is valuable for the scholarship of assessment.

What I love about this book is that it brings the conversation about learning assessment into context with the conversation about graduation and degree quality. This connection is much needed and drawing attention to it is valuable for the scholarship of assessment. I am energized by the idea of institutions developing real-time assessment data because I readily see how it would create a situation in which assessment can fulfill its promise to be the powerful learning-enhancing tool it could be. I agree with Maki that the audience for the book is anyone involved in higher-education assessment, including academic technology specialists and students. I suspect that institutional leaders, administrators, and faculty leaders will be most interested in the book and will share the book to initiate and frame campus discussions about how to respond to Maki's compelling, insightful, and urgent call for real-time student assessment.

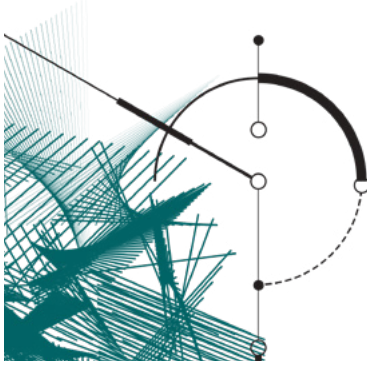
References

Maki, P. L. (2017). *Real-time student assessment: Meeting the imperative for improved time to degree, closing the opportunity gap, and assuring student competencies for 21st century needs*. Sterling, VA: Stylus Publishing.

Maki, P. L. (2010). *Assessing for learning: Building a sustainable commitment across the institution*. Sterling, VA: Stylus Publishing.

Notes in Brief

Implementation fidelity data indicate to what extent the delivered educational intervention (e.g., pedagogies, curricula) differs from the designed intervention (Gerstner & Finney, 2013; O'Donnell, 2008). Fidelity data help practitioners make more accurate inferences regarding program effectiveness (Dumas, Lynch, Laughlin, Smith, & Prinz, 2001). However, implementation fidelity research is underused in higher education (Berman & McLaughlin, 1976; Dhillon, Darrow, & Meyers, 2015). Institutional and programmatic assessment cycles typically omit implementation fidelity processes. Moreover, there are too few didactic examples of how to engage in implementation fidelity (O'Donnell, 2008). Thus, we provide actionable steps for gathering implementation fidelity data. Practitioners who adopt these steps will be well-positioned to conduct fidelity research as part of assessment processes. They will also be able to draw more valid inferences from assessment data and make more informed decisions regarding interventions. Fidelity research can help higher education evolve from an assessment culture to a learning improvement culture.



AUTHORS

Kristen L. Smith, Ph.D.
University of North Carolina-Greensboro

Sara J. Finney, Ph.D.
James Madison University

Keston H. Fulcher, Ph.D.
James Madison University

Actionable Steps for Engaging Assessment Practitioners and Faculty in Implementation Fidelity Research

Imagine you are an assessment practitioner at a university. Several faculty members from a program on your campus decide to use their assessment results to improve students' learning. Specifically, previous assessment results indicated that students' ethical reasoning skills were weak. Faculty created a new educational intervention (i.e., new curricula and pedagogies) to improve students' ethical reasoning skills. The faculty attempted to implement the new educational intervention across several courses.

With your help, faculty collect student learning outcomes data via an ethical reasoning performance assessment instrument (e.g., a rubric). These data collection procedures involve standardized, rigorous, longitudinal methodology. The assessment instrument has been studied previously and found to be psychometrically sound (i.e., adequate reliability and validity evidence exist for scores). Therefore, student learning outcomes data are expected to be trustworthy and of high quality.

You analyze and present the student learning outcomes assessment data and results to faculty members. For some classes, students' ethical reasoning skills improved dramatically over the course of the semester (i.e., from before students experienced the new intervention to after they completed the new intervention). For other classes, students' ethical reasoning skills did not improve over time. The faculty ask: "Why? Please explain to us why students in some courses experienced great change in their ethical reasoning skills whereas students in other courses did not. What do the data indicate regarding *why* this occurred?"

CORRESPONDENCE

Email
k_smith8@uncg.edu

To many assessment practitioners, these questions are all too familiar and the response "I don't know *why*" is often difficult to offer faculty who spent a great deal of time and

energy collecting student learning outcomes data, with the goal of improving their program. Both faculty and assessment practitioners often leave these meetings feeling as though there is little that can actually be inferred or acted upon with respect to program improvement; student learning outcomes data (e.g., scores on performance assessment instruments, scores on multiple-choice tests) by themselves do not appear helpful.

Unfortunately, in these situations, this feeling is completely appropriate, as there is no data collected to help faculty understand *why* outcomes assessment results differed across classes. You have no information about what actually occurred in the classrooms when the new intervention was (supposed to be) implemented. Thus, the intervention that these faculty delivered can be thought of as a "black box." Inside this black box could be the intervention as it was designed or intended or an intervention that severely deviated from what was intended. Perhaps the intervention was delivered with higher quality in some classes compared to others, or students were more responsive in some classes but not in other classes. The black box obfuscates inferences about the designed program from the student learning outcomes assessment data. However, a specific line of inquiry exists to unlock this "black box" and facilitate more accurate inferences from student learning outcomes assessment data: implementation fidelity research.

More specifically, we describe the steps a group of faculty took when engaging in implementation fidelity research. By detailing these steps we aim to promote more implementation fidelity research within higher-education contexts.

Implementation fidelity has been defined as "the degree to which a program model [educational intervention] is instituted as intended" (Dhillon, Darrow, & Meyers, 2015, p. 9). Other names for implementation fidelity include enacted curriculum, program integrity, treatment integrity, and clinical effectiveness (Dhillon, Darrow, & Meyers, 2015; Mellard, 2010). Implementation fidelity data indicate to what extent the *delivered* educational intervention (e.g., pedagogies, curricula) differs from the *designed* or *planned* educational intervention (Gerstner & Finney, 2013; O'Donnell, 2008). The five components of implementation fidelity data include:

- specific features and components of the intervention (i.e., program differentiation),
- whether each feature or component was actually implemented (i.e., adherence),
- quality with which features and components of the intervention were implemented,
- perceived student responsiveness during implementation, and
- duration of implementation.

O'Donnell (2008) and Gerstner and Finney (2013) provide more detailed definitions of implementation fidelity. Implementation fidelity data is often collected via class observations using a *fidelity checklist*, which is a behaviorally based data collection instrument, as described in greater detail by Swain, Finney, and Gerstner (2013). Although the aforementioned articles offer definitions, describe data collection tools, and emphasize the need to collect implementation data, few resources guide practitioners through the entire process of implementation assessment via an applied example.

Purpose

This article's purpose is to provide a didactic example guiding practitioners and faculty through the process of gathering implementation fidelity data. More specifically, we describe the steps a group of faculty took when engaging in implementation fidelity research. By detailing these steps we aim to promote more implementation fidelity research within higher-education contexts.

Implementation fidelity data can be gathered for virtually any educational content area at an institution of any size (Durlak & DuPre, 2008). We describe how implementation fidelity data were gathered for a campus-wide ethical reasoning initiative at James Madison University (JMU), *The Madison Collaborative: Ethical Reasoning in Action*. Because the focus of this article is implementation fidelity research, not ethical reasoning, we do not elaborate on how ethical reasoning was defined or assessed by the Madison Collaborative. Nevertheless, for readers interested in these details, we refer them to Ames et al. (2016) and Sanchez, Fulcher, Smith, Ames, and Hawk (2017).

The implementation fidelity practices described in this article contributed to large-magnitude student learning improvement across multiple courses, disciplines, and student developmental levels at JMU. That is, gains in student learning were greater in courses where the educational intervention was implemented with high fidelity compared to courses where the intervention was implemented with lower fidelity. In a forthcoming, separate article, we describe how fidelity data were integrated with student learning outcomes assessment data to facilitate and demonstrate learning improvement. Our goal is to detail the implementation fidelity process itself, showcasing how fidelity data—on their own—can be powerful for understanding the educational intervention students receive and necessary if learning improvement is the goal (Finney & Smith, 2016).

Importance of Implementation Fidelity Research

Prior to detailing the steps of gathering implementation fidelity data employed by the faculty on our campus we explain the importance of fidelity data. In brief, fidelity data are crucial for modifying educational interventions and demonstrating learning improvement (Fisher, Smith, Finney, & Pinder, 2014). Implementation fidelity data provide important information that can enhance the accuracy of the inferences made from student learning outcomes assessment data (Dumas, Lynch, Laughlin, Smith, & Prinz, 2001). Fidelity data also allow faculty to more systematically understand the educational intervention that their students actually received. Moreover, Durlak and DuPre (2008) concluded that high fidelity of implementation contributes to the success of educational programming [interventions].

More specifically, when student learning outcomes assessment data or results are unfavorable faculty are left wondering why. With implementation fidelity data practitioners and faculty are equipped to explain "why" and make informed changes to the educational intervention. That is, perhaps student learning outcomes assessment data were unfavorable because an intervention feature was not actually implemented or an intervention feature was delivered with low quality (Dhillon, Darrow, & Meyers, 2015). If so, implementation fidelity data can help faculty backward design courses (Fink, 2003), enhancing alignment between assessment, pedagogy, curriculum, and student learning. Alternatively, when student learning outcomes assessment data are favorable (e.g., students' scores improve), implementation fidelity data can "provide a roadmap for replication" and help identify "critical ingredients of program success" (Bond, Evans, Salyers, Williams, & Kim, 2000, p. 79). Understanding the effectiveness of intervention features allows faculty to be more pedagogically efficient and intentional. They can avoid "wasting" time on features of an intervention that have been shown to be ineffective for student learning improvement.

Fidelity data also allow faculty to more systematically understand the educational intervention that their students actually received.

In contrast, without implementation fidelity data, it is difficult to determine whether unfavorable assessment results are due to a poorly designed intervention or incomplete/inadequate delivery of the designed intervention (Dhillon, Darrow, & Meyers, 2015). Lack of fidelity data can lead faculty to make one of two costly errors:

- abandoning effective interventions (that perhaps were not implemented with high fidelity), or
- continuing to implement ineffective interventions (Gerstner & Finney, 2013).

In addition to these errors, lack of fidelity data can contribute to invalid inferences. For example, if student learning outcomes assessment data indicate that students' ethical reasoning skills improved from the beginning to the end of the semester faculty may (incorrectly) conclude that their new educational intervention was effective. In reality, the faculty did not implement the new intervention with high fidelity, and thus the new intervention cannot be credited with contributing to improvements in students' knowledge or abilities. Implementation fidelity provides important information that enhances the accuracy of the inferences made from student learning outcomes assessment data (Dumas, Lynch, Laughlin, Smith, & Prinz, 2001; Fisher, Smith, Finney, & Pinder, 2014). As described by Durlak and DuPre, "without data on implementation, research cannot document precisely what program [educational intervention] was conducted, or how [student learning] outcome data should be interpreted" (2008, p. 340).

For higher education practitioners who require external funding to support assessment efforts implementation fidelity data is becoming increasingly important. The U.S. Department of Education requires grant recipients to measure and report implementation fidelity to gauge educational program impact (Goodson, Price, & Darrow, 2015). In addition, public and private organizations are funding research to examine fidelity in educational contexts, develop best practices for fidelity research, and refine how fidelity is measured (Dhillon, Darrow, & Meyers, 2015; Hulleman & Cordray, 2009). Medical researchers have been measuring implementation fidelity for years (Bond et al., 2000; Rogers, Eveland, & Klepper, 1977).

Although fidelity data are imperative for assessment best practices, the collection, analysis, and integration of implementation fidelity data are completely absent from most institutional and programmatic assessment cycles.

Unfortunately, implementation fidelity continues to be underused in educational research—especially higher education (Berman & McLaughlin, 1976; Dhillon, Darrow, & Meyers, 2015). Although fidelity data are imperative for assessment best practices, the collection, analysis, and integration of implementation fidelity data are completely absent from most institutional and programmatic assessment cycles. Instead, assessment practitioners incorrectly assume that the "delivered" or "implemented" intervention is the same as the "designed" or "intended" intervention. Moreover, assessment practitioners mistakenly infer that student learning outcomes were achieved as a result of the "intended" educational intervention, not what actually occurred in the classrooms (i.e., the "delivered" intervention). This misconception is not surprising.

Method

O'Donnell (2008) notes a lack of literature guiding practitioners through the implementation fidelity process. How can assessment practitioners engage in implementation fidelity research more frequently and effectively without instructive examples of how to do so? In response to that question we detail the following steps our faculty followed to collect implementation fidelity data for an ethical reasoning educational intervention. We also highlight how fidelity data on their own (i.e., before they are integrated with student learning outcomes assessment data) can help faculty understand which features of the intervention students received (Finney & Smith, 2016).

Step 1: Allocate Adequate Time, Space, and Expertise

On our campus, six faculty from diverse disciplines and backgrounds participated in a week-long training institute related to implementation fidelity and student learning improvement.

Implementation fidelity research requires several inputs from assessment practitioners and faculty (e.g., a targeted student learning outcome, an educational intervention, a fidelity checklist, an assessment instrument or tool). Adequate time must be set aside for creation of these materials. On our campus, six faculty from diverse disciplines and backgrounds participated in a week-long training institute related to implementation fidelity and student learning improvement. The institute took place during the summer. As detailed in Appendix A, activities for the institute included:

- helping faculty members understand implementation fidelity research processes,
- helping faculty understand the assessment instrument (i.e., the rubric) used to evaluate students' ethical reasoning skills and how that was related to fidelity research,
- providing examples of fidelity research studies,
- reviewing implementation fidelity checklists and fidelity data collection processes,
- helping faculty articulate their program theory,
- allowing faculty to draw from their own experiences and learning activities to create a new learning intervention, and so forth.

During the institute, assessment practitioners used group discussions, "think. pair. share." and other activities to engage faculty. Readers are encouraged to review Appendix A for a more detailed explanation of the institute's structure and specific content. The activities included in Appendix A can be used as a template to help practitioners provide adequate time, space, and expertise to faculty as they begin engaging in fidelity research.

Faculty were compensated for their time during the institute; however, they were also intrinsically motivated to participate (e.g., faculty indicated interest in participating in the training institute before knowing a stipend would be provided). Assessment experts used group activities, peer-to-peer feedback, and other tools to promote a collaborative and safe environment for faculty. Figure 1 provides a snapshot of the processes used during the training institute to help faculty engage in implementation fidelity research.

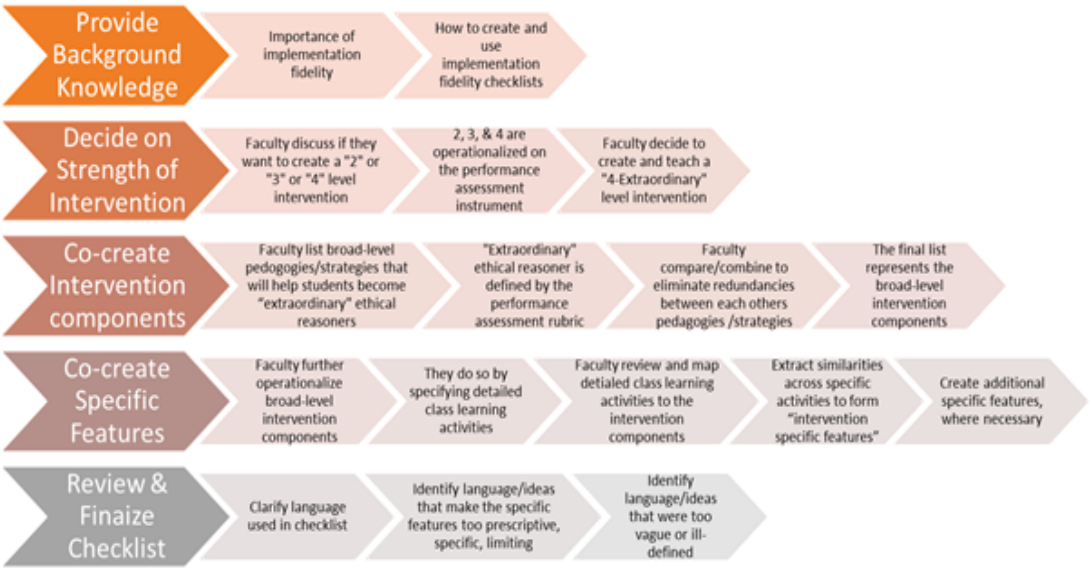


Figure 1. Visualization of Process Used During Training Institute to Help Faculty Create an Ethical Reasoning Intervention and Accompanying Fidelity Checklist

We created the following learning outcomes for the training institute. As a result of participating in the 2016 Implementation Fidelity Research Training Institute faculty will:

- Explain how assessment practice and teaching and learning are connected or related.
- Identify the five components of implementation fidelity.
- Explain the steps of collecting implementation fidelity data.
- Articulate why implementation fidelity data is important for demonstrating student learning improvement.
- Discuss and agree upon the specific features of an effective ethical intervention aligned with one of the James Madison University's Madison Collaborative ethical reasoning student learning outcomes.
- Design an ethical reasoning intervention based on the agreed upon features that aligns with the targeted Madison Collaborative ethical reasoning student learning outcome and that can be applied in various classes.
- Create a general implementation fidelity checklist aligned with the ethical reasoning intervention and the targeted university ethical reasoning student learning outcome.

These faculty learning outcomes were used to prepare and deliver institute activities and content. As shown in Appendix A, each institute activity was mapped back to at least one of the faculty learning outcomes.

In addition to the learning outcomes, the institute had two main deliverables. First, faculty were charged with detailing the specific components and features of an ethical reasoning educational intervention that they all agreed to implement within their respective classes. Second, faculty were asked to create an accompanying implementation fidelity checklist. This

checklist would later be used to capture the extent to which the designed intervention was actually delivered to students.

Faculty participants needed appropriate assessment and subject matter expertise to accomplish the learning objectives, create an educational intervention, and build a fidelity checklist. A team of two assessment practitioners, along with an ethical reasoning subject matter expert, worked closely with the faculty. At least one assessment and/or subject matter expert guided faculty through various presentations and working sessions each day of the institute[1]. The assessment practitioners were affiliated with our campus Center for Assessment and Research Studies. Our campus also has a separate Center for Faculty Innovation who provided the physical space where the summer training institute took place. The assessment practitioners worked closely with the Center for Faculty Innovation on other projects, and thus had received some cross-training in faculty development best practices.

Step 2: Facilitate Faculty Understanding of Targeted Student Learning Outcome and Assessment Instrument(s)

Before faculty could create an educational intervention, they needed to understand the learning outcome targeted for improvement. That is, effective educational interventions are intentionally created to impact particular skills or abilities. An intervention built to impact one outcome may not be effective at impacting another outcome. These ideas were discussed with the faculty.

The student learning outcome that faculty targeted for improvement concerned students' abilities to *apply their ethical reasoning skills to their personal, professional, and civic lives*. Thus, faculty needed to discuss and process what it means for students to apply their ethical reasoning skills. To facilitate this processing, an assessment expert began by explaining the importance of this outcome at the program- and university-levels.

Faculty then familiarized themselves with the rubric used to assess students' achievement of the targeted learning outcome. The rubric was a locally developed instrument designed to measure students' application of ethical reasoning skills (see Appendix B). Researchers had previously studied this rubric. As a result of those studies, reliability and validity evidence for rubric scores was provided (Smith, Pyburn, & Ames, 2016; Sanchez, Fulcher, Smith, Ames, & Hawk, 2017).

At this point, faculty understood the targeted student learning outcome they wanted to improve (i.e., ethical reasoning) and had studied the assessment instrument used to measure those skills (i.e., the ethical reasoning rubric provided in Appendix B). Now they were ready to study best practices in implementation fidelity.

An assessment expert provided copies of the ethical reasoning performance assessment rubric and reviewed all rubric criteria with the faculty. Multiple faculty members previously used the assessment rubric to rate students' ethical reasoning essays. These faculty members were asked to share their experiences using the rubric, their insights about what the rubric was measuring, and their interpretations of the rubric's criteria. The assessment expert also shared previous years' assessment results to help faculty understand the extent to which students were achieving the targeted learning outcome. Using the assessment rubric to help define and clarify the outcome promoted alignment between the assessment instrument (i.e., the rubric) and the educational intervention. The assessment rubric also provided a common language and crucial reference point for faculty who were approaching ethics education from diverse backgrounds and experiences.

Step 3: Facilitate Faculty Understanding of Implementation Fidelity

At this point, faculty understood the targeted student learning outcome they wanted to improve (i.e., ethical reasoning) and had studied the assessment instrument used to measure those skills (i.e., the ethical reasoning rubric provided in Appendix B). Now they were ready to study best practices in implementation fidelity. First, we explained the importance of implementation fidelity research, as well as identified and extensively discussed the five

[1]Note, this type of faculty development does not necessarily have to occur during a week-long institute. However, providing adequate time and space for faculty education, discussion, creation, etc. is imperative to engaging in implementation fidelity research. Moreover, consider working with external consultants if you do not have access to assessment practitioners or a subject matter expert on your campus. It is important to provide faculty with appropriate expertise as they create fidelity checklists and engage in implementation fidelity research.

components of implementation fidelity (O'Donnell, 2008; Gerstner & Finney, 2013). Faculty reviewed at least three different examples of fidelity checklists and asked questions to clarify their understanding of the five components of implementation fidelity research. As discussed previously, these five components included specific features and components of the intervention, whether each feature or component was actually implemented, the quality with which features and components of the intervention were implemented, perceived student responsiveness during implementation, and duration of implementation. We introduced faculty to five different implementation fidelity data collection methodologies and discussed how implementation fidelity data on their own (i.e., before they are integrated with student learning outcomes assessment data) are extremely useful for articulating the educational intervention students actually receive (Finney & Smith, 2016). We then provided numerous examples of how implementation fidelity data can be coupled or integrated with student learning outcomes assessment data to make more accurate inferences about student learning. This information was conveyed primarily through presentations and small group discussions.

After faculty participants understood implementation fidelity research best practices the assessment and subject matter experts guided faculty as they built a new ethical reasoning educational intervention.

Several faculty participants had no prior knowledge of implementation fidelity research. Thus, it was important to spend adequate time familiarizing them with these processes. Moreover, we reiterated that implementation fidelity data would not be used in an evaluative or punitive way (e.g., to evaluate their teaching prowess, make decisions about tenure). This frank discussion helped alleviate faculty concerns about potential uses of fidelity data while continuing to support an innocuous environment.

Step 4: Guide Faculty Through Creation of Educational Intervention and Fidelity Checklist

After faculty participants understood implementation fidelity research best practices the assessment and subject matter experts guided faculty as they built a new ethical reasoning educational intervention. Recall, the newly created intervention was constructed under the guiding framework of the ethical reasoning performance assessment rubric (see Appendix B). The assessment expert asked faculty what strength of educational intervention they wanted to create, in reference to the ethical reasoning performance assessment rubric. For instance, she asked faculty if they wanted to create an intervention that would facilitate their students being able to demonstrate "3-Excellent" ethical reasoning skills or "4-Extraordinary" ethical reasoning skills. Note, this questioning was intentionally and explicitly linked to the criteria and elements detailed on the assessment rubric. The intent was to facilitate alignment between the educational intervention, the targeted learning outcome, and the assessment instrument (i.e., the ethical reasoning rubric provided in Appendix B).

Faculty decided that they wanted to create an intervention that would help students demonstrate "4-Extraordinary" ethical reasoning skills. Thus, while faculty were creating their ethical reasoning intervention they had clear and common criteria detailed in the assessment rubric to guide them (see Appendix B). They understood they now needed to create an intervention that supported students becoming "4-Extraordinary" ethical reasoners as defined by the characteristics and skills noted in the rubric. They returned to the rubric continually as they built and rebuilt the new educational intervention.

At this stage, it was important to help faculty think through their program theory. Program theory provides a model of how a given educational intervention is expected to work (Rogers, Petrosino, Huebner, & Hacsi, 2000). Expanding on Bickman's (1987) conceptualizations, faculty members should create and articulate a program theory which details the specific aspects of their educational intervention and how that intervention is supposed to work—in theory—to enhance student learning, help students acquire a certain skillset, and more. The program theory is in reference to specific outcomes (i.e., criteria). The purpose of conducting outcomes assessment is to understand if the educational intervention—which is operationalizing a clearly articulated theory of how students should acquire certain knowledge and skills—is effective. The program theory explains *why* and/or *how* certain intervention specific features should result in students achieving certain learning outcomes.

To help articulate their program theory faculty generated a list of intervention components or "broadly-stated activities, pedagogies or approaches" that they could integrate into their classes that *should* help students become level "4-Extraordinary" ethical reasoners (See "Co-Create Intervention Components" in Figure 1). The assessment expert asked faculty

to provide rationale for *why* these components *should*—in theory—help students improve their ethical reasoning skills. Faculty referenced literature from cognitive psychology to help provide such rationale (e.g., Halpern & Hakel, 2003).

Faculty then participated in a series of "think. pair. share." exercises to co-create the components of the intervention. Intervention components are broadly specified activities, pedagogies, or curriculum, and the specific features operationalize or detail the activities under each component. Faculty compared and contrasted each other's intervention components to eliminate redundancies where appropriate.

The goal was to create an intervention that, if effective, could be easily understood and implemented by other faculty, in a variety of classes. The specific features and components of the intervention were finalized and used to create an implementation fidelity checklist.

To then operationalize these broad intervention components faculty began by sharing specific activities, assignments, demonstrations, case studies, or other learning opportunities they implemented in their classes in the past, or planned to implement in the future, to help students achieve "4-Extraordinary" ethical reasoning skills. Each faculty shared these activities with one partner, refined them, and then presented to the larger group. Then we helped faculty categorize, or map, all of the specific activities (i.e., intervention-specific features) to the intervention components that they previously articulated. The intervention-specific features were edited to be more generalizable, such that each specific feature of the intervention would be general enough to be applied across the different courses and disciplines of each faculty member.

For example, "case studies" was one of the intervention components that faculty thought would be important for helping students become "4-Extraordinary" ethical reasoners. Several faculty shared specific assignments and/or activities from their classes that would be aligned with the case studies component. As a group, faculty took these course- and discipline-specific assignments and/or activities and pulled out any underlying commonalities or similarities. These common threads became the specific features on the intervention implementation fidelity checklist (see Appendix C).

We encourage readers to review the fidelity checklist in Appendix C to understand the specific features of the intervention that faculty co-created during the summer training institute. The checklist is a vital tool for fidelity research because it details the specific features of the educational intervention, and aligns those features to student learning objectives (Swain, Finney, & Gerstner, 2013). Readers can use the checklist provided in Appendix C as a template for helping faculty articulate their own program theory and build well-aligned educational interventions. Furthermore, the checklist can serve as a template for numerous constructs or content areas of interest other than ethical reasoning.

Once the intervention components and specific features were articulated faculty critically reviewed them. They clarified language in the intervention components and specific features, identifying any instances where language/ideas were too prescriptive, specific, or limiting, as well as instances where language/ideas could be further detailed. The goal was to create an intervention that, if effective, could be easily understood and implemented by other faculty, in a variety of classes. The specific features and components of the intervention were finalized and used to create an implementation fidelity checklist, as shown in Appendix C. The fidelity checklist was general enough to be used across a wide variety of classes to collect fidelity data related to students' abilities to apply their ethical reasoning skills (i.e., the student learning outcome targeted for improvement).

Step 5: Co-create a Fidelity Data Collection Plan with Faculty

Fidelity researchers used the checklist throughout the fall 2016 semester to collect fidelity data from all six faculty participants' classes. The implementation fidelity checklist was converted into an excel worksheet, facilitating electronic gathering and storage of fidelity data. Collecting and storing fidelity data in electronic format, as opposed to paper-pencil, simplified the process of adjudicating and integrating the fidelity data with the student learning outcomes assessment data.

Note, in accordance with institutional IRB procedures, faculty participants signed an informed consent form granting consent for fidelity researchers to observe their classrooms and collect fidelity data using the fidelity checklist. During these class observations researchers

applied the checklist to at least six specified class sessions and/or specified class assignments throughout the semester. Fidelity researchers discussed and adjudicated fidelity data (e.g., came to agreement, averaged scores) to ensure that one researcher did not overlook any specific features that were implemented, or that one researcher did not rate quality of implementation too low, etc.

In addition to fidelity researchers observing class sessions to gather fidelity data, each faculty member filled in the checklist for him or herself as a self-report indication of fidelity (i.e., a "self-audit") for at least three class sessions throughout the semester. Researchers asked faculty to complete self-audits for several reasons. First, self-audits provided additional implementation fidelity data points (i.e., in addition to those collected by the fidelity researchers who observed class sessions). Additional data points promoted greater accuracy of fidelity data. Data from faculty self-audits were used in data adjudication processes described previously. Second, self-audits allowed faculty to further engage in the fidelity research process by collecting fidelity data on their own class sessions. Faculty were able to contribute their own fidelity data points to the larger pool of data points being collected by the fidelity researchers. Lastly, asking faculty to complete self-audits encouraged them to review the checklist and remain familiar with the specific intervention features they had decided to implement.

Faculty were permitted to complete the self-audit checklists during two different occasions, depending on what was most feasible for them. For example, instructors could fill out the checklist for themselves on occasions when the fidelity researchers were not able to attend the class to collect fidelity data. Thus, faculty were able to capture fidelity data points that would have otherwise been missed due to fidelity researchers not being able to observe the class session. Alternatively, faculty could fill out the checklist for themselves during class sessions where fidelity researchers were able to observe and collect fidelity data. In this instance, the faculty self-audit fidelity data provided additional data points that were adjudicated with fidelity researchers' data to enhance accuracy. Additionally, some faculty filled out the checklist via paper-pencil, whereas others filled it out electronically using an excel worksheet depending on their preference. An assessment specialist converted all paper-pencil data to electronic form.

Having faculty complete self-audits, in addition to fidelity researchers collecting fidelity data, is considered best practice (Gerstner & Finney, 2013). First, self-audit practices provide additional data points that enrich interpretation of results. When fidelity data from faculty self-audits are consistent with fidelity data collected via fidelity researchers' observations there is initial evidence of data trustworthiness. Second, self-auditing can protect against program drift by explicitly reminding faculty of the specific features they intended to include in the intervention (Gerstner & Finney, 2013). Third, engaging in self-auditing promoted faculty buy-in for implementation fidelity processes. Faculty also demonstrated greater interest in student learning outcomes assessment results, given their personal time spent collecting fidelity data via self-audits.

Collecting and storing fidelity data in electronic format, as opposed to paper-pencil, simplified the process of adjudicating and integrating the fidelity data with the student learning outcomes assessment data.

Step 6: Share and Discuss Fidelity Data with Faculty

To promote transparency all fidelity data were shared with faculty for review. After reviewing the fidelity data for each class faculty provided feedback to ensure data accuracy. For example, faculty made note of any specific features that were implemented that the fidelity researcher might have missed and commented on whether perceived student responsiveness ratings seemed accurate. The fidelity data review processes were important given fidelity data are observational. Even the best-trained fidelity researchers occasionally miss an intervention feature being implemented, misinterpret student responsiveness during class, etc.

After faculty reviewed fidelity data for accuracy a fidelity researcher synthesized and summarized all fidelity data for each faculty member across the entire semester. The fidelity researcher shared these summaries (e.g., graphs and tables of fidelity data) with the faculty, individually and as a larger group. Thus, faculty could easily evaluate the degree to which the designed educational intervention was actually implemented across their various classes. Moreover, faculty could understand what intervention features their students actually

received compared to the features that students from other classes received. Given the faculty were involved in articulating the intervention features, creating the checklist, collecting the fidelity data, etc., they understood and appreciated the quality of these fidelity data and their utility for drawing accurate conclusions about the efficacy of the new learning intervention.

The fidelity data suggested that the intervention was implemented with varying degrees of fidelity across different faculty members' classes. For example, five of the six faculty were able to implement (i.e., "adhered to") most of the specific features on the checklist (see Appendix C). However, there was some variability in student perceived responsiveness and quality of implementation. One faculty member's class had exceptionally high levels of student perceived responsiveness compared to other faculty. This same faculty member implemented the intervention with high fidelity and used the greatest variety of activities, exercises, and so forth to implement the intervention-specific features. Duration of implementation and adherence differed notably among faculty members. Some faculty members implemented the specific features with greater frequency compared to other faculty. For instance, one faculty member implemented the specific features with much greater frequency than the other faculty, but their students were perceived to be less responsive during implementation. Fidelity data also indicated that certain intervention-specific features were very rarely (or in some classes never) implemented. Perhaps faculty need further development or training to effectively implement these features, or perhaps these features are not salient to an effective educational intervention.

Overall, fidelity data demonstrated that the new educational intervention could be implemented with moderate to high fidelity across a variety of disciplines, course-types (e.g., large v. small, lecture v. community service learning, etc.), and contexts. Fidelity data on their own (i.e., before they were integrated with student learning outcomes assessment data) were powerful for understanding the educational intervention that students received (Finney & Smith, 2016). Nevertheless, the next step was to integrate fidelity data with student learning outcomes assessment data (i.e., students' scores on the performance assessment rubric provided in Appendix B) to evaluate student learning associated with the new intervention (Fisher, Smith, Finney, & Pinder, 2014). Through the process of integration, fidelity data allowed faculty to examine why students' rubric scores improved differentially across various classes over the course of the semester. This speaks directly to the hypothetical situation described at the opening of this article.

Fidelity data illuminated the black box. Faculty and researchers were able to link differential improvements in students' learning back to what students actually experienced in the classroom. Once fidelity data were integrated with the student learning outcomes assessment data (i.e., students' pre- and post-test scores on the performance assessment rubric) faculty were able to understand how certain features of the educational intervention may have contributed to students' learning improvements. In a forthcoming article we explain in greater detail how fidelity data were integrated with student learning outcomes assessment data.

Once fidelity data were integrated with the student learning outcomes assessment data (i.e., students' pre- and post-test scores on the performance assessment rubric) faculty were able to understand how certain features of the educational intervention may have contributed to students' learning improvements.

Conclusion

When differential learning gains were observed across classes, instead of asking "*why*," assessment practitioners and faculty turned to implementation fidelity data to explain the results. This was empowering, as it allowed faculty to identify which intervention-specific features were implemented with high fidelity, how students' perceived responsiveness contributed to learning gains, etc. The implementation fidelity practices described in this article also contributed to large-magnitude student learning improvement across multiple courses, disciplines, and contexts. That is, students' ethical reasoning rubric scores (see Appendix B), on average, improved two standard deviations (*Cohen's d = 2*) from the beginning to the end of the semester. According to Cohen (1988) the threshold for a large effect is 0.8. In this context, the magnitude of improvement in students' ethical reasoning abilities was exceptionally large. The new ethical reasoning intervention—articulated and studied via the implementation fidelity steps described previously—was found to be effective.

Yet, the overall effect does not tell the whole story. In a subsequent article we will explicitly describe how we integrated implementation fidelity data with student learning outcomes assessment data. This integration process allowed us to make fine-tuned inferences about which intervention features worked and which did not. Further, we were able to make recommendations regarding how the intervention could be strengthened and how faculty could deliver it more effectively in the future.

Granted, it is not enough to explain the steps needed to begin engaging in implementation fidelity research. Assessment practitioners and faculty need further guidance on how to:

- analyze fidelity data,
- integrate fidelity data with student learning outcomes assessment data, and
- present integrated data to faculty in a way that is meaningful and actionable (Coburn, Hill, & Spillane, 2016; O'Donnell, 2008).

This was beyond the scope of the current article but will be provided in forthcoming work. We encourage assessment practitioners to engage with implementation fidelity to help others understand how it can facilitate learning improvement. We know that assessment—as currently practiced—has produced few examples of learning improvement (Banta & Blaich, 2011; Blaich & Wise, 2011; Fulcher, Good, Coleman, & Smith, 2014). Perhaps we should stop griping about the under-use of learning outcomes assessment results and start investigating the efficacy of educational interventions via implementation fidelity research.

We encourage assessment practitioners to engage with implementation fidelity to help others understand how it can facilitate learning improvement.

Implementation fidelity cracks open the black box of higher-education curriculum and pedagogy. With such a link, assessment can help close the learning improvement loop. Without the connection to curriculum and pedagogy—provided by fidelity data—assessment merely perseverates a data collection loop.

References

Ames, A. J., Smith, K. L., Sanchez, E. R. H., Pyle, L. K., Ball, T. C., & Hawk, W. J. (2016). Impact and persistence of ethical reasoning education on student learning: Results from a module-based ethical reasoning educational program. *International Journal of Ethics Education*. doi:10.1007/s40889-016-0031-x.

Banta, T. W., & Blaich, C. (2011). Closing the assessment loop. *Change: The Magazine of Higher Learning, 43*(1), 22–27.

Berman, P., & McLaughlin, M. W. (1976). Implementation of educational innovation. *The Educational Forum, 40*(3), 345–370.

Bickman, L. (1987). The functions of program theory. *New directions for program evaluation, 33*, 5–18.

Blaich, C. F., & Wise, K. S. (2011, January). *From gathering to using assessment results: Lessons from the Wabash National Study* (NILOA Occasional Paper No. 8). Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment.

Bond, G. R., Evans, L., Salyers, M. P., Williams, J., & Kim, H. W. (2000). Measurement of fidelity in psychiatric rehabilitation. *Mental Health Services Research, 2*(2), 75–87.

Coburn, C. E., Hill, H. C., & Spillane, J. P. (2016). Alignment and accountability in policy design and implementation: The common core state standards and implementation research. *Educational Researcher, 45*(4), 243–251.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York, NY: Lawrence Erlbaum Associates.

Dhillon, S., Darrow, C., & Meyers, C. V. (2015). Introduction to implementation fidelity. In C. V. Meyers and W. C. Brandt (Eds.), *Implementation fidelity in education research* (pp. 8–22). New York, NY: Routledge.

Dumas, J., Lynch, A., Laughlin, J., Smith, E., & Prinz, R. (2001). Promoting intervention fidelity: Conceptual issues, methods, and preliminary results from the early alliance prevention trial. *American Journal of Preventative Medicine, 20*(1S), 38–47.

Durlak, J. A., & DuPre E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology, 41*, 327–350.

Finney, S. J., & Smith, K. L. (2016). Ignorance is not bliss: Implementation fidelity and learning improvement. *National Institute for Learning Outcomes Assessment: Guest Viewpoints*. Retrieved from: https://illinois.edu/blog/view/915/309716

Fisher, R., Smith, K. L., Finney, S. J., & Pinder, K. (2014). The importance of implementation fidelity data for evaluating program effectiveness. *About Campus, 19*, 28–32.

Fulcher, K. H., Good, M. R., Coleman, C. M., & Smith, K. L. (2014, December). A simple model for learning improvement: Weigh pig, feed pig, weigh pig. (NILOA Occasional Paper No. 23). Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment.

Gerstner, J. J., & Finney, S. J. (2013). Measuring the implementation fidelity of student affairs programs: A critical component of the outcomes assessment cycle. *Research and Practice in Assessment, 8*, 15–28.

Goodson, B., Price, C., & Darrow, C. (2015). Measuring fidelity. In C. V. Meyers and W. C. Brandt (Eds.), *Implementation fidelity in education research* (pp. 176–193). New York, NY: Routledge.

Halpern, D. F., & Hakel, M. D. (2003). Applying the science of learning. *Change Magazine, 35*(4). 36–41.

Hulleman, C. S., & Cordray, D. S. (2009). Moving from the lab to the field: The role of fidelity and achieved relative intervention strength. *Journal of Research on Educational Effectiveness, 2*(1), 88–110.

Mellard, D. (2010). *Fidelity of implementation within a response to intervention framework*. Retrieved from: http://ped.state.nm.us/ped/RtIdocs/Fidelity%20of%20Implementation%20guidev5.pdf

O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K-12 curriculum intervention research. *Review of Educational Research, 78*(1), 33–84.

Rogers, E. M., Eveland, J. D., & Klepper, C. (1977). *The Innovation Process in Organizations*. Department of Journalism, University of Michigan, Ann Arbor, MI.

Rogers, P. J., Petrosino, A., Huebner, T. A., & Hacsi, T. A. (2000). Program theory evaluation: Practice, promise, and problems. *New Directions for Evaluation, 87*, 5–13.

Sanchez, E. R. H., Fulcher, K. H., Smith, K. L., Ames, A. J., & Hawk, W. J. (2017, March-April). Defining, teaching, and assessing ethical reasoning in action. *Change: The Magazine of Higher Learning, 49*(2), 30–36.

Smith, K. L., & Pyburn, L., & Ames, A. J. (2016). *Madison Collaborative Annual Assessment Report #3*, James Madison University, Harrisonburg, VA.

Swain, M. S., Finney, S. J., & Gerstner, J. J. (2013). A practical approach to assessing implementation fidelity. *Assessment Update, 25*(1), 5–7, 13.

Appendix A

Faculty Summer Training Institute
Schedule At-a-Glance

	Activities/Curriculum	Faculty Learning Outcomes
Day 1 → Implementation Fidelity Basics **Understanding and Applying Implementation Fidelity**	• Brief introduction to the research project: explain why we are all here; the need for this work; review faculty signed MOUs and faculty responsibilities/roles in the project • Brief introduction to assessment cycle • Introduce implementation fidelity through examples from James Madison University's campus and introduce very general idea of backward design • Discuss the five components of Implementation Fidelity o Think. Pair. Share- Work with partner to fill in a blank implementation fidelity checklist for one intervention that you do in your class (can pick any intervention/activity/assignment, etc.) o What was the hardest part about creating the checklist? What components require further clarification? o Explain how implementation fidelity information can be useful pedagogically and useful for demonstrating learning improvement • Describe the typical Implementation Fidelity data collection process o The James Madison University Orientation Program o The James Madison University LID CIS project • Group discussion about the implementation fidelity matrix of possible inferences (Gerstner & Finney, 2013) o Work through four (hypothetical) examples set in an academic contexts using the fidelity matrix (Gerstner & Finney, 2013) to convey the importance of fidelity data when making inferences based on outcomes assessment data *Day 1 Wrap Up:* *Call back to why we are here: to apply implementation fidelity principles to Ethical Reasoning Instruction; to give faculty members development opportunities and skills that they can use beyond this research project. Tomorrow we will review the James Madison University ethical reasoning objective and discuss ethical reasoning educational interventions*	• Describe the steps of the assessment cycle • Explain how assessment practice and teaching and learning are connected or related • Identify the five components of implementation fidelity • Explain the steps or process of collecting implementation fidelity data • Articulate why implementation fidelity data is important for demonstrating student learning improvement • Create a "general" implementation fidelity checklist aligned with the ER intervention and JMU's ethical reasoning student learning outcome
Days 2,3, & 4 → Application of Implementation Fidelity to Ethical Reasoning (ER) Education **Creating an ethical reasoning Intervention & accompanying fidelity checklist mapped to James Madison University's ethical reasoning student learning outcome**	• Brief review of the "program differentiation" component of implementation fidelity • Brief review of the Ethical Reasoning 8 Key Questions • Review the university ethical reasoning student learning outcome targeted for improvement & the pre-existing institution-wide interventions that are mapped to each • Think. Create. Pair. Share: Individually, articulate the key features of what you believe would be a "highly effective" ethical reasoning intervention aligned with targeted university ethical reasoning student learning outcome that you could do in your classroom. Discuss in small groups and as larger group o In order for students to be able to do the targeted learning outcome, what do we need to have them practice in our classrooms? What general things or "key features" must students do in order to achieve the targeted ethical reasoning learning outcome? How can these be generalized across disciplines? How can I teach students these things or integrate these "key features" into my course? • Integrate these key features into a clear, agreed upon list of key intervention features o General "Key features" must be agreed upon by all faculty participants • Provide "blank" fidelity checklist and have faculty fill in with agreed upon key features o This will be the final checklist used for data collection	• Discuss and agree upon key components or features of an effective ethical reasoning intervention aligned with JMU's ethical reasoning student learning outcome • Based on those agreed upon components, design an ethical reasoning intervention aligned with JMU's ethical reasoning student learning outcome that can be applied in various classes • Create a "general" implementation fidelity checklist aligned with the ethical reasoning intervention and JMU's ethical reasoning student learning outcome
Day 5 -→ **Finalizing ethical reasoning intervention, checklist, & creating fidelity data collection plan**	• Faculty complete filling in fidelity checklist with agreed upon key features • Create implementation fidelity data collection procedures for Fall 2016 • Create schedule for when researchers will observe classes to collect implementation fidelity data • Discuss expectations for faculty "self-audit" using the fidelity checklist	

Appendix B

Performance Assessment Rubric Used to Rate Student Ethical Reasoning Essays

Insufficient 0	**Marginal 1**	**Good 2**	**Excellent 3**	**Extraordinary 4**
No reference to decision option(s).	Implicit reference to decision options AND/OR little context given regarding decision option(s).	Explicit but unorganized reference to decision option(s) and context.	Clear description of decision option(s) and context.	Meets criteria for *Excellent* AND… • Context treated with nuance. • Builds tension with organization and word choice.
Reference to zero or only one key question.	Vague references to key questions OR only two key questions referenced.	References four key questions.	References six key questions.	References all eight key questions.
No rationale provided for the applicability or inapplicability of any Key Questions to the ethical situation.	Provides a rationale for the applicability or inapplicability of two key questions to the ethical situation.	Provides a rationale for the applicability or inapplicability of four key questions to the ethical situation.	Provides a rationale for the applicability or inapplicability of six key questions to the ethical situation.	For all eight questions provides a rationale for its applicability or inapplicability to the ethical situation.
No attempt to analyze any of the referenced key questions.	Analysis attempted using two or more key questions. Typically incorrect ascription of the key questions to the ethical situation. Account is unclear, disorganized, or inaccurate.	Analysis attempted using three or more key questions. Basically accurate ascription of the key questions to the ethical situation. Account is unclear or disorganized.	Analysis attempted using three or more key questions. Accurate ascription of the key questions to the ethical situation. Account is clear and organized.	Meets criteria for *Excellent* AND… Nuanced treatment of key questions, for example: • elucidates subtle distinctions • uses analogies or metaphors • considers different issues within same key question
No judgment is presented OR judgment presented with no rationale.	Uses products of the analysis and provides some weighing to make a decision. Account is unclear, disorganized, or inaccurate.	Conveys weighing approach using analysis products. Provides an intelligible basis for judgment.	Meets criteria for *Good* AND…. Logically terminates in decision that will be reached.	Meets criteria for *Excellent* AND… Products of analysis weighed to make judgment compelling.

Appendix C

Ethical Reasoning Intervention Implementation Fidelity Checklist

Fidelity Researcher: _____ Date of Data Collection: _____

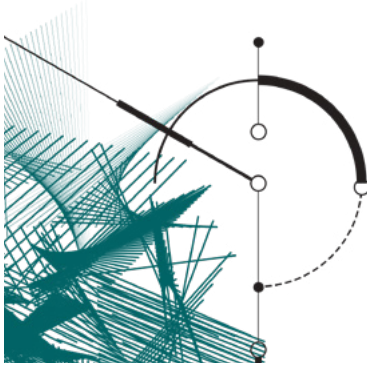
Targeted Objective for Learning Improvement	**Intervention component**	**Duration in minutes** (Actual)	**Responsiveness** 1 = Low (unengaged) 3 = Medium 5 = High (engaged)	**Specific Features**	**Adherence** Y/N	**Quality** 1 = Low (confusing) 3 = Medium 5 = High (clear)	**Comments/ Additional Observations**
Students will be able to apply their ethical reasoning skills to their personal, professional, and civic lives	Introduction/ Building Foundation to 8 Key Question (i.e., James Madison University's ethical reasoning framework)			Elaborate or unpack each of the 8 Key Questions ethical reasoning framework (e.g., reviewing the handbook, lecturing, PowerPoint slides, video clip, discussion)			
				Read/Review JMU's ethical reasoning student learning outcome			
				Read/Review rubric			
				Students experience a "check point" to check their own knowledge of the 8 Key Questions ethical reasoning framework (maybe use ethical reasoning content expert's multiple choice items??; crossword puzzle or word find; ball activity, news stories)			
				Map 8 Key Questions to some other work (can be something disciplinary like standards or something societal like policies or media or something practical, or something personal, news stories, onto class community or rules of engagement, etc.)			
				Critique/edit/comment/annotate the 8 Key Questions framework (e.g., could be wiki, could be collectively done in class, what do you like about 8 Key Questions? What would you change about them?; collective knowledge building)			
				Provide/discuss/present example of a decision-making process with AND without ethical reasoning ("ethical reasoning" is defined as being able to use 2+ Key Questions)			
Students will be able to apply their ethical reasoning skills to their personal, professional, and civic lives	Ethical Case Study			Review/Refresh 8 Key Questions ethical reasoning framework			
				Identify where/how each of the 8 Key Questions are/ are not applied within the case			
				Give/discuss rationale for how each of the 8 Key Questions are/are not applied			
				Engage in reflection (e.g., could be formal or informal, written, oral, group, what issues did you have, what was easy/hard)			
				Identify/discuss which (if any) aspects of the case are "compelling?" To what extent or degree was the case "compelling?"			

Appendix C, continued

Targeted Objective for Learning Improvement	**Intervention component**	**Duration in minutes** (Actual)	**Responsiveness** 1 = Low (unengaged) 3 = Medium 5 = High (engaged)	**Specific Features**	**Adherence** Y/N	**Quality** 1 = Low (confusing) 3 = Medium 5 = High (clear)	**Comments/ Additional Observations**
Students will be able to apply their ethical reasoning skills to their personal,	Examples			Have students together review/build a "strong" or "effective" example of ethical reasoning (e.g., show senior ethical reasoning faculty members students' videos in class and talk about what they could have done differently)			
professional, and civic lives				Identify and explain how characteristics or features make the case (in)effective referencing JMU's ethical reasoning student learning outcome and/or rubric?			
Students will be able to apply their ethical reasoning skills to their personal, professional, and civic lives	Multi-modal Analysis Visualization			Students experience (either visually or through some other sensory modality like touch, feel, movement, etc.) analysis processes- this can be "shown" by professor or created by students (e.g., block exercise, using color or size, show pre-made PowerPoint slides, students personify Key Question using their bodies as visuals, concept map- decision trees, Pictionary type game, role playing, collages, etc.)			
Students will be able to apply their ethical reasoning skills to their personal, professional, and civic lives	Analysis of 8 Key Questions and/or Analysis with 8 Key Questions			Students experience some sort of analysis (or breaking a part) of at least one Key Question; should get at nuances if possible			
				Identify obstacles or pitfalls to analysis (e.g., only analyzing 1 Key Question, confirmation bias, privilege)			
				Consider contextual factors (e.g., could include or "get at" multiple perspectives)			
				Expose/demonstrate/suggest how multiple perspectives can compete/interact w/one another within the same Key Question			
Students will be able to apply their	Weighing & Deciding using the 8 Key Questions as rationale			Students process something (debate, case, discussion, etc.) using the 8 Key Questions			
ethical reasoning skills to their personal, professional, and civic lives				Students must arrive at or grapple with a particular conclusion or decision point			
				Multiple stakeholders and/or multiple perspectives are identified or considered			

Notes in Brief

Bias is part of the human condition and becoming aware of how to avoid bias will help to ensure greater accuracy in the work of assessment. In this paper the authors discuss three different theoretical frameworks that can be applied when assessing student work for cognitive skills such as critical thinking and problem solving. Each of the frameworks highlights the importance of underlying response structure, rather than specific perspective expressed, in evaluating the quality of the response. The authors provide examples of how focusing on the structure of the response within each framework will help those assessing student work to minimize bias in their scoring and discuss how recent developments in higher education necessitate more work in this area.



AUTHORS

Pamela Steinke, Ph.D.
University of St. Francis

Peggy Fitch, Ph.D.
Central College

Minimizing Bias When Assessing Student Work

The work of faculty, assessment professionals, and scholarship of teaching and learning (SOTL) researchers often requires assessing the qualitative, open-ended work of students and in some way codifying it by outcome criteria into meaningful levels to determine how well students are meeting the outcomes. This could be part of grading for a course, doing course-embedded program assessment, or assessing student products from across departments and disciplines as occurs with scoring for general education assessment or for research purposes. Most often, some kind of rubric is used to assist with this coding or scoring of materials. The rubric could represent levels of outcome criteria that are part of a grade for an assignment, program outcomes, institutional-wide standards, or the demonstration of specific skills or beliefs. In all cases, the possibility exists that the scorer may be influenced by the perspective or point of view of the writer. The writer's perspective will further affect the content emphasized and source materials used. When this perspective does not match the perspective of the scorer and emphasizes different content and source materials than the scorer would, there exists a chance of bias resulting in scoring that is based not only on the outcome criteria being assessed but also on the perspective of the writer.

As psychologists who are involved in assessment and interested in cognitive skills and intellectual development, the authors have realized that bias has the potential to affect assessment at all levels. Often cited in this regard is what social psychologists refer to as confirmation bias, the human tendency to agree with and assess as more valid those facts and opinions that are consistent with one's own beliefs (Nickerson, 1998). Furthermore, social psychologists have demonstrated that humans have the tendency to exhibit attitudinal bias, even without awareness, for a number of distinctions including race, gender, age, and nationality (e.g., Greenwald & Krieger, 2006). Research also supports that under some conditions negative emotions can increase implicit bias (Dasgupta, DeSteno, Williams, & Hunsinger, 2009). In short, bias happens. Moreover, it is difficult to recognize, especially in oneself.

CORRESPONDENCE

Email
pamsteinke@gmail.com

We argue that the potential for bias is a concern when assessing student work and that when it does occur scorers are often not aware that the bias is operating. We present two common experiences that illustrate the point, both of which have been observed by the authors on multiple occasions, and we suspect by many readers as well. The first is a situation that occurs in grading. It is common for students, particularly first- and second-year students, to have a strong reaction to some topics that are presented in class even when they are presented in a fair and balanced manner (e.g., environmental issues, racism or sexism, religion). In these situations, some students will use an assignment to loudly voice an opinion that they perceive to be going against that of the instructor by countering a major perspective that was part of the class. When done well this approach can demonstrate critical thinking skills; even when it is not done well it is crucial that the instructor stays focused on the quality of the argument and is not biased by the student's perspective. In some cases, the student may have even misinterpreted the points being made by the instructor but the emotional tenor of the work suggests the student's perspective is deeply held. In these situations, it becomes even more important that the instructor is not exhibiting bias. However, just being aware of the need to remain unbiased does not provide instructors with tools or guidance for helping them to do so.

We argue that the potential for bias is a concern when assessing student work and that when it does occur scorers are often not aware that the bias is operating.

The second example is one that may be seen in program or institutional assessment and in research when two or more independent scorers read and score the same student sample and come together to reconcile their scores. In this situation, it is not uncommon for the scores to be very similar until the scorers encounter work on a controversial topic wherein the student's perspective is either completely consistent or completely opposite that of one of the scorers. Typically, when confronted with the extreme difference in scoring with a partner for which there is usual agreement, the discrepant scorer will then re-read the student product and recognize that the scoring was too generous or too harsh.

Importance to Current State of Higher Education

The issue addressed in this article is how to minimize bias when assessing student work for outcomes related to thinking skills (e.g., problem solving, critical thinking) which are not relevant to the student's perspective. Implementing systematic strategies to avoid bias has become even more crucial in the current climate where tension between groups with opposing viewpoints is high and "liberal lean" is being identified as a problem in higher education (e.g., Abrams, 2017). Two important changes in higher education highlight the need for more work on bias.

First, higher education serves and will continue to serve an increasingly diverse student population (e.g., Bok, 2013; McGee, 2015). As Bok (2013) notes, the current audience for higher education has expanded in the last 40 years to include a much greater variety of students including more older, low-income, and international students and more students who are working full time. McGee (2015) refers to demographic, economic, and cultural transitions that indicate in the future even more students will be first-generation, low-income, or students of color and particularly Hispanic or Latino/a. Discussing the potential for bias when assessing student work can help raise awareness among faculty scorers about the ways in which perspectives traditionally underrepresented in higher education could get discounted.

Second, higher education has recently suffered a loss of respect among some groups. For example, based on a recent study from the Pew Research Center (July 2017), the majority of Republicans view the effects of colleges and universities to be negative and part of this negativity seems to be related to differences in ideology. Focusing on potential bias could help to address this concern.

Relevant Theoretical Frameworks

In this paper, the authors demonstrate how three theoretical frameworks can help avoid bias when assessing student products for intellectual competencies such as critical thinking and problem solving: 1) Cognitive Structures in Developmental Theories, 2) Knowledge Structures, and 3) Argument Structures. Each framework provides some

specific insights into strategies for minimizing bias and the authors provide examples of how those strategies can be applied to assessment. Although this article focuses specifically on the assessment of cognitive and intellectual skills such as critical thinking and problem solving, some of the strategies discussed here could be applied to other skills as well (e.g., communication skills).

Cognitive Structures in Developmental Theories

Developmental theories that can be helpful to addressing bias in assessment include those that focus on intellectual development such as Perry (1968/1970) and others who built on his work (Baxter Magolda, 1992; King & Kitchener, 1994), moral/ethical development (Kohlberg, 1964; Rest, 1979), and development of intercultural sensitivity (Bennett, 1993). These theories share a common underlying structure comprised of stages that move from simplistic to increasingly complex ways of knowing, thinking, and perceiving.

Developmental theories. Perry's (1968/1970) scheme of intellectual and ethical development describes the evolution of college students' conceptions of the nature of knowledge and truth and how they come to reason in an increasingly complex manner. Nine positions or stages trace the student's journey from Dualism (all knowledge is known, right and wrong answers exist for everything), through Multiplicity (diversity of opinion and uncertainty with respect to knowledge become legitimate and more extensive), into Contextual Relativism (all knowledge is contextual, students perceive themselves as makers of meaning), and finally, Commitment within Relativism (Commitments, as affirmations of self, must be made in a relativistic world).

Kohlberg's (1964) theory illustrates the development of moral reasoning across six stages that are grouped in pairs to form three broad levels. Pre-conventional reasoning defines right and wrong based on obedience to authority, punishment and reward, and cooperation that benefits oneself. Conventional reasoning involves reciprocity, approval of others, and the rule of law to protect the social order. Post-conventional reasoning recognizes multiple ways of arranging a stable social order, acknowledges the existence of basic human rights, applies procedures for establishing systems of social cooperation, and appeals to abstract principles that a rational, fair-minded society would choose to govern its moral system.

Bennett (1993) extended Perry's scheme of intellectual development to describe changes in how people construe cultural difference. His developmental model of inter-cultural sensitivity includes six stages where the first three reflect ethnocentric perspectives and the last three reflect ethnorelative perspectives. In Denial people do not recognize that cultural differences even exist. In Defense others who are culturally different are categorized as "them" in contrast to "us." In Minimization superficial cultural differences are acknowledged but do not matter because all people are human. In Acceptance people are aware of their own culture as one of many and they may enjoy exploring cultural differences. In Adaptation they apply their knowledge of different cultures to shift intentionally from one frame of reference to another and modify behavior appropriately. Finally, Integration involves contextually interpreting a variety of cultural frames of reference, some of which are in conflict with each other and may not be fully reconciled.

The issue addressed in this article is how to minimize bias when assessing student work for outcomes related to thinking skills (e.g., problem solving, critical thinking) which are not relevant to the student's perspective.

Structure. All of these theories are in the cognitive developmental family and share some common assumptions, including the fundamental idea that there is an underlying structural organization to how one interprets the world and understands and solves problems. These cognitive structures function as filter systems to organize experience and thought. Structural organization leads to another assumption of this family of theories: cognitive development is a process that is content-free. That is, because development is defined as the increase in complexity of the cognitive structures used by an individual to interpret and order the outside world, then it can be conceptualized as an on-going process and not a fixed-content outcome. Therefore, what matters with respect to development is not what or how much an individual experiences but how the individual thinks about,

interprets, and orders his or her experience in qualitatively different ways. In his original publication Perry (1968/1970) emphasizes that development "takes place in the forms in which a person perceives his world rather than in the particulars of 'content' of his attitudes and concerns. The advantage in mapping development in the forms of seeing, knowing, and caring lies precisely in their transcendence over content" (p. ix).

With respect to avoiding bias when assessing student work, a strategy implied by cognitive developmental theories like Perry, Kohlberg, and Bennett would be to assess how a student's reasoning evolves from black-and-white thinking to recognizing multiple viewpoints and understanding the role of context in framing critical analysis and problem solving.

The content/process issue is addressed repeatedly throughout the literature on cognitive developmental theory. Learning is viewed as the acquisition of increasingly abstract concepts and occurs independent from the content or specific nature of the concepts involved. The stages in Kohlberg's (1964) model of moral development are based on the assumption that content and process are distinct from each other. Indeed, in his original dilemma whether a person agreed or disagreed that Heinz should steal the drug to save his wife was irrelevant; only the underlying structure of moral reasoning mattered. As Rest (1979) explains, "Each stage is described in terms of formal structures of reasoning, not in terms of the content of judgments and values generated" (p. xi). It should be noted that an underlying assumption of the cognitive developmental approach is that an increase in cognitive complexity implies more adequate and mature reasoning. For example, when confronted with an ethical dilemma, a more complex reasoner would consider such issues as the consequences of one's behavior and the effects on others while a more simplistic reasoner would primarily be concerned with simple reward and punishment.

Strategies. With respect to avoiding bias when assessing student work, a strategy implied by cognitive developmental theories like Perry, Kohlberg, and Bennett would be to assess how a student's reasoning evolves from black-and-white thinking to recognizing multiple viewpoints and understanding the role of context in framing critical analysis and problem solving. The AAC&U Problem Solving Rubric (Association of American Colleges and Universities, 2009b) reflects a similar underlying structure and acknowledgment of the role of context. Contrasting examples are illustrated in the Define Problem criterion where the lowest level reads, "Demonstrates a limited ability in identifying a problem statement or related contextual factors" versus the highest level, "Demonstrates the ability to construct a clear and insightful problem statement with evidence of all relevant contextual factors." The Problem Solving Analysis Protocol (P-SAP) poses a problem or issue that students analyze by responding to a series of questions (Steinke & Fitch, 2003). The P-SAP has been revised over the years; the most recent version can be found at http://departments.central.edu/psychology/faculty/psap/. The P-SAP can be used to assess the underlying structure of student analysis to the extent that students frame the problem and potential solutions simplistically or from a limited perspective, versus analyzing it in a more complex manner from various perspectives. For example, in response to an issue about parents being blamed for how their kids turn out, students' analyses of the problem could vary in complexity from low (example 1) to high (example 2):

> **Example 1.** Kids might think they have bad parents.
>
> **Example 2.** Peers and media often have a stronger influence in children's lives than their parents because children often spend more time with their friends and listening to music, watching television, and playing video games. School has a very strong impact on children's behavior as well because teachers and other students often treat each person differently or a classroom may be categorized as a whole and individual differences aren't recognized.

When assessing students' responses on the P-SAP, another strategy implied by two of these cognitive developmental theories is to assess how a student's analysis of the problem shifts from a focus on individual, personal factors (as in Kohlberg's Pre-conventional reasoning or Bennett's Ethnocentrism) to include broader systemic factors (Kohlberg's Conventional reasoning) and finally to integrated individual and systemic factors (Kohlberg's Post-conventional reasoning and Bennett's Ethnorelativism). Three examples below illustrate these differences in the underlying cognitive structure of students' interpretations and analyses; students responded to a P-SAP prompt asking for potential solutions to the problem of reliance on standardized tests as the most important measure of student success.

Example 1. People could look more at the student's performance throughout the year.

Example 2. Research needs to be done in order to find out the best way to measure success.

Example 3. Schools need standardized tests that accommodate all learning styles, a variety of interests, and a variety of testing styles. Plus teaching and learning occur at local levels and they do vary city to city, state to state. Standardized tests need to take into account specific emphasis schools and teachers place on certain subjects and create local testing that matches local teaching and then set up a national guideline of materials to be covered.

The first example posits a solution based solely on the individual student. The second implies that there is a best method for measuring success and proposes research as the way to discover it, a solution focused entirely on the system. The third example integrates both individual (learning styles, interests) and systemic factors (local variations in teaching by school, city, and state; national guidelines for materials) when addressing solutions. Using the framework of developmental theories, the latter response is a more cognitively complex analysis of solutions to the problem posed.

Knowledge Structures

The second theoretical framework that provides direction in coding was developed by cognitive scientists to describe how knowledge is organized and processed (e.g., Graesser & Clark, 1985; Schank, 1986; Schank & Abelson, 1977). Classic work by Schank and Abelson (1977) identified the importance of knowledge structures in the form of scripts to human understanding and planning. This work also drew attention to the important role of goals in comprehension and the need to identify different types of goals. Graesser & Clark's Generic Knowledge Structure (GKS) approach was developed to further explain text comprehension including the causal and superordinate goal inferences used to provide coherence to a text. For our purposes, an important aspect of this family of theories is how they are used to identify types of knowledge, relationships between nodes of knowledge/meaning units, and inferences made in order to connect knowledge. This focus on the abstract knowledge structure rather than the knowledge content is what makes the application of this theoretical framework useful to minimizing bias when assessing student work. For example, Graesser & Clark identify four different types of knowledge nodes (i.e., state, event, goal, and style) with arcs representing the structural relationships between nodes (i.e., consequence, reason, outcome, initiate). These structures contribute to response coherence.

The P-SAP can be used to assess the underlying structure of student analysis to the extent that students frame the problem and potential solutions simplistically or from a limited perspective, versus analyzing it in a more complex manner from various perspectives.

Strategies. With respect to avoiding bias when assessing student work, a strategy implied by the knowledge structure approach is to focus on the structural coherence of the student's explanatory response. Schank (1986) suggests explanations are types of knowledge structures enacted when a pre-existing knowledge structure is not available. In the search to find a relevant knowledge structure that might work for the explanation, the respondent calls up relevant knowledge structures and puts them together in a coherent pattern to provide an explanation. Paying attention to types of knowledge and types of relationships between knowledge nodes allows the scorer to focus on the coherence of the knowledge structure itself. When applied to coding students' responses, scorers can focus on the coherence of student knowledge nodes connected with arcs and held together with inferences into a logical causal or goal structure. For example, one of the descriptors for scoring a response on the highest level of complexity in the P-SAP rubric is, "at least two different factors explained/elaborated and situated in context with causal connections either between or within the factors." Similarly, the highest level of the last criterion in the AAC&U Critical Thinking Rubric (Association of American Colleges and Universities, 2009a) is, "Conclusions and related outcomes (consequences and implications) are logical and reflect student's informed evaluation and ability to place evidence and perspectives discussed in priority order." These descriptors highlight the connections between nodes of knowledge or the need to develop a coherent knowledge structure.

Four different example responses from the same P-SAP prompt illustrate differences in complexity of causal structures and coherence of explanations. All respondents are being asked to explain the cause of parents being blamed for how their kids turn out.

> **Example 1.** Other people in communities tend to cause this problem. When students misbehave in school, act out in the community, etc. parents get the blame and get looked down on. People think obviously the parents must have done something wrong.
>
> **Example 2**. Most people believe that the parents have the greatest effect on children, but while they do have a big role, they are not the only role in the development of that child. While the parents may have some influence, they may not be the whole problem with how the child "turns out."
>
> **Example 3**. I think society instantaneously blames the parents and dismisses themselves or peers because "origin" and background is a huge means of defining status and character/personality, thus we look to this first to blame. It only makes sense, at first, to think of the effects of parents.
>
> **Example 4.** Parents have become the targets of blame for the way children turn out because it is easy to blame parents. Parents become a scapegoat because no one else wants to be at fault and throughout history people have always seen parents as being responsible for their child's behavior. It is simple to blame parents and it is complex to blame a number of factors, so parents usually get blamed.

Sometimes it is not a matter of the student citing what "they" said but acknowledging that the other view will have something to say that must be considered. For example, in the P-SAP Locus rubric, greater elaboration of a single perspective, including a recognition of the need to gather more information, is an indicator that the individual is moving higher up on the scale toward the ability to clearly articulate an elaboration of two different perspectives.

While each response addresses the prompt, in the first two examples the statements that make up the responses are not connected causally to form a coherent explanation. In the third and fourth examples, however, the responses include clear causal connections between different propositional content. The coherence of the explanation can be seen in the pattern that emerges from the successful integration of different nodes of knowledge (i.e., origin as indication of character leads to looking at parents first; ease of identifying parental role throughout history leads to avoiding complex answers).

Argument Structures

The last theoretical framework that provides direction is one that was not developed by psychologists to capture intellectual development or knowledge structures but rather one that was developed by English professors Graff and Birkenstein (2014) to identify argument structures that help students enter the world of academic discourse through their writing. We include the "they say / I say" framework (along with its associated templates) because, as the authors assert, it "represents the deep, underlying structure, the internal DNA as it were, of all effective argument" (p. xix). From a psychologist's viewpoint, the authors are claiming that the template reflects an internal cognitive structure for effective argumentation that could easily be identified as a component of critical thinking, much like causal knowledge structures discussed previously in this paper.

Strategies. With respect to avoiding bias when assessing student work, a strategy implied by this framework is to take out the content altogether and determine whether the structure of an argument exists; if so, then evaluate the quality and complexity of the argument structure itself. Although there was no original connection to the work of Graff and Birkenstein (2014), the development of the P-SAP protocol and rubric reflects this same framework. As noted previously, the rubric reflects the importance of students recognizing both systemic and individual aspects of a problem at the highest levels of complexity. This recognition is often revealed through a dialogue in which the student accepts parts of some views but not all, a version of the "they say / I say" template. An example demonstrates how the P-SAP encourages this dialogue in a response to a question about the solution to increased reliance on standardized tests in education: "Do away with tests all together, *that is what some people may think*. *I think that* standardized tests are important, but not what a child's educational standing should be solely based on." Awareness of the structure of the argument, independent of content, will help to ensure that scoring is not affected by

a scorer's agreement or disagreement with the content. The above has a clear "they say / I say" framework in the italicized portions, independent of content. In fact, the content could be switched and it would have the same level of cognitive complexity as in the following example: "Standardized tests are important, but should not be the sole basis of a child's educational standing *is what some people think*. *I think* that we should do away with standardized tests altogether."

Sometimes it is not a matter of the student citing what "they" said but acknowledging that the other view will have something to say that must be considered. For example, in the P-SAP Locus rubric, greater elaboration of a single perspective, including a recognition of the need to gather more information, is an indicator that the individual is moving higher up on the scale toward the ability to clearly articulate an elaboration of two different perspectives. When "they say" is acknowledged separate from "I say" the scorer should be looking for either an understanding of the importance of getting the "they say" right or a demonstration by the author of the ability to fully articulate the other view. As Graff and Birkenstein (2014) argue, "When a writer fails to provide enough summary or to engage in a rigorous or serious enough summary, he or she often falls prey to what we call 'the closest cliché syndrome,' in which what gets summarized is not the view the author in question has actually expressed but a familiar cliché that the writer *mistakes* for the author's view" (p. 33). In combination with the strategies discussed earlier, becoming more aware of the structure and quality of the argument, regardless of content, will help to minimize bias when assessing student work.

Faculty awareness of the importance of structure may be heightened by incorporating discussion of these frameworks into faculty development at the program or institutional level, especially prior to the scoring of student work.

Conclusion and Implications for Future Work

In this paper we have demonstrated how three different theoretical frameworks can be applied to the assessment of student work to help minimize bias. The frameworks are not meant to be exhaustive and much more could be done to demonstrate how each of the three presented here can be applied to assessing student work. The intent of this paper is to increase awareness of how a focus on structure can help to minimize bias. In doing so, the authors are not arguing that content is unimportant. To the contrary, content is crucial for evaluating the coherence of the structure. However, within a coherent structure, the perspective and resultant content of the respondent may not be relevant to the scoring of student work when evaluating thinking skills such as problem solving and critical thinking.

Student awareness of the importance of structure may be heightened by sharing these theoretical frameworks and the strategies implied by them with students at the course level. Sharing the frameworks would also help students develop cognitive skills for critical thinking and problem solving. Indeed, Graff and Birkenstein (2014) explicitly recommend teaching students how to use the "they say / I say" template as a strategy for helping them learn how to develop effective arguments. With respect to Perry's (1968/1970) scheme, the value for sharing this framework with students comes less from teaching them about the developmental stages than from helping them learn how to use questions that prompt growth from one stage to the next and challenge them to think and problem solve in more complex ways. Questions such as, "Are there other ways to define this problem?" could challenge a student in Dualism toward Multiplicity, or "What evidence would support your analysis of this problem?" or "What strengths and limitations does your proposed solution have that might not apply everywhere?" could prompt the shift from Multiplicity into Contextual Relativism. Doing so has the potential to develop students' critical thinking and problem-solving skills regardless of their specific perspective on the issue.

Faculty awareness of the importance of structure may be heightened by incorporating discussion of these frameworks into faculty development at the program or institutional level, especially prior to the scoring of student work. Incorporating these frameworks into training sessions to prepare for scoring can also have the added benefit of increasing inter-rater reliability. Moreover, Perry's scheme has implications for the design of courses, pedagogy, curriculum, and assessment (Knefelkamp, 1974; Moore, 2000). Training faculty about these stages of intellectual and ethical development can raise their awareness

of the underlying structure of students' reasoning about complex issues and, ideally, help them avoid getting distracted by the specific perspective expressed that might differ from their own.

Beyond these specific insights, there are some broader implications for increasing awareness of bias when assessing student work and reasons for furthering research in this area. One of the major challenges for higher education is how to welcome all voices and all perspectives whether or not they are expressed in the traditional language of the academy. This point was highlighted in the landmark work on Women's Ways of Knowing (Belenky, Clinchy, Goldberger, & Tarule, 1986) that increased awareness of how gender can influence intellectual development. Ross (2016) argues that communication mismatch theory helps to explain why so many "new majority" college students (e.g., low income, immigrant, first-generation) do not complete college or do not perform to their full potential. This theory states that how a person uses spoken and written language, as well as the attitudes and body language exhibited, will unconsciously be interpreted differently when experienced by someone who is not of the same background or culture as the communicator, and that this misinterpretation may have unintended consequences. According to Ross, the resultant misunderstanding or miscommunication is often never consciously acknowledged or analyzed but can have a major impact on how well higher education supports new majority students. Our claim is that it can also impact how biased we are when assessing student work. Encouraging faculty to acknowledge the bias inherent in any perspective and to actively find ways to maintain high academic standards while countering that bias may encourage more diverse thinking in higher education to the benefit of all.

References

Association of American Colleges and Universities. (2009a). *Critical thinking VALUE rubric*. Retrieved from http://www.aacu.org/value/rubrics/critical-thinking

Association of American Colleges and Universities. (2009b). *Problem solving VALUE rubric*. Retrieved from http://www.aacu.org/value/rubrics/problem-solving

Abrams, S.J. (2017, March 5). Why colleges' liberal lean is a problem. *The Chronicle of Higher Education*. Retrieved from http://www.chronicle.com/article/Why-Colleges-Liberal-Lean/239355

Baxter Magolda, M. B. (1992). *Knowing and reasoning in college: Gender-related patterns in students' intellectual development*. San Francisco: Jossey-Bass.

Bennett, M. J. (1993). Towards ethnorelativism: A developmental model of intercultural sensitivity. In R. M. Paige (Ed.), *Education for the intercultural experience* (2nd ed., pp. 21–71). Yarmouth, ME: Intercultural Press.

Belenky, M. F., Clinchy, B.M., Goldberger, N.R., & Tarule, J.M. (1986). *Women's ways of knowing: The development of self, voice, and mind*. New York, NY: Basic Books.

Bok, D. (2013). *Higher education in America*. Princeton, NJ: Princeton University Press.

Dasgupta, N., DeSteno, D., Williams, L.A., & Hunsinger, M. (2009). Fanning the flames of prejudice: The influence of specific incidental emotions on implicit prejudice. *Emotion, 9*(4), 585–591.

Graesser, A.C., & Clark, L.F. (1985). *Structures and procedures of implicit knowledge*. Norwood, NJ: Ablex.

Graff, G., & Birkenstein, C. (2014). *They say / I say: The moves that matter in academic writing* (3rd ed.). New York, NY: W.W. Norton & Company.

Greenwald, A.G., & Krieger, L.H. (2006). Implicit bias: Scientific foundations. *California Law Review, 94*(4), 945–967.

King, P. M., & Kitchener, K. S. (1994). *Developing reflective judgment: Understanding and promoting intellectual growth and critical thinking in adolescents and adults*. San Francisco: Jossey-Bass.

Knefelkamp, L. L. (1974). *Developmental instruction: Fostering intellectual and personal growth of college students*. (Unpublished doctoral dissertation). University of Minnesota: Minneapolis.

Kohlberg, L. (1964). Development of moral character and moral ideology. In M. Hoffman (Ed.), *Review of child development research* (Vol. 1, pp. 381–431). New York, NY: Russell Sage Foundation.

McGee, J. (2015). *Breakpoint: The changing marketplace for higher education*. Baltimore, MD: Johns Hopkins University Press.

Moore, W. S. (2000). Understanding learning in a postmodern world: Reconsidering the Perry scheme of intellectual and ethical development. In B. K. Hofer & P. R. Pintrich (Eds.), *Personal epistemology: The psychology of beliefs about knowledge and learning* (pp. 17–35). Hillsdale, NJ: Lawrence Erlbaum Associates.

Nickerson, R.S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*(2), 175–220.

Perry, W. G. (1999). *Forms of intellectual and ethical development in the college years: A scheme*. San Francisco: Jossey-Bass. (Original work published 1968/1970)

Pew Research Center (2017, July). *Sharp partisan divisions in views of national institutions*. Retrieved from http://www.people-press.org/2017/07/10/sharp-partisan-divisions-in-views-of-national-institutions/

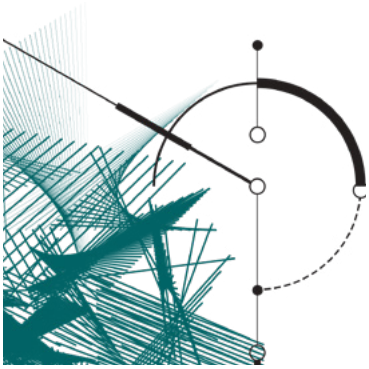
Rest, J. R. (1979). *Development in judging moral issues*. Minneapolis, MN: University of Minnesota Press.

Ross, K.A. (2016). *Breakthrough strategies: Classroom-based practices to support new majority college students*. Cambridge, MA: Harvard University Press.

Schank, R.C. (1986). *Explanation patterns: Understanding mechanically and creatively*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Schank, R.C. & Abelson, R.C. (1977). *Scripts, plans, goals and understanding: An inquiry into human knowledge structures*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Steinke, P., & Fitch, P. (2003). Using written protocols to measure service-learning outcomes. In S.H. Billig & J. Eyler (Eds.), *Advances in service-learning research. Deconstructing service-learning: Research exploring context, participation, and impacts* (Vol. 3, pp. 171–194). Greenwich, CT: Information Age.



Notes in Brief

The aims of this study were to determine faculty's ability to accurately and reliably categorize exam questions using Bloom's Taxonomy, and if modified versions would improve the accuracy and reliability. Faculty experience and affiliation with a health sciences discipline were also considered. Faculty at one university were asked to categorize 30 sample exam questions using either Bloom's Taxonomy or one of two modified versions of Bloom's Taxonomy. Overall accuracy improved when a modified version of Bloom's Taxonomy was used. Collapsing the six categories of Bloom's into three (knowledge; comprehension and application; analysis, synthesis, and evaluation) showed higher levels of accuracy than when each category was collapsed with its neighbor. There was no difference between health science and nonhealth science faculty in accuracy. Overall interrater reliability was low regardless of experience or health science affiliation.

AUTHORS

Adam C. Welch, Pharm.D.
East Tennessee State University

Samuel C. Karpen, Ph.D.
East Tennessee State University

L. Brian Cross, Pharm.D.
East Tennessee State University

Brandie N. LeBlanc, Pharm.D.
Vanderbilt University Medical Center

A Multidisciplinary Assessment of Faculty Accuracy And Reliability with Bloom's Taxonomy

Published in 1956, Bloom's Taxonomy is a hierarchy of six categories (knowledge, comprehension, application, analysis, synthesis, and evaluation) that can be used to classify the depth of students' learning (Bloom, 1956; Krathwohl, 2002). While the taxonomy has maintained its hierarchal structure it has undergone several revisions and extensions including application in the affective and psychomotor domains, and, in 2002, a revision to better align the levels with their intended outcomes (Krathwohl, 2002). Due to its ability to classify the depth of learning, Bloom's Taxonomy can be applied when creating learning objectives for a course or when creating assessments but only to the extent of the accuracy and reliability of faculty use (Adams, 2015). Several findings call into question faculty ability to apply Bloom's Taxonomy. For example, faculty sometimes misalign their course objectives with the difficulty of exam questions. Misalignment commonly occurs when expectations of learning are at a higher level than the assessment questions that are written (i.e. the test is too easy; Momsen, Long, Wyse, & Ebert-May, 2010; Jideani & Jideani, 2012). Furthermore, faculty may not be formally trained in educational pedagogy (George, 2016; Engle et al., 2014).

Several colleges, especially in the health sciences, use an electronic platform to create, deliver, and assess exam questions. (ExamSoft For Your Program, 2017). This electronic testing platform allows for individual exam questions to be tagged to a particular outcome. Bloom's Taxonomy serves as one of those outcomes on this platform. By tagging questions to Bloom's faculty can identify potential areas of student weakness and consider curriculum changes if needed (Terry, 2016). However, this form of assessment is only effective if the faculty member can appropriately distinguish between the levels of Bloom's Taxonomy.

This study aimed to determine the accuracy and reliability of faculty's ability to use Bloom's Taxonomy to categorize sample exam questions. A secondary aim was to determine if other factors would have an effect on accuracy and reliability, such as having experience

CORRESPONDENCE

Email
welcha1@etsu.edu

with Bloom's Taxonomy or using a modified version of Bloom's. In fact, several collapsed versions of Bloom's Taxonomy have been described in the literature (Cecilio-Fernandes, Kerdijk, Jaarsma, & Tio, 2016; Igbaria, 2013; Kibble & Johnson, 2011; Phillips, Smith, & Straus, 2013). This study was completed at East Tennessee State University (hereby referred to as institution), an R3 doctoral university located in the southeast (Carnegie, 2017). Since this institution contains several health sciences colleges, another secondary aim was to assess whether being a member of a health sciences discipline would have an effect on accuracy and reliability.

Methods

This study aimed to determine the accuracy and reliability of faculty's ability to use Bloom's Taxonomy to categorize sample exam questions. A secondary aim was to determine if other factors would have an effect on accuracy and reliability, such as having experience with Bloom's Taxonomy or using a modified version of Bloom's.

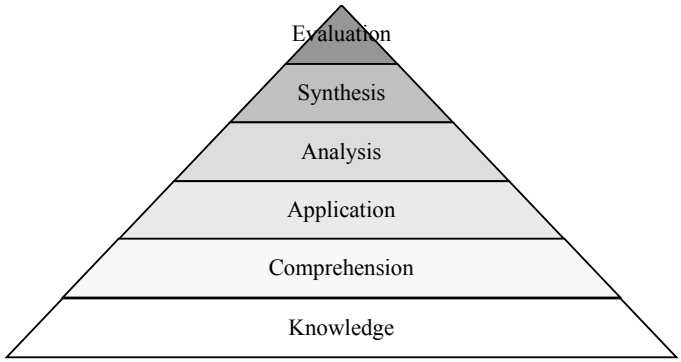
Three versions of the thirty-minute online survey were developed in Formstack (Formstack–Indianapolis, IN) and fielded to all faculty at the institution. This study was approved by the institutional review board. There were 1,202 faculty included in the sample. Academic Health Sciences Center (AHSC) faculty included nursing, public health, physical therapy, medicine, and clinical/rehabilitative health sciences. Non-AHSC participants were faculty from colleges of education, arts and sciences, business and technology, graduate and continued studies, and the honors college. The surveys were fielded for two weeks and email reminders were sent every four days.

Each version of the survey required participants to categorize 30 sample exam questions according to Bloom's 1956 Taxonomy. This taxonomy, rather than the 2002 revision, was used because it is included in the exam management software (Vandre & Ermie, 2017) used by several colleges at institution. The exam questions, which were written to be clear examples of each level of Bloom's Taxonomy, were taken from the teacher resources section of the University of California at Berkley's Center of Teaching and Learning Web site (University of California–Berkeley, 2015). See Table 1 for a subset of the questions used in the survey and their corresponding Bloom's levels. Participants were given one of three versions of the survey. The first version of the survey required participants to categorize each sample exam question to one of the original six levels of Bloom's Taxonomy (knowledge, comprehension, application, analysis, synthesis, and evaluation), hereafter known as Original. The second and third versions collapsed the categories by combining them into three categories. The three versions were called Original, Collapse One, and Collapse Two. Figure 1 outlines the original and modifications to Bloom's Taxonomy that were used in this study.

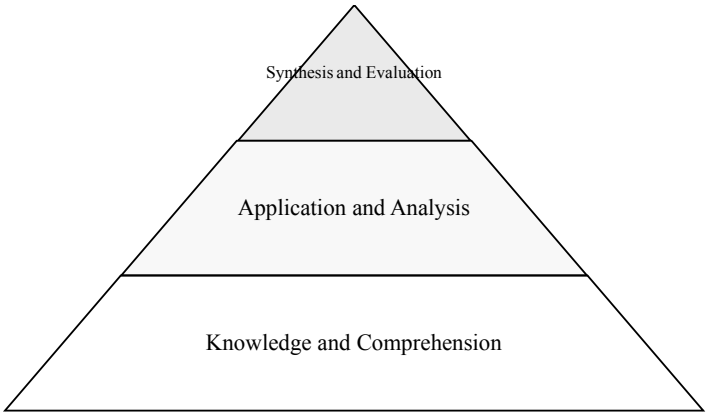
In all versions participants indicated whether or not they believed that they had categorized each item correctly. The Collapse Two version was based on Karpen and Welch (2016) who found that faculty tended to categorize knowledge items accurately but tended to confuse comprehension with application and analysis, synthesis, and evaluation with one another. The Collapse One version was based on other researchers who had merged each level of Bloom's with its neighbor (Plack et al., 2007; Gonzalez-Cabezas, Anderson, Wright, & Fontana, 2015). We sought to determine which collapsed scheme produced more accurate and reliable responses. Before categorizing the 30 items participants were provided with a brief explanation of Bloom's Taxonomy that included a description of each level and a corresponding example.

After categorizing the 30 sample exam questions participants estimated the number of items that they categorized correctly, reported their primary department affiliation, and reported how frequently they used Bloom's Taxonomy on a six-point scale (1=*Never* to 6=*At least once per week*; see Tables 2 and 3 for a demographic description of the sample). Krippendorff's alpha was used to determine the interrater reliability of the participant's classification. For this, greater than .600 is considered substantial and greater than .800 is considered almost perfect (Landis & Koch, 1977; Krippendorff, 1970).

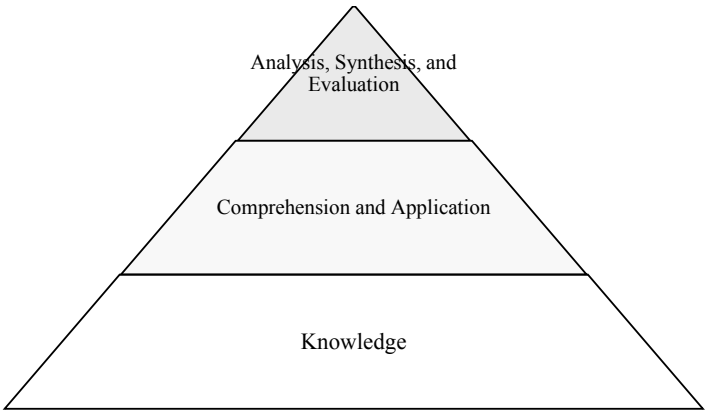
Figure 1. Original and collapsed versions of Bloom's Taxonomy used in this study.



Original—Bloom's Taxonomy. Of note, more recent versions of the taxonomy use the following terms in order: remember, understand, apply, analyze, evaluate, and create (Krathwohl, 2002).



Collapse One—Each category collapsed with its neighbor, thus creating three categories.



Collapse Two—A unique collapsing of the original categories into three categories.

Table 1

Example questions used in the three surveys

	Original	*Collapse One*	*Collapse Two*
Define short term memory	Knowledge	Knowl/Comp	Knowledge
What are the five sections of a research report?	Knowledge	Knowl/Comp	Knowledge
In one sentence give the point of a written passage.	Comprehension	Knowl/Comp	Comp/App
Describe in prose what is shown in graph form.	Comprehension	Knowl/Comp	Comp/App
Apply shading to produce depth in a drawing.	Application	App/Analysis	Comp/App
Determine the volume of an irregularly shaped object.	Application	App/Analysis	Comp/App
Given an argument for the abolition of guns, enumerate the positive and negative points presented.	Analysis	App/Analysis	Analysis/Synth /Eval
Identify the assumptions underlying a geometric proof.	Analysis	App/Analysis	Analysis/Synth /Eval
Write a logically organized argument in favor of a given position.	Synthesis	Synth/Eval	Analysis/Synth /Eval
Given two opposing theories, design an experiment to compare them.	Synthesis	Synth/Eval	Analysis/Synth /Eval
Given an argument for any position, enumerate the logical fallacies in that argument.	Evaluation	Synth/Eval	Analysis/Synth /Eval
In a given clinical situation, determine best treatment and predict the main effects and possible side effects.	Evaluation	Synth/Eval	Analysis/Synth /Eval

Results

There were 131 participants responding to the survey (10.9%) with 56 affiliated with colleges in the AHSC. Participants were given one of three versions of the survey: Original survey had 46 participants, Collapse One had 42, and Collapse Two had 42 participants. One participant was dropped due to incomplete information.

Interrater reliability

Interrater reliability for the Original version was α=.308 [95% CI (.341–.419)]. The Collapse One reliability was α=.423 [95% CI (.324–.515)] and Collapse Two was α=.426 [95% CI (.328–.524)]. Within each version Krippendorff's alpha was determined for frequent Bloom's Taxonomy users (participants who use Bloom's Taxonomy at least several times per semester), novices (participants who had not used Bloom's Taxonomy prior to the survey), AHSC faculty (nursing, public health, physical therapy, medicine, and clinical/rehabilitative health sciences), and nonAHSC faculty (arts and sciences, business and technology, education, graduate studies, and honors college). Alphas by version and subgroup are displayed in Table 4. The difference between the frequent users and novices was not significant for any version. Likewise, health science affiliation did not result in any statistical change in reliability. For all versions and all subgroups Krippendorff's Alpha was below the preferred threshold of .800.

Table 2

Participants' college affiliation

	Number	Percent of Sample
Arts & Sciences	36	27.5%
Business & Technology	12	9.2%
Clinical and Rehabilitative[1]	11	8.4%
Education	24	18.3%
Graduate Studies	1	0.8%
Honors College	1	0.8%
Medicine[1]	25	19.1%
Nursing[1]	14	10.7%
Public Health[1]	7	5.3%

[1] College affiliated with the Academic HealthSciences Center (AHSC)

Table 3

Participants' identified demographics for each survey version

	Department		*Usage*	
	Health Science	Non-Health Science	Novice	Non-Novice
Original	38.3%	61.7%	34.0%	66.0%
Collapse One	50.0%	50.0%	40.5%	59.5%
Collapse Two	40.5%	59.5%	33.3%	66.7%

Table 4

Interrater reliability expressed as α (95% CI)

	Overall	*Novices*	*Frequent Users*	*Health*	*Nonhealth*
Original	.308(.341-.419)	.356(.317-.397)	.448(.358-.537)	.379(.339-.418)	.392(.353-.430)
Collapse One	.423(.324-.515)	.350(.247-.451)	.398(.298-.492)	.393(.294-.491)	.451(.355-.547)
Collapse Two	.426(.328-.524)	.374(.271-.477)	.531(.428-.633)	.440(.341-.535)	.406(.303-.507)

Accuracy

Overall accuracy (the percent of items classified correctly) was 60.6% for the Original, 67.6% for Collapse One, and 77.6% for Collapse Two. Collapse Two yielded significantly higher accuracy than Collapse One $t(82)=4.46$, $p<.001$. Novices and nonnovices (participants who had used Bloom's Taxonomy prior to the survey) performed similarly in each version. Since the accuracy analyses required a larger sample size than the interrater reliability analyses - which only require two cases—frequent users could not be used; consequently, participants who had some prior experience with Bloom's Taxonomy were combined into one group: nonnovices. Health science participants' level of accuracy did not differ from nonhealth science participants' level of accuracy in any version (see Table 5 for a summary of the overall accuracy results).

Table 5

Accuracy of categorizing questions based on Bloom's Taxonomy based on demographic groups

	Overall	*Novices*	*Non-Novice*	*Health*	*Non-Health*
Original	60.6%	60.0%	61.4%	58.3%	61.9%
Collapse One	67.6%[1]	65.5%	69.6%	67.3%	68.6%
Collapse Two	77.6%[1]	75.0%	78.9%	78.9%	76.7%

1 – When comparing versions with three categories Collapse Two participants attained significantly higher accuracy levels than Collapse One participants ($p<.001$) using a Z test for proportions.

Accuracy for each Bloom's category varied substantially. In the Original participants were able to categorize Knowledge (85.9%) and Application (76.2%) items more accurately than any other type of item. Comprehension and Analysis items were the most difficult for participants to categorize at 40.9% and 45.1% accuracy, respectively. Table 6 summarizes the responses of participants.

Table 6

Original. Accuracy of responses for correct question category for Original version of Bloom's Taxonomy

		Actual Question Classification					
		Knowledge	Comprehension	Application	Analysis	Synthesis	Evaluation
Responses	Knowledge	**85.9%**	1.7%	0.4%	6.8%	0.0%	0.4%
	Comprehension	12.0%	**40.9%**	6.4%	14.9%	2.1%	3.1%
	Application	0.9%	21.7%	**76.2%**	13.6%	18.3%	7.8%
	Analysis	0.4%	17.9%	10.2%	**45.1%**	11.5%	24.4%
	Synthesis	0.4%	15.7%	5.1%	6.0%	**60.0%**	8.6%
	Evaluation	0.4%	2.1%	1.7%	13.6%	8.1%	**55.7%**

Note. Correct responses are bolded.

Accuracy in the Collapse One is shown in Table 7. Participants categorized 63.1% of Knowledge/ Comprehension items, 72.4% of Analysis/Application items, and 68.7% of Synthesis/Evaluation items correctly.

Table 7

Collapse 1. Accuracy of responses for correct question category for Collapse One version

		Actual Question Classification		
		Knowl/Comp	App/Analysis	Synth/Eval
Responses	Knowl/Comp	**63.1%**	10.7%	2.6%
	App/Analysis	27.4%	**72.4%**	28.7%
	Synth/Eval	9.5%	16.9%	**68.7%**

Note. Correct responses are bolded

Accuracy in Collapse Two is shown in Table 8. Participants categorized 92.3% of Knowledge items correctly, 71.9% of Comprehension/Application items correctly, and 69.5% of Analysis/Synthesis/Evaluation items correctly.

Table 8

Collapse 1. Accuracy of responses for correct question category for Collapse Two version

	Collapse Two	*Actual Question Classification*		
		Knowledge	Comp/App	Analysis /Synth/Eval
Responses	Knowledge	**92.3%**	5.0%	3.9%
	Comp/App	6.7%	**71.9%**	26.5%
	Analysis/Synth/Eval	0.9%	23.1%	**69.5%**

Note. Correct responses are bolded

Self-Assessment

More experienced participants over-estimated their ability to a greater extent than less experienced participants. Thus, experience using Bloom's Taxonomy may have a larger impact on perceived ability than on actual ability.

Absolute bias—the degree to which performance estimates differ from actual performance—was used as an index of self-assessment. In this study absolute bias is the difference between the proportion of items that participants believed that they categorized correctly and the proportion of items that they actually categorized correctly. In the Original version participants estimated that they categorized 68.0% of the items correctly and actually categorized 60.6% correctly, difference -7.4%, $t(46)=3.11$, $p=.003$. No absolute bias was observed in the two Collapsed versions. Collapse One participants estimated that they categorized 67.9% correctly and actually categorized 67.6% correctly, difference -0.3%, $t(42)=1.17$, $p=.247$. Collapse Two participants estimated that they categorized 74.8% correctly and actually categorized 77.6% correctly, difference 2.8%, $t(42)=1.10$, $p=.277$.

When all three surveys were combined to allow for adequate sample size—a repeated measures ANOVA with frequency of usage (Non-novice vs. novice) or health science affiliation (Non-health sciences vs. health sciences) as the between-subjects factor and predicted vs. actual percent correct as the within-subjects factor—they revealed that non-novice participants over-estimated their performance to a greater extent than novices, $F(1,128)=5.55$, $p=.020$. Non-health science participants showed significantly more optimistic absolute bias than health science participants, $F(1,128)=7.77$, $p=.006$ (see Table 9).

Table 9

Self-predicted accuracy based on participation demographic using repeated measure analysis of variance

Participant	Predicted Correct	Actual Correct	Difference: Actual–Predicted
Non-Novice[1]	74.8%	69.8%	-5.0% [$t(83)=2.89$, $p=.005$]
Novice	64.3%	66.1%	1.8% [$t(46)=.780$, $p=.439$]
Non-Health Science[2]	74.5%	68.7%	-5.8% [t(74)=3.35, p=.001]
Health Science[2]	66.1%	68.1%	2.0% [$t(54)=.880$, $p=.383$]

1 – Non-Novices made up 64.0% of the Non-Health Science group and 64.3% of the Health Science group.

2 - Including frequency of usage as a covariate did not alter these results.

Discussion

In line with Karpen and Welch (2016), interrater reliability was low in both the original and collapsed versions for both health science and non-science participants. Additionally, more experienced Bloom's users did not have significantly better reliability or accuracy than less experienced Bloom's users. Although, regarding accuracy, experience with Bloom's was analyzed nominally (yes or no) and did not necessarily equate to training. Overall, this study suggests that Collapse Two yields higher accuracy results than Collapse One. It is possible that faculty in any discipline think of assessments, regardless of nomenclature, as trichotomous: easy, medium, and hard questions. Collapsing categories, however, may dilute the data for assessment purposes. For example, combining Comprehension and Application may hide a desired distinction in abilities. If the original six-category hierarchy is desired by faculty then perhaps some alternative to collapsing, such as faculty development, may be useful.

Being able to assess a student's level of learning by an exam question relies on a faculty member's ability to accurately and reliably identify that level of learning. In this study the accuracy and reliability of categorizing Bloom's Taxonomy to exam questions were low.

Knowledge and application-level questions were categorized most accurately, perhaps because knowledge is the most basic category on the taxonomy; it represents a simple transfer of information. Application questions may have higher accuracy due to familiarity. They are commonly used in the health sciences—which represented a large portion of this study's sample (Blanco, Capello, Dorsch, Perry, & Zanetti, 2014). However, it is also suggested that multiple choice questions in general, cannot assess cognitive processes beyond knowledge recall (Scully, 2017).

Overall, participants overestimated their ability to use Bloom's Taxonomy. In both collapsed versions, however, the perceptions were similar to the outcomes, as fewer categories should make for easier accuracy estimation (Phillips et al., 2013). More experienced participants overestimated their ability to a greater extent than less experienced participants. Thus, experience using Bloom's Taxonomy may have a larger impact on perceived ability than on actual ability. Health science and non-health science participants also differed in their estimation accuracy such that health science participants more accurately estimated their performance than non-health science participants.

Conclusions

Being able to assess a student's level of learning by an exam question relies on a faculty member's ability to accurately and reliably identify that level of learning. In this study the accuracy and reliability of categorizing Bloom's Taxonomy to exam questions were low. Faculty are hired because of knowledge and expertise in a particular field and teaching abilities may come secondary to research or practice abilities in that field (Blanco et al., 2014; Ehrlich & Fu, 2012; Robinson & Hope, 2013). Using a collapsed version of Bloom's Taxonomy may be one way to improve accuracy in identifying learning. This approach may be useful to faculty of various disciplines and varying degrees of familiarity with Bloom's Taxonomy. However, collapsing Bloom's Taxonomy minimizes its distinction abilities. Faculty development may serve as one method to better understand their exam question hierarchy, though faculty development is challenging with pressures and demands on faculty (Szybinski & Jordan, 2010). Further research is needed to better identify ways to improve college faculty's abilities to identify levels of student learning through exam questions.

Authors' Note: Special acknowledgement to Ms. Emily Weyant of the Quillen Medical Library at East Tennessee State University for developing the search strategies to support this manuscript and Dr. John Bossaer for review of the manuscript during preparation. At a time of research, Brandie N. LeBlanc was a student was a student at East Tennessee State Univeristy. The authors declare no conflicts of interest or financial interests in any product or service mentioned in this article, including grants, employment, gifts, stock holdings, or honoraria.

References

Adams, N.E. (2015). Bloom's taxonomy of cognitive learning objectives. *Journal of the Medical Library Association JMLA, 103*(3), 152–153. doi: 10.3163/1536-5050.103.3.010

Blanco, M.A., Capello, C.F., Dorsch, J.L., Perry G (Jerry), & Zanetti, M.L. (2014). A survey study of evidence-based medicine training in US and Canadian medical schools. *Journal of the Medical Library Association JMLA, 102*(3), 160–168. doi: 10.3163/1536-5050.102.3.005

Bloom, B.S. (1956). *Taxonomy of Educational Objectives, the classification of educational goals—Handbook I: Cognitive Domain*. New York: McKay.

Carnegie Classification of Institutions of Higher Education (2017). Institution lookup. Retrieved from http://carnegieclassifications.iu.edu/

Cecilio-Fernandes, D., Kerdijk, W., Jaarsma, A.D.D.C., & Tio, R.A. (2016). Development of cognitive processing and judgments of knowledge in medical students: Analysis of progress test results. *Medical Teacher*, 1125–1129. doi:10.3109/0142159X.2016.1170781

Ehrlich, T., & Fu, E. (2013, November 12) College professors: Before you teach, learn how! *Forbes*. Retrieved from http://www.forbes.com/sites/ehrlichfu/2013/11/12/college-professors-before-you-teach-learn-how/#3d26237057f5

Engle, J.P., Erstad, B.L., Anderson Jr., D.C., Bucklin, M.H., Chan, A., Donaldson, A.R., et al. (2014). Minimum qualifications for clinical pharmacy practice faculty. *Pharmacotherapy, 34*(5), e38–44.

ExamSoft For Your Programs. (2017). Retrieved from http://learn.examsoft.com/exam-programs George, T. (2016). Why take the certified nurse educator exam? *Nursing, 46*(3), 21–24. doi: 10.1097/01.NURSE.0000480615.77543.06

Gonzalez-Cabezas, C., Anderson, O.S., Wright, M.C., & Fontana, M. (2015). Association between dental student-developed exam questions and learning at higher cognitive levels. *Journal of Dental Education, 79*(11), 1295–1304.

Igbaria, A.K. (2013). A content analysis of the WH-Questions in the EFL textbook of horizons. *International Education Studies, 6*(7), 200. doi: 10.5539/ies.v6n7p200

Jideani, V.A., & Jideani, I.A. (2012). Alignment of assessment objectives with instructional objectives using revised Bloom's Taxonomy—the case for food science and technology education. *Journal of Food Science Education, 11*(3), 34–42. doi: 10.1111/j.1541-4329.2012.00141.x

Karpen, S., & Welch, A.C. (2016). Assessing the interrater reliability and accuracy of pharmacy faculty's Bloom's Taxonomy classifications. *Currents in Pharmacy Teaching and Learning*. Retrieved from http://dx.doi.org/10.1016/j.cptl.2016.08.003

Kibble, J.D., & Johnson, T. (2011). Are faculty predictions or item taxonomies useful for estimating the outcome of multiple-choice examinations? *Advances in Physiology Education, 35*(4), 396–401. doi: 10.1152/advan.00062.2011

Krathwohl, D.R. (2002). A revision of Bloom's taxonomy: an overview. *Theory Into Practice, 41*(4), 212-218. doi: 10.1207/s15430421tip4104_2

Krippendorff, K. (1970). Estimating the reliability, systematic error and random error of interval data. Educational and *Psychological Measurement, 30*(1), 61–70.

Landis, J.R., & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174.

Momsen, J.L., Long, T.M., Wyse, S.A., & Ebert-May, D. (2010). Just the facts? Introductory undergraduate biology courses focus on low-level cognitive skills. *CBE - Life Sciences Education, 9*(4), 435–440. doi: 10.1187/cbe.10-01-0001

Phillips, A.W., Smith, S.G., & Straus, C.M. (2013). Driving deeper learning by assessment: an adaptation of the revised Bloom's Taxonomy for medical imaging in gross anatomy. *Academic Radiology, 2*0(6), 784–789. doi:10.1016/j.acra.2013.02.001

Plack, M. M., Driscoll, M., Marquez, M., Cuppernull, L., Maring, J., & Greenberg, L. (2007). Assessing Reflective Writing on a Pediatric Clerkship by Using a Modified Bloom's Taxonomy. *Ambulatory Pediatrics, 7*(4), 285–291. doi:10.1016/j.ambp.2007.04.006

Robinson, T.E., & Hope, W.C. (2013). Teaching in higher education: Is there a need for training in pedagogy in graduate degree programs? *Research in Higher Education Journal, 21*, 1-11

Scully, D. (2017). Constructing multiple-choice items to measure higher-order thinking. Practical Assessment, *Research & Evaluation, 22*, 4. Retrieved May 23, 2017, from http://pareonline.net/getvn.asp?v=22&n=4

Szybinski, D., & Jordan, T. (2010). Navigating the future of the professoriate. *Peer Review, 12*(3). Retrieved from https://www.aacu.org/publications-research/periodicals/navigating-future-professoriate

Terry, C. (2016, September 19). Bloom's Taxonomy (Part 2): Using Bloom's Taxonomy in assessment. Retrieved from http://resources.examsoft.com/examsofts-blog/bloom-s-taxonomy-part-2-using-bloom-s-taxonomy-in-assessment

University of California–Berkeley Center for Teaching and Learning. (2015). Bloom's Taxonomy [Internet]. Retrieved from http://teaching.berkeley.edu/blooms-taxonomy

Vandre, D.D., & Ermie, E. (2017). Improving student learning outcomes. ExamSoft [Internet]. Retrieved from http://resources.examsoft.com/white-papers/improving-student-learning-outcomes