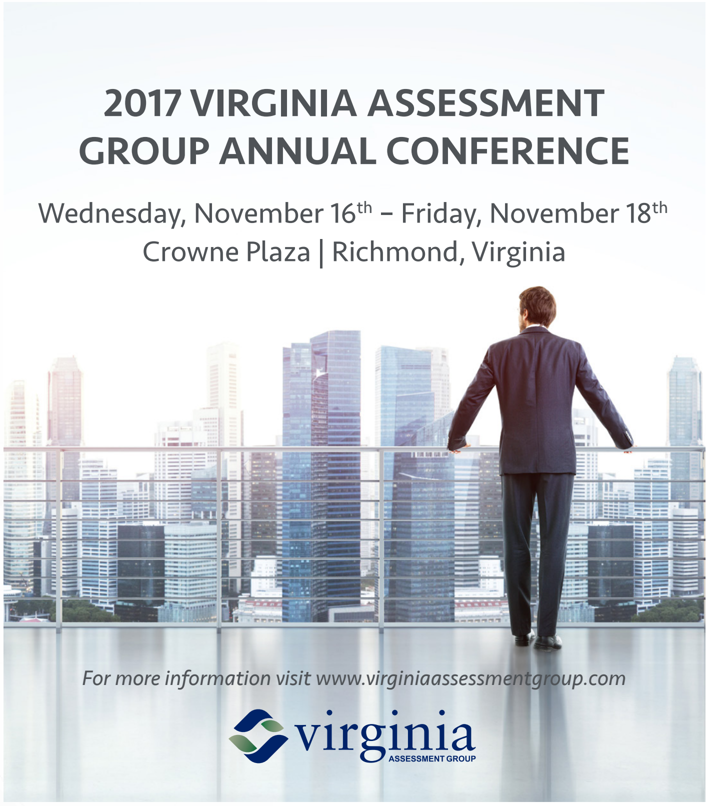
RESEARCH & PRACTICE IN ASSESSMENT

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

RPA

A PUBLICATION OF THE VIRGINIA ASSESSMENT GROUP

RPA RESEARCH & PRACTICE IN ASSESSMENT



CALL FOR PAPERS

Research & Practice in Assessment is currently soliciting articles and reviews for its Winter 2017 issue. Manuscripts submitted to RPA may be related to various higher education assessment themes, and should adopt either an assessment measurement or an assessment policy/foundations framework. Contributions are accepted at any time, but submissions received by August 1 will receive consideration for the winter issue. Manuscripts must comply with the RPA Submission Guidelines and be sent electronically to: editor@rpajournal.com

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

TABLE OF CONTENTS

IN MEMORY OF

Deborah L. Moore
RPA Review Board and *Virginia Assessment Group*

Kendra Jeffcoat
RPA Review Board

FROM THE EDITOR

Choosing Wisely

Faculty, student affairs educators, and administrative leaders in higher education regularly seek, generate, and use information to inform their decision making and their practices. Some of the information used by these individuals is related directly to student learning outcomes, and some of the information informs programs and services that advance student persistence and achievement. The accuracy of the data, the reliability of the measures, and the validity of the inferences remain paramount to assessment research and practice.

The contributions presented in this issue of *Research & Practice in Assessment* demonstrate the importance of data in decision making and selecting appropriate measures to demonstrate student learning and achievement outcomes. It is likely that you have faced challenges with the use of such measures.

The Summer 2017 issue includes three peer-reviewed articles that exemplify the importance of data use and utilization of strong measures to advance assessment practices in higher education. Addressing the challenge of predicting student readiness for college, Soland utilizes data reduction techniques to identify the strongest identifiers of college readiness. Gray, Brown, and Connolly present evidence of using the Association of American Colleges and Universities (AAC&U) Quantitative Literacy VALUE rubric to assess a single construct, Empirical and Quantitative Skill, in performance-based assessment. Prinski and Winterrowd examine the implications of measure selection when assessing the impact of campus counseling services on students' academic functioning.

Stitt-Bergh reviews *Learning Assessment Techniques: A Handbook for College Faculty*, a text designed for faculty to incorporate assessment techniques that not only measure, but also facilitate student learning.

This issue also includes a Notes in Brief that highlights the importance of teaching graduate students how to engage in robust assessment practices and affords the opportunity for faculty to reflect and examine their own teaching practices. I hope the scholarship and practices presented in this issue will be applicable to your work.

On a more solemn note, it is with deep sympathy, that I take this opportunity to remember and recognize two valued members of the RPA family, Kendra Jeffcoat and Deborah "Deb" Moore. Kendra Jeffcoat served as a member of the RPA Review Board and Deb Moore was a member of the Virginia Assessment Group as well as a member of the RPA Review Board. As colleagues and volunteers, they both gave selflessly of their time expertise. Kendra and Deb will be remembered fondly and missed deeply.

Regards,

Katie Busby

University of Mississippi

AUTHORS

James Soland, Ph.D.
NWEA

Abstract

Research shows college readiness can be predicted using a variety of measures, including test scores, grades, course-taking patterns, noncognitive instruments, and surveys of how well students understand the college admissions process. However, few studies provide guidance on how educators can prioritize predictors of college readiness across instruments, constructs, and frameworks to optimally identify students not on track for college. Using a nationally representative dataset with thousands of measures, I employ data reduction techniques to identify a handful of variables that are the strongest predictors of college readiness and understand what they measure. Based on my models, enrolling in college and persisting for a semester can be predicted with almost 90 percent accuracy using a small set of predictors. Evidence suggests these predictors measure academic preparation, postsecondary aspirations, teacher perceptions of readiness, and socioeconomic status. Educators can use results to help identify appropriate supports for students not on track for college.

Combining Academic, Noncognitive, and College Knowledge Measures to Identify Students Not on Track For College: A Data-Driven Approach

Considerable research has been devoted to identifying measures that predict whether a student is ready for college. These types of measures include (but are not limited to) test scores, grade-point average (GPA), secondary course-taking patterns, assessments of so-called "noncognitive" factors like grit, and surveys of how well students understand the college admissions process. Within each of these types of measures, there may be dozens of assessments or instruments shown to forecast college readiness. For example, there are hundreds of noncognitive surveys, many of which predict postsecondary success (Dweck, Walton, & Cohen, 2011). While a number of papers suggest how these different measures relate to each other (Conley, 2008; Dweck, Walton, & Cohen, 2011; Farrington et al., 2012), little guidance exists for practitioners on how to combine and prioritize these measures with the goal of optimally identifying students who will enter, and persist in, college. As a result, educators are left with a potentially overwhelming array of measures to choose from when attempting to figure out which students are not on track for college. I help address this issue by using data reduction techniques designed to select the measures that maximize the accuracy with which students are identified as being ready to enroll in college and persist for at least a semester. Results show that students not on track for a postsecondary education can be identified with 90 percent accuracy using only a handful of predictors. These predictors tend to measure four broad constructs—academic preparation, educational aspirations and expectations, socioeconomic status, and teacher perceptions of student performance—that educators can use to inform development of college readiness instruments and interventions for students, either before or after college begins.

CORRESPONDENCE

Email
jim.soland@nwea.org

These issues are examined by mining a national dataset with thousands of college readiness-related variables to answer two specific research questions. First, can a small subset of measures be combined into a single model that accurately and consistently identifies students not on track for college? Second, what are the most important predictors of college readiness across models and what do they measure? When trying to identify students not on track to enroll and persist in college, practitioners can use answers to these questions to maximize the usefulness of the data they already have—and find new readiness predictors that are simple to measure but can greatly increase the accuracy with which off-track students are identified.

Literature on College Readiness & Data Reduction

In this section, I review studies on college readiness that are most pertinent to the research questions and data, including relevant measures and frameworks. I also examine the growing use of data reduction techniques in education. Despite their promise, these data reduction methods have not been used to identify the strongest predictors of college readiness nor to help identify what educators should measure if trying to support students who are not on track for college.

College Readiness Inputs

David Conley (2005, 2007, 2008, 2010) has suggested that there are four main components to being college ready, including being prepared academically, mastering various cognitive strategies, understanding the college process, and holding certain attitudes. The first, being prepared for college academically, involves both developing content knowledge and using that content knowledge to solve novel problems. Students must also use analytical strategies that are not content specific, such as the ability to reason, argue, and interpret. Conley (2007) argues that students may seem ready for college based on content mastery but still lack these other competencies, suggesting that college eligibility and college readiness are not the same. His research influences and is informed by work showing that the courses students take and their performance in them, especially course failures and GPA, are much better predictors of high school graduation and college readiness than standardized test scores (Allensworth, 2013; Allensworth & Easton, 2005; Balfanz & Boccanfuso, 2007; Camara, 2013; Neild, Balfanz, & Herzog, 2007; Roderick et al., 2006).

…little guidance exists for practitioners on how to combine and prioritize these measures with the goal of optimally identifying students who will enter, and persist in, college.

Beyond academics, research shows students need to have a basic understanding of the admissions process and how to survive in postsecondary settings that emphasize autonomy, an understanding many term “college knowledge” (Conley, 2005, 2008; Hooker & Brand, 2010; York-Anderson & Bowman, 1991). For example, a student may be highly qualified for college academically but not know how to secure financial aid, without which a postsecondary education would be prohibitively expensive. The importance of college knowledge has been reinforced by randomized control trials around interventions related to the college admissions process. One recent study randomly selected low- to middle-income families filing their taxes with H & R Block to have federal student loan applications filled out for them, as well as receive information on eligibility for financial aid and estimated award amounts (Bettinger, Long, Oreopoulos, & Sanbonmatsu, 2012). Students from families receiving these supports were substantially more likely to submit the aid application, enroll in college the following fall, and receive financial aid (Bettinger, Long, Oreopoulos, & Sanbonmatsu, 2012). How well students understand college processes and contexts can also influence whether and where they decide to go to school (Perna, 2000), and how comfortable they feel when they get there (Hurtado & Carter, 1997).

Researchers further show that students’ attitudes and beliefs are important to success in college. Resilience, self-regulation, and beliefs about intelligence are all predictive of college grades and completion (Cury, Elliot, Da Fonseca, & Moller, 2006; Durlak, Weissberg, Dymnicki, Taylor, & Schellinger, 2011; Dweck et al., 2011). For instance, students who believe they belong in their college’s academic community are likelier to persist and graduate, especially among minority and first-generation students (Walton & Cohen, 2007, 2011). In some instances, these measures of attitudes and beliefs have predicted long-term academic achievement better than grades and test scores (Deke & Haimson, 2006; Duckworth, Grant, Loew, Oettingen, & Gollwitzer, 2011; Duckworth & Quinn, 2009; Duckworth, Peterson, Matthews, & Kelly, 2007; Good & Dweck, 2006).

Sedlacek's (2004) work on noncognitive measures shows how results of studies on student attitudes can be used in the college admissions process. For example, Noonan, Sedlacek, and Veerasamy (2005) find that noncognitive questionnaire responses predict GPA for community college students and should therefore be used in admissions criteria. Noncognitive outcomes also vary in how much they improve college readiness forecasts by race. Tracey and Sedlacek (1982) show that different noncognitive questions predict college GPA for white students (self-confidence, preference for long-range goals, and realistic self-appraisal) compared to black students (positive self-concept and realistic self-appraisal). In more recent work, Sedlacek (2004) argues that using noncognitive outcomes in college admissions is not only useful, but also practically feasible.

College Readiness Frameworks

Given the range of measures and constructs that research suggests are associated with postsecondary success, college readiness frameworks abound. Some of these frameworks, like Conley's (2005, 2007, 2008, 2010), look across all types of measures to identify the constructs most important to college readiness (Borsato, Nagaoka, & Foley, 2013). Others look specifically at certain types of measures, especially noncognitive factors. For example, Dweck, Walton, and Cohen (2011) highlighted a few attitudes and beliefs that predict achievement: tenacity, mindset, social belonging, and self-regulation. They argued that, in combination, having these noncognitive factors in place generates motivated students. Similarly, a literature review produced by the Chicago Consortium on School Research (Farrington et al., 2012) synthesized different categories of noncognitive skills into a framework that shows how these skills relate to each other and influence academic achievement. They suggested that academic behaviors (e.g. completing school work in a timely way) are most proximal to achievement and that these behaviors are driven by social skills, perseverance, and learning strategies (Farrington et al., 2012).

Despite their promise, these data reduction methods have not been used to identify the strongest predictors of college readiness nor to help identify what educators should measure if trying to support students who are not on track for college.

These frameworks are often cited and likely useful to educators. However, they do not always give concrete guidance on how to prioritize and blend different measures of college readiness when trying to identify students most in need of related supports and interventions. That is, they show that certain measures and constructs are important but often emphasize different aspects of readiness and do not look across frameworks to identify the relative importance of predictors in forecasting college readiness. I use data reduction techniques to help close this gap.

Data Reduction Techniques and College Readiness

Over the last few decades, data reduction techniques have been developed to help make decisions when researchers or practitioners are awash in data, including cases where there are more variables than observations. These data reduction techniques, specifically decision trees, have advantages relative to traditional regression-based methods that can help identify the best predictors of college readiness. For example, these methods are straightforward, produce results that are easily interpretable (including to lay audiences), and generate clear decision points for action (Magee, 1964; Murthy, 1998; Quinlan, 1990). A range of articles demonstrate, and argue for, the importance of data reduction techniques when identifying students who are not ready for college or are unlikely to persist (Baker & Corbett, 2014; Denley, 2014; O'Reilly & Veeramachaneni, 2014). However, this literature remains sparse.

Fong, Si, and Biuk-Aghai (2009) used decision trees to predict university admissions in Macau and, thereby, better understand how students ended up in the schools they ultimately attended. Specifically, they predicted which university a student was likeliest to enter by analyzing achievement, student background, and university admissions criteria. These models were shown to accurately predict which school a student would enter over 95 percent of the time, and were easily tailored to students' interests and experiences (Fong et al., 2009). Other studies, meanwhile, used decision trees to identify which students entering college were likely to complete their postsecondary education. Dekker, Pechenizkiy, and Vleeshouwers (2009) used data reduction to identify which college freshmen in an electrical engineering program were likeliest to drop out before receiving their degrees. They were able to classify students with 80 percent accuracy and, in the process, identify dropout indicators that were not being used as warning flags by student counselors (Dekker et al., 2009). Similarly, when examining high school students, Quadri and Kalyankar (2010) showed that dropout indicators included a range of measures like GPA, gender, attendance, and parental income.

Data

I use the National Education Longitudinal Survey (NELS) in my analyses. This dataset tracks students from their 8th-grade year in 1988 through 2000—which means postsecondary outcomes are recorded. In addition to providing several college readiness outcomes, the dataset is nationally representative and includes thousands of student survey items. (One reason I use NELS instead of the Education Longitudinal Survey (ELS), which is similar and more recent, is that the former includes a broader range of student survey items.) I use various measures from students' 8th- and 10th-grade years (the two grades at which data were collected) to predict college readiness.

Researchers further show that students' attitudes and beliefs are important to success in college.

One disadvantage to NELS survey items is that they are not organized by construct. Therefore I have no concise way to summarize what, exactly, these thousands of survey items measure. However, there are some distinct themes in the questions, which relate (among other things) to postsecondary aspirations and expectations, home environment, language proficiency, attitudes and beliefs about schooling, occupational expectations, self-esteem and locus of control, quality of life, and college choice factors. These items span different aspects of college readiness, including academic preparation, attitudinal factors, and college knowledge.

Researchers for NELS collected baseline data on 12,144 students, of whom 9,601 (79 percent) attended college[1]. Of those 12,144, roughly 8,800 persisted for a semester. I use college attendance for a semester rather than completion as an outcome despite the latter being a much better measure of college readiness. This approach is taken because available measures of enrollment after a semester are imperfect, let alone after two to four years. Further, sample sizes dwindle considerably when college completion is the outcome, and there are only limited (and technically complicated) options for dealing with missingness in the dependent variable. Therefore, I try to balance the accuracy and completeness of the outcome variable with how well it measures college readiness.

There are, however, several basic approaches available to address the substantial missingness in the predictors. First, when a variable is nominal (i.e. unordered categorical), NELS always includes a category for missing. In a data reduction framework, I can divide nominal variables into a series of binary variables, including a separate binary variable for missing versus nonmissing. Second, I impute missing values based on all of the covariates in the sample using a fully conditional estimation strategy appropriate for categorical variables (Buuren & Groothuis-Oudshoorn, 2011; Van Buuren, Brand, Groothuis-Oudshoorn, & Rubin, 2006). Imputation has been used many times before with NELS (Bokossa & Huang, 2001) and is appropriate given the missingness in the data appears to be covariate-dependent.

A measure to identify students not on track for college is only as good as its ability to predict which students fall in each category when we do not know their outcome. Educators are primarily interested in signs that a student is off-track well before the end of high school in order to provide those students with the supports they need.

Training and Test Datasets

A measure to identify students not on track for college is only as good as its ability to predict which students fall in each category when we do not know their outcome. Educators are primarily interested in signs that a student is off-track well before the end of high school in order to provide those students with the supports they need. To simulate this process, I use an approach well documented in the data reduction literature, namely dividing the full dataset into training and test data. The training data are used to fit models, then predictions relying on those model parameters are made for the test data. This approach means we can test the model on a set of students not used to derive that model, as well as protect against overfitting the model (mistaking the noise for signal). Roughly 70 percent of the 12,144 students in the sample were randomly assigned to the training set with the remaining 30 percent assigned to the test set. Results did not change substantively when different proportions of students were assigned to each group.

[1] College is defined as any two- or four-year institution.

Methods

In this section, I discuss methods on a question-by-question basis. Those questions are as follows:

1) Can a small subset of measures be combined into a single model that accurately and consistently identifies students not on track for college?
2) What are the most important predictors of college readiness across models and what do they measure?

Can a small subset of measures be combined into a single model that accurately and consistently identifies students not on track for college?

Decision trees start with an outcome of interest—in our case attending college and persisting for a semester—then use an algorithm that cycles through each variable in the dataset, select the one that maximizes information about the outcome for individuals in the sample, then repeat the process until some stopping criterion is met. In more mathematical terms, we want to maximize the information we have about an outcome *Y* given a variable, C: I[C:Y]. We can do this using the sum of squared errors *S* for person *i* at variable[2] (or "leaf") *c* for tree *T*:

$$S = \sum_{c \in leaves\,(T)} \sum_{i \in c} (y_i - m_c)^2 \qquad (1)$$

where $m_c = \frac{1}{n_c} \sum_{i \in c} y_i$, the prediction for leaf *c*. I can rewrite (1) as:

$$S = \sum_{c \in leaves\,(T)} n_c V_c \qquad (2)$$

where V_c is the within-leaf variance for leaf c. All splits are made to minimize *S*. Since every split produces two leaves, for the first two leaves we are really minimizing:

$$min[{}^{min}_{C1}[n_1 V_1] + {}^{min}_{C2}[n_2 V_2]\,] \qquad (3).$$

One measure that can be used to determine the information gain from a split is the Gini index, which is directly related to within-leaf variability. The Gini index can be thought of as the variance of a response variable summed over the *k* classes of that variable (for more information see Friedman, Hastie, & Tibshirani, 2001). I use the mean reduction in the Gini index (MDGI) as a measure of the importance of a variable in classifying students (Archer & Kimes, 2008).

As previously described, models are fit using training data and then applied to the test data. That is, students in the test data are classified as expected to go to college or not based on the parameters from the training data. I gauge the accuracy of these predictions using two measures. First, I consider the overall classification accuracy using a two-by-two contingency table with predictions of whether the student attended college against actual outcomes. I then divide the number of students on the diagonals by the total number of students. Second, I examine the ratio of false positives to false negatives.

The issue of balancing false positives and negatives in the predictions cuts across models and research questions. If one were to simply guess that every student will go to college then the accuracy rate would be 79 percent, which is roughly the rate touted for other models from the high school graduation and college readiness literature (Bowers, Sprott, & Taff, 2013). Such a model provides no useful information because it does not identify any students as needing support to be college-ready. While this example is extreme, it highlights the fact that there may be a bigger practical downside to wrongly identifying students as being ready

Decision trees start with an outcome of interest—in our case attending college and persisting for a semester—then use an algorithm that cycles through each variable in the dataset, select the one that maximizes information about the outcome for individuals in the sample, then repeat the process until some stopping criterion is met.

Given the range of measures and constructs that research suggests are associated with postsecondary success, college readiness frameworks abound.

[2] Unordered categorical variables are coded into a series of dummy variables in these models.

These data reduction techniques, specifically decision trees, have advantages relative to traditional regression-based methods that can help identify the best predictors of college readiness.

for college than the opposite. That is, one could argue there is a bigger downside to not giving a student support who needs it than to helping a student who is actually college ready. To acknowledge this argument, I built additional trees that were weighted to de-emphasize false positives. I accomplished this weighting using a built-in feature of the ctree package (Hothorn, Hornik, & Zeileis, 2006) in the R programming language.

What are the most important predictors of college readiness across models and what do they measure?

In the first question, I built a single model designed to maximize the accuracy with which students were classified as not on track for college. In this question, I attempt to identify the best predictors of college readiness across a range of models and determine what the predictors measure in tandem. I approach this question in two steps. First, I fit a series of data reduction models, then see which predictors are most important across those models. Second, I do a factor analysis of those important predictors to help determine what constructs they appear to be measuring. By understanding what they measure, I can provide general guidance on what educators may wish to assess when trying to identify students who need college supports, both those who are far off course and those who are more borderline.

I also build models that do not use decision trees because one could plausibly argue that a decision tree misrepresents the data. Decision trees assume the data can be divided into smaller datasets that are modeled in turn.

First, I generate a variety of models to identify the best predictors of college readiness. In all cases, these models are built using the training data then applied to the test data. One model replicates the original decision tree many times. This type of model is called a random forest. Random forests address a problem with decision trees, namely that they can be sensitive to the first split. For example, 100 decision trees might produce 50 different first splits. Since the rest of the tree is conditional on that first split each would produce discrepant results. Random forest models were developed to counteract this sensitivity to the first split and generally protect against overfitting. Random forests replicate decision trees hundreds of times in order to identify the covariates that do the best job of classifying students on average across all the replications. (For more details on this method, please see Elith, Leathwick, & Hastie, 2008). I identify variables that show the highest MDGI across these replicated decision trees as important predictors. Specifically, I decide which variables to retain by estimating the random forest models with a dataset that includes ten variables generated randomly with known means and variances. I then keep any variable that had a higher MDGI than these randomly generated variables.

I also build models that do not use decision trees because one could plausibly argue that a decision tree misrepresents the data. Decision trees assume the data can be divided into smaller datasets that are modeled in turn. By contrast, the data could be represented altogether in a single model, the approach used in most regression methods. To acknowledge this alternative conceptualization of the data I use lasso, a regression-based method. Lasso essentially runs a series of regressions omitting each variable, determines how much the omission impacts the residual sum of squares, then shrinks coefficients to zero when they have little predictive power. Mathematically, this model can be described for student i and covariate j as:

$$\hat{\beta}^{lasso} = \min_{\beta} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$$

With the below restriction:

$$\sum_{j=1}^{p} |B_j| \leq t$$

I identify t using a cross-validation process on the training data outlined in Tibshirani (1996)[3]. Any variables with non-zero coefficients are deemed strong predictors.

In the second step, I take all the variables identified as important in the random forest and lasso models and explore what they measure. I do this by conducting a factor analysis on those highly predictive variables. In particular, I do an exploratory factor analysis on the training data followed by a confirmatory factor analysis on the test data to make sure the model fits data from a different group of students.

[3] Per Gelman (2008), all variables are scaled by dividing by two standard deviations to make them comparable to binary variables. Nominal variables are coded as a series of dummy variables.

In addition to a standard factor analysis model, I also fit higher-order factor and bifactor models. The higher-order and bifactor models acknowledge that college readiness could involve both a single construct and several subconstructs. In practical terms, these models recognize that there might be a general aspect of college readiness that cuts across measures as well as specific aspects of readiness that may or may not relate to that general aspect. The higher-order model treats factors as nested within that general construct. The bifactor model, meanwhile, includes a general factor but treats it as distinct from the other constructs being measured. All of these models are well established in the literature so I will not show the underlying matrix algebra here. For details, please see Holzinger and Swineford (1937). Analyses in this paper were implemented using the omega command in the Psych package in R (Revelle, 2011; Revelle & Zinbarg, 2009).

The higher-order and bifactor models acknowledge that college readiness could involve both a single construct and several subconstructs. In practical terms, these models recognize that there might be a general aspect of college readiness that cuts across measures as well as specific aspects of readiness that may or may not relate to that general aspect.

To determine which factor analysis model to use, fit statistics including Chi-square statistics and the root mean squared error of approximation (RMSEA) were compared across models, the latter of which is not sample-size dependent (Koh & Zumbo, 2008). Based on these statistics the bifactor model fit the data best. For example, the chi-square statistics for the bifactor model is 2403 and the RMSEA is .056 (the RMSEA for the higher-order model was .059). Therefore, results from the bifactor model are reported in the findings section. The number of factors retained was determined using scree plots and Very Simple Structure analysis (Revelle & Rocklin, 1979).

Limitations

Despite the benefits of this methodological approach there are limitations worth mentioning. First, the methods are designed to maximize classification accuracy—not necessarily how actionable a measure may be for educators. While a survey question may be quite good at dividing students into those who are and are not on track for college, the question itself might have little to do with the root causes of being off-track—which would limit its use in informing instructional practice. Even if a measure does help identify an underlying cause of being off track that cause might not be especially actionable. For example, if aspects of poverty cause students to be less ready for college there are few options for educators to improve students' socioeconomic statuses.

This approach also does little to improve the measures themselves. For example, decision tree models do not generally account for measurement error, though the factor analyses do, to some extent. Further, survey responses can still suffer from self-report biases and grades can still rely on subjective judgments from teachers. Using such measures to make decisions with implications for students and teachers would likely only increase these deficiencies.

Even if a measure does help identify an underlying cause of being off track that cause might not be especially actionable. For example, if aspects of poverty cause students to be less ready for college there are few options for educators to improve students' socioeconomic statuses.

Relatedly, despite the wide range of questions asked and constructs measured by the NELS surveys, not all constructs of interest are included. For example, NELS largely predates Duckworth and Quinn's (2009) work on developing and validating a short grit scale and much of Dweck's (2006) research on growth mindset. Therefore, additional research will be needed to continue to understand the relative importance of constructs in determining which students are on track for college. Findings should also be replicated using newer data and different sets of survey questions.

Results

I show that a small number of measures can be combined in a single model to identify which students are on track for college with great accuracy. Across models, these predictors appear to measure academic achievement, college aspirations and expectations, socioeconomic status, and teacher perceptions of postsecondary readiness.

Can a small subset of measures be combined into a single model that accurately and consistently identifies students not on track for college?

Figure 1 displays the decision tree produced using weights[4]. For these models, roughly 12 percent of students in the training data and 14 percent in the test data were misclassified. The balance between false negatives and positives was much better than in the unweighted models. This tree's first split is based on a question asking students how far in school they think they will go. Splits that immediately follow primarily use questions about students' educational plans, standardized test scores, and socioeconomic status. The covariates in the tree include a blend of test scores, administrative data, and student and teacher surveys.

In some cases, a very small number of splits can be used to identify students with a high risk of not going to, and persisting in, college.

In some cases, a very small number of splits can be used to identify students with a high risk of not going to, and persisting in, college. For example, looking at the left-most branch of the decision tree, among the 712 students who did not think they would make it to college and were in the bottom quartile of SES, only 28 percent went to college and persisted. The tree also helps identify students who are high risk but may be less obvious absent the tree. For instance, going down the right-most branch, students who plan to go to college, are in the upper two standardized test quartiles, but who do not plan to go straight into college, occasionally cut class, and have low social studies grades have only a 50 percent chance of being college-ready.

What are the most important predictors of college readiness across models and what do they measure?

Table 1 presents results from lasso and random forest models as well as loadings from the factor analysis. Roughly 40 variables were deemed strongly predictive based on random forest and lasso models. I report only the coefficients from the latter because they are very highly correlated with the MDGI values and are easier to interpret. The coefficients can be interpreted as the change in the log odds of going to, and persisting in, college associated with every one-unit increase in the independent variable, each of which has been standardized. On average, random forest models accurately classified 91 percent of students in the training set and 88 percent in the test set—with far fewer false positives than in unweighted models. Classification rates were similar for the lasso models.

While there are not many studies describing the classification accuracy of models predicting college enrollment there are dozens showing the classification accuracy for high school completion. The accuracy rates of the models are higher than those from 97 percent of the studies predicting high school completion catalogued by Bowers, Sprott, and Taff (2013), a statistic made more meaningful by the fact that models using only K-12 data usually do a better job of predicting high school completion than college enrollment (Soland, 2013). Variables that were predictive in both random forest and lasso models included measures of postsecondary aspirations and expectations, socioeconomic status, GPA, standardized test scores, and teacher survey questions.

This study finds both GPA and test scores to be important forecasters of college readiness whereas other research has found test scores to be less predictive of outcomes like high school graduation when GPA is accounted for.

Turning to the factor loadings in Table 1, I show how individual variables load on the general factor and each subfactor. Only loadings of .2 or higher are reported, and variables that do not have loadings of .2 or higher on any variable are omitted for parsimony. Beyond the general factor, the variables load on roughly four factors. The first of these factors is associated primarily with test scores and other measures of academic preparation like grades. Variables related to postsecondary aspirations and expectations load on the second factor, as do factors in the college application process like the importance of financial aid to the student. The third factor is correlated with measures of socioeconomic status, including parents' level of educational attainment and family income. Finally, the fourth factor is associated with measures related to teachers' perceptions of the student's college readiness, including survey questions asking whether the student is likely to finish high school and go on to college.

Discussion

There are two major findings from this work, both of which can be useful to practitioners. First, results show that, while there are thousands of measures that forecast college readiness, a student can be classified as on track to enroll in, and persist at, a postsecondary institution with a high degree of accuracy (over 90 percent) using a small number of variables. Second, these predictors of college readiness tend to measure four things: academic preparation, educational aspirations and expectations, socioeconomic status, and teacher perceptions of student performance (GPA, teachers' confidence the student will go to college, etc.).

[4] A copy of the decision tree using the original NELS variable names is available upon request.

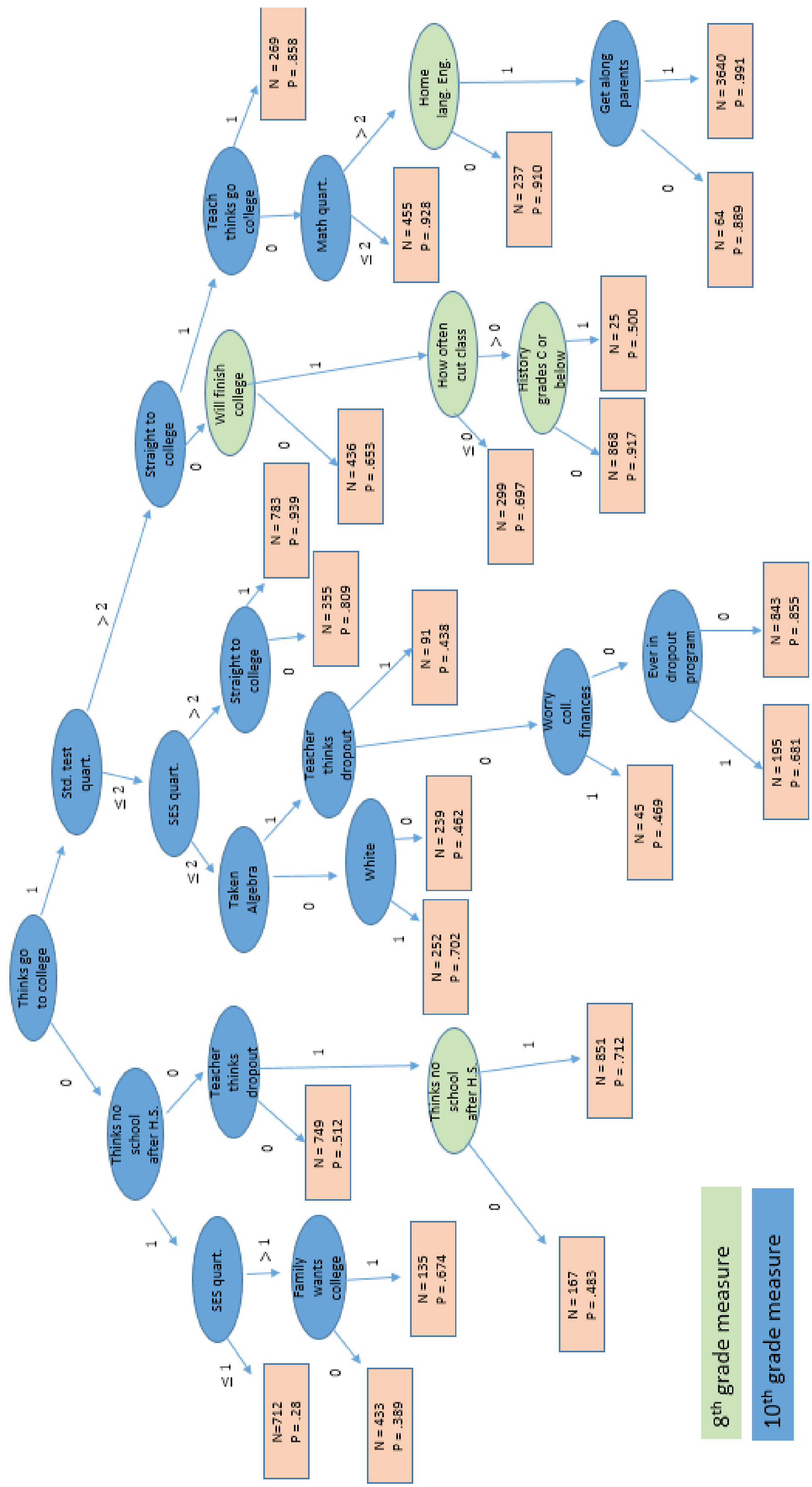


Figure 1. Decision Tree Results

One implication of these findings is that educational systems can likely mine their data to make more effective and efficient decisions about student needs. While individual K-12 and higher education administrators are unlikely to have the time or expertise to build decision tree models, one could certainly imagine such models being developed by a state department of education or university system. Decision trees could be built using available data at minimal cost—especially in states with K-12 and higher education data systems that are integrated. Such trees could then be used by practitioners at specific institutions given how easy results are to interpret. Models could be developed not just to predict who goes to college but also which students who enroll are on track to receive a diploma or training certificate.

Still, findings highlight the need for more research on how much teachers control the postsecondary fates of their students and how those fates are influenced.

Another implication is that results can be used to help practitioners decide which measures to prioritize. For example, many K-12 school districts now administer climate surveys to students and parents to help understand how conducive the local context is to academic achievement and attainment. Results from this study suggest that including two or three questions about college expectations and aspirations could be helpful when those districts attempt to identify which students need additional supports to be ready for a postsecondary education. Postsecondary institutions might benefit from polling their incoming students on how likely they think they are to complete the degree program and other topics related to their long-term outlook.

The prominence of variables assessing student aspirations and expectations in the models is noteworthy. Increasingly, K-12 schools and school systems are devoting resources to making students aware of college opportunities and developing an understanding of the college process (Karp, 2012). Though not causal, this study's findings are consistent with the rationale for programs that try to generate excitement among students about their college prospects. I also find that survey questions about aspirations (how far a student wants to go) and chances (how far a student actually thinks she will go) are both predictive. What a student wants and what a student deems realistic are both strong forecasters of college readiness.

Beyond individual measures, the four broad constructs identified by this study's factor analyses could be used by practitioners to identify different sources of risk, which could in turn be used to develop a more complete picture of students' readiness. For example, schools and colleges could benefit from finding ways to measure a student's level of risk in each of the four areas and combining results from those measures into an overall index of risk. That index could then be used to triage students most at risk of not attending college, as well as support more borderline students who may only show marginal risk in one or more areas. Measuring the level of risk in each area need not use complicated assessments or statistical techniques. For instance, a teacher, guidance counselor, or faculty advisor could informally talk to students about their expectations and whether their parents went to college (a proxy for socioeconomic status identified by this study) in order to supplement administrative data on student achievement.

These models can help identify students with the highest risk of not attending college as well as students who are more borderline and may not appear to be at risk.

Findings from this study provide actionable information for practitioners while the questions they raise are also useful. For example, results generate further questions about what, exactly, GPA measures, and how it relates to test scores. This study finds both GPA and test scores to be important forecasters of college readiness whereas other research has found test scores to be less predictive of outcomes like high school graduation when GPA is accounted for. The factor analyses presented in Table 1 show that grades are correlated both with measures of achievement and teacher perceptions of college readiness, and that the correlation with teacher perceptions tends to be higher. How much do grades measure content mastery versus teacher perceptions of readiness that extend beyond academic content? One potential avenue for additional research would be to use the data reduction techniques in this article to see what best predicts GPA so that guidance and admissions counselors have a better sense of what grades capture.

One should also note that grades and teacher perceptions of student readiness forecast college enrollment and persistence even when academic mastery, aspirations, and student background are accounted for in the model. There is no way to be sure whether teacher perceptions are predictive because teachers have information that these other measures dos

Table 1. *Lasso coefficients and factor loadings for best predictors*

Variables	NELS variable description	Lasso coeff.	Factor Gen.	F1	F2	F3	F4
F12XCOMP	STANDRDIZED TEST COMPOSITE (READING, MATH)	0.839	0.78	0.54			
F12XQURT	STANDARDIZED TEST QUARTILE (1=LOW)	0.477	0.76	0.52			
F12XRSTD	READING STANDARDIZED SCORE	0.589	0.71	0.52			
F12XMQ	MATHEMATICS QUARTILE (1=LOW)	0.689	0.72	0.48			
F12XSSTD	SCIENCE STANDARDIZED SCORE	0.410	0.61	0.48			
F12XHSTD	HISTORY/GEOGRAPHY STANDARDIZED SCORE	0.584	0.53	0.47			
BYGRADSQ	QUARTILE CODING OF GPA COMPOSITE	0.646	0.58	0.27			0.34
BYS81D	SOC. STUDIES GRADES FROM GRADE 6 UNTIL NOW	0.089	0.48	0.26			0.27
F1S70G	IMPORTANT TO HAVE STEADY BOY/GIRLFRIEND	0.098		0.25			
F1S24G	HOW MUCH COURSEWORK IN FOREIGN LANGUAGE	0.351	0.46	0.22	0.23		
F1S39B	DESCRIBE RESPONDENT'S ENGLISH GRADES	0.010	0.5	0.2			0.37
F1S70L	AMONG FRIENDS, HOW IMPORTANT TO HAVE JOB	0.126	0.21	0.2			-0.2
F1S52Bc6	HOW IMPORTANT IS FINANCIAL AID	0.316	0.27		0.79		
F1S52Fc6	HOW IMPORTANT ATTEND COLLEGE AND LIVE AT HOME	0.295	0.27		0.79		
F1S18B	RESPONDENT SURE TO FURTHER EDUCATION AFTER H.S	0.363	0.44		0.61		
F1S49	HOW FAR IN SCHOOL RESPONDENT THINKS HE WILL GET	1.501	0.56		0.59		
F1S51c2	DOES RESPONDENT PLAN TO GO TO COLLEGE AFTER H.S.	1.058	0.44		0.51		
BYPSEPLN	POST-SECONDARY EDUCATION PLANS	0.004	0.53		0.45	0.22	
BYS45	HOW FAR IN SCHOOL DO YOU THINK YOU WILL GET	0.470	0.53		0.45	0.22	
BYS47	HOW SURE RESPONDENT IS TO GO FURTHER THAN H.S.	0.470	0.42		0.43		
F1S48Ac2	HOW FAR IN SCHOOL FATHER WANTS RESPONDENT TO GO	0.074			0.26		
F1S64A	CHANCES THAT RESPONDENT WILL GRADUATE FROM H.S.	0.000	0.43		0.2		0.29
BYSESQ	SES QUARTILE	0.300	0.53			0.71	
BYPARED	PARENTS' HIGHEST EDUCATION LEVEL	0.233	0.53			0.66	
BYS34B	MOTHER'S HIGHEST LEVEL OF EDUCATION	0.001	0.46			0.61	
BYFAMINC	YEARLY FAMILY INCOME	0.289	0.44			0.59	
F1S93A	NO. BROTHER(S) LIVING IN SAME HOUSEHOLD	0.079				0.3	
BYS57C	RESPONDENT THREATENED AT SCHOOL	0.062				0.21	
F1S12F	FEEL IT'S OK TO GET INTO PHYSICAL FIGHTS	0.086				-0.22	
F1T1_22c2	STUDENT IS AT RISK OF DROPPING OUT H.S.	0.080	0.27				0.52
F1T1_4c1	STUDENT WILL PROBABLY GO TO COLLEGE - ELA TEACH	0.229	0.5				0.49
F1T5_4c1	STUDENT WILL PROBABLY GO TO COLLEGE - MATH TEACH	0.036	0.42				0.41
F1S28D	OFTEN FEEL CHALLENGED IN SCIENCE CLASS	0.001	0.2				0.4
F1S26Dc1	OFTEN ASKED TO SHOW UNDERSTAND SCIENCE	0.039					0.36
F1S39C	DESCRIBE RESPONDENT'S HISTORY GRADES	0.017	0.37				0.35
F1T5_22c1	STUDENT IS AT RISK OF DROPPING OUT H.S	0.041	0.2				0.34
F1T1_16c5	HOW OFTEN STUDENT IS ABSENT	0.000					0.31
F1T1_17c6	HOW OFTEN STUDENT IS TARDY	0.015					0.21
F1T5_2c2	STUDENT IS AT RISK OF DROPPING OUT H.S	0.055	0.23				0.2
BYS62c1	DID PARENTS WANT RESPONDENT TO TAKE ALGEBRA	0.011	0.34				
F1S36B2	TIME SPENT ON MATH HOMEWORK OUT OF SCHOOL	0.258	0.22				

Note: This table omits variables with loadings less than .2 on all measures. Loadings are left blank if the loading is less than .2 on some factors but not others. NELS variables with "BYS" and "FIS" prefixes are from student surveys in 8th and 10th-grade, respectively. NELS variables with "FIT1" and "FIT5" prefixes are from surveys of 10th-grade teachers in English/history and mathematics/science respectively.

Second, results show that the strongest predictors of college readiness appear to be measuring four broad categories: academic preparation, college goals and aspirations, teacher perceptions and evaluations of college readiness, and socioeconomic status, including parental education.

not capture or because the teachers have a hand in determining who goes to college and who does not. There is ample literature on self-fulfilling prophecies generated by teachers (Jussim, Eccles, & Madon, 1996; Madon, Jussim, & Eccles, 1997), and even some evidence that the effects of these self-fulfilling prophecies on postsecondary outcomes are modest (Soland, 2013). Still, findings highlight the need for more research on how much teachers control the postsecondary fates of their students and how those fates are influenced.

In sum, this study's results provide useful guidance for educators in the K-12 and higher education systems. First, decision rules can be established that identify students not on track for college with greater accuracy than in much of the prior research. These models can help identify students with the highest risk of not attending college as well as students who are more borderline and may not appear to be at risk. While decision tree methods are somewhat nuanced, they can be developed at the system level and are straightforward to interpret. Second, results show that the strongest predictors of college readiness appear to be measuring four broad categories: academic preparation, college goals and aspirations, teacher perceptions and evaluations of college readiness, and socioeconomic status, including parental education. Educators interested in identifying students for college readiness supports may benefit from basing determinations on measures of these constructs, whether formal or informal.

References

Archer, K. J., & Kimes, R. V. (2008). Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis, 52*(4), 2249–2260.

Baker, R. S., & Corbett, A. T. (2014). Assessment of robust learning with educational data mining. *Research & Practice in Assessment, 9*(2), 38–50.

Bettinger, E. P., Long, B. T., Oreopoulos, P., & Sanbonmatsu, L. (2012). The role of application assistance and information in college decisions: Results from the H&R block FAFSA experiment. *The Quarterly Journal of Economics, 127*(3), 1205–1242.

Bokossa, M. C., & Huang, G. (2001). *Imputation of test scores in the national education longitudinal study of 1988* (NELS: 88). US Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics.

Borsato, G. N., Nagaoka, J., & Foley, E. (2013). *College readiness indicator systems framework*. Stanford, CA: The John W. Gardner Center for Youth and Their Communities.

Bowers, A. J., Sprott, R., & Taff, S. A. (2013). Do we know who will drop out?: A review of the predictors of dropping out of high school: Precision, sensitivity, and specificity. *The High School Journal, 96*(2), 77–100.

Bradford, J. P., Kunz, C., Kohavi, R., Brunk, C., & Brodley, C. E. (1998). Pruning decision trees with misclassification costs. In *Machine Learning: ECML-98* (pp. 131–136). Berlin, Germany: Springer.

Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*(3).

Conley, D. T. (2005). College knowledge: Getting in is only half the battle. *Principal Leadership, 6*(1), 16–21.

Conley, D. T. (2007). The challenge of college readiness. *Educational Leadership, 64*(7), 23.

Conley, D. T. (2008). *College knowledge: What it really takes for students to succeed and what we can do to get them ready*. Hoboken, NJ: Wiley.

Conley, D. T. (2010). *College and career ready: Helping all students succeed beyond high school*. Hoboken, NJ: Wiley.

Cury, F., Elliot, A. J., Da Fonseca, D., & Moller, A. C. (2006). The social-cognitive model of achievement motivation and the 2-times-2 achievement goal framework. *Journal of Personality and Social Psychology, 90*(4), 666.

Dattatreya, G. R., & Sarma, V. V. S. (1981). Bayesian and decision tree approaches for pattern recognition including feature measurement costs. *IEEE transactions on pattern analysis and machine intelligence, 3*, 293–298.

Dekker, G., Pechenizkiy, M., & Vleeshouwers, J. (2009, July). *Predicting students drop out: A case study*. Paper presented at the International Conference on Education Data Mining, Cordoba, Spain.

Denley, T. (2014). How predictive analytics and choice architecture can improve student success. *Research & Practice in Assessment, 9*(2), 61–69.

Duckworth, A. L., & Quinn, P. D. (2009). Development and validation of the Short Grit Scale (GRIT–S). *Journal of Personality Assessment, 91*(2), 166–174.

Durlak, J. A., Weissberg, R. P., Dymnicki, A. B., Taylor, R. D., & Schellinger, K. B. (2011). The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions. *Child Development, 82*(1), 405–432.

Dweck, C. (2006). *Mindset: The new psychology of success*. New York, NY: Random House.

Dweck, C., Walton, G. M., & Cohen, G. L. (2011). Academic tenacity: Mindset and skills that promote long-term learning. *Seattle, WA: Bill & Melinda Gates Foundation*.

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology, 77*(4), 802–813.

Farrington, C. A., Roderick, M., Allensworth, E., Nagaoka, J., Keyes, T. S., Johnson, D. W., & Beechum, N. O. (2012). *Teaching adolescents to become learners: The role of noncognitive factors in shaping school performance–A critical literature review*. Chicago, IL: Consortium on Chicago School Research.

Fong, S., Si, Y.W., & Biuk-Aghai, R. P. (2009, December). *Applying a hybrid model of neural network and decision tree classifier for predicting university admission*. Paper presented at the International Conference on Information, Communications and Signal Processing, Macau, China. doi: 10.1109/ICICS.2009.5397665

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1). vol. 536 (2nd ed). Berlin, Germany: Springer.

Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine, 27*(15), 2865–2873.

Hauser, R. M., & Andrew, M. (2007). Reliability of student and parent reports of socioeconomic status in NELS-88. *Presented at the ITP Seminar, University of Wisconsin–Madison*.

Hooker, S., & Brand, B. (2010). College knowledge: A critical component of college and career readiness. *New Directions for Youth Development, 2010*(127), 75–85.

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics, 15*(3), 651–674.

Hurtado, S., & Carter, D. F. (1997). Effects of college transition and perceptions of the campus racial climate on Latino college students' sense of belonging. *Sociology of Education*, 324–345.

Jussim, L., Eccles, J., & Madon, S. (1996). Social perception, social stereotypes, and teacher expectations: Accuracy and the quest for the powerful self-fulfilling prophecy. *Advances in Experimental Social Psychology*, 28, 281–388.

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

Koh, K. H., & Zumbo, B. D. (2008). Multi-group confirmatory factor analysis for testing measurement invariance in mixed item format data. *Journal of Modern Applied Statistical Methods, 7*(2), 12.

Madon, S., Jussim, L., & Eccles, J. (1997). In search of the powerful self-fulfilling prophecy. *Journal of Personality and Social Psychology, 72*(4), 791.

Magee, J. F. (1964, July). *Decision trees for decision making*. Harvard Business Review.

Murthy, S. K. (1998). Automatic construction of decision trees from data: A multi-disciplinary survey. *Data Mining and Knowledge Discovery, 2*(4), 345–389.

Noonan, B. M., Sedlacek, W. E., & Veerasamy, S. (2005). Employing noncognitive variables in admitting and advising community college students. *Community College Journal of Research and Practice, 29*(6), 463–469.

O'Reilly, U.-M., & Veeramachaneni, K. (2014). Technology for mining the big data of MOOCs. *Research & Practice in Assessment, 9*(2), 29–37.

Perna, L. W. (2000). Differences in the decision to attend college among African Americans, Hispanics, and Whites. *Journal of Higher Education*, 117–141.

Quadri, M. M., & Kalyankar, N. V. (2010). Drop out feature of student data for academic performance using decision tree techniques. *Global Journal of Computer Science and Technology, 10*(2), 2-5.

Quinlan, J. R. (1990). Decision trees and decision-making. IEEE Transactions on Systems, *Man and Cybernetics*, 20(2), 339–346.

Revelle, W. (2011). *An introduction to psychometric theory with applications in R*. New York, NY: Springer.

Revelle, W., & Rocklin, T. (1979). Very simple structure: An alternative procedure for estimating the optimal number of interpretable factors. *Multivariate Behavioral Research, 14*(4), 403–414.

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika, 74*(1), 145–154.

Sedlacek, W. E. (2004). *Beyond the Big Test: Noncognitive Assessment in Higher Education*. Indianapolis, IN: Wiley.

Soland, J. (2013). Predicting high school graduation and college enrollment: Comparing early warning indicator data and teacher intuition. *Journal of Education for Students Placed at Risk (JESPAR), 18*(3–4), 233–262.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Tracey, T. J., & Sedlacek, W. E. (1984). Noncognitive variables in predicting academic success by race. *Measurement and Evaluation in Guidance,16*(4), 171–78.

Van Buuren, S., Brand, J. P., Groothuis-Oudshoorn, C. G. M., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation, 76*(12), 1049–1064.

Walton, G. M., & Cohen, G. L. (2007). A question of belonging: Race, social fit, and achievement. *Journal of Personality and Social Psychology, 92*(1), 82.

Walton, G. M., & Cohen, G. L. (2011). A brief social-belonging intervention improves academic and health outcomes of minority students. *Science, 331*(6023), 1447–1451.

York-Anderson, D. C., & Bowman, S. L. (1991). Assessing the college knowledge of first-generation and second-generation college students. *Journal of College Student Development, 32*(2), 116-122.

Abstract

Data-driven decision making is increasingly viewed as essential in a globally competitive society. Initiatives to augment standardized testing with performance-based assessment have increased as educators progressively respond to mandates for authentic measurement of student attainment. To meet this challenge, multidisciplinary rubrics were developed as a method of scoring student work samples. The current study utilized confirmatory factor analysis to examine ratings of student work (N = 245) using the Quantitative Literacy VALUE Rubric from the Association of American Colleges and Universities. The study examined a conceptual model of the six skill measures from the rubric to validate whether, taken together, they are reliable measures of a single general construct—Empirical and Quantitative Skill (EQS), a Texas Core Curriculum objective. The model confirmed that the six measures in the rubric (*Interpretation, Representation, Calculation, Application/Analysis, Assumptions, and Communication*) appeared to describe a single construct. Results support using the Quantitative Literacy VALUE Rubric for assessing EQS.

AUTHORS

Julie S. Gray, Ph.D.
The University of Texas at Arlington

Melissa A. Brown, M.Ed.
The University of Texas at Arlington

John P. Connolly, Ph.D.
The University of Texas at Arlington

Examining Construct Validity of the Quantitative Literacy VALUE Rubric in College-level STEM Assignments

An individual's quantitative literacy and competence with data evaluation is helpful in all areas of life, including academia. Because data-driven decision making is increasingly viewed as essential in a globally competitive society, educational objectives often emphasize learning outcome elements such as data analysis and how to use the data to draw conclusions. Data analysis without an understanding of the implications limits appropriate actions that can be taken by individuals and businesses (Tufte, 1997). Further, hiring managers seek individuals with empirical and quantitative skills because they have the ability to see connections and systemic problems (National Association of Colleges and Employers, 2016). Indeed, findings from the Spellings Commission panel stated that, "better data about real performance and lifelong working and learning ability is absolutely essential if we are to meet national needs and improve institutional performance" (U.S. Department of Education, 2006, p. 30).

Over the past two decades educational policies in the United States were changed by congressional legislation (e.g., No Child Left Behind Act [NCLB], 2001; Every Student Succeeds Act [ESSA], 2015). McGuinn (2006) maintains that the NCLB was implemented in response to public sentiment to hold educators accountable for the instruction students receive. More recently, initiatives to augment standardized testing with performance-based assessment (PBA) have increased as educators progressively respond to mandates for authentic measurement of student attainment. This progression is particularly reflected in the recently legislated ESSA (Gewertz, 2015), which is anticipated to go into full effect during the 2017–2018 academic year. The next section briefly reviews some of the policy implications for assessment professionals.

CORRESPONDENCE

Email
julie.gray@uta.edu

Impact of Policy Changes on Assessment Professionals

NCLB in particular affected the responsibilities of educational assessment professionals in requiring that each state must measure student progress for an academic year using single summative tests (Gewertz, 2015). As a result, a reliance upon standardized tests quickly developed to assess student attainment and inform process improvements in educational service delivery (Supovitz, 2009). Such testing often took the form of high-stakes, multiple-choice examinations. However, in the last decade, initiatives to extend assessment methods to include performance-based student work have gained momentum at many institutions (State Higher Education Executive Officers Association [SHEEO], 2016). As ESSA implementation moves toward completion, assessment professionals and state officials anticipate that it will provide them with options that include multiple measures during an academic year, including merging results from both standardized tests and performance-based tests (Gewertz, 2015). While many call the assessment of performance-based work a more authentic method of rating student attainment (Montgomery, 2002; Peden, Reed, & Wolfe, 2017; Rhodes, 2010; Rhodes & Finley, 2014), efforts to validate the way it is rated or scored present challenges for educators (Montgomery, 2002).

Unlike standardized tests, PBAs typically consist of written student work samples (e.g., essays, experimental or research lab summaries, and presentations).

PBA Challenges

PBA implies that in response to the assignment prompt, a student reveals the skills they have attained to date. That is, the student response contains authentic agreement between what the student knows and their ability to demonstrate that knowledge (Cobb, 2014). Unlike standardized tests, PBAs typically consist of written student work samples (e.g., essays, experimental or research lab summaries, and presentations). However, while PBA holds an advantage of authenticity it also presents a disadvantage. Montgomery (2002) lists concerns reported in the literature, including the difficulty of avoiding rater subjectivity when scoring authentic student work samples.

In contrast, normed scores for standardized tests for specific student populations typically guide comparisons based on equity and excellence. PBA often requires the introduction of a rubric to increase rater objectivity. Use of rubrics adds structure and consistency to the performance level assessment and comparisons (Montgomery, 2002).

VALUE Rubric Development as a Solution-Oriented Assessment Approach

Indeed, Montgomery (2002) recommended the use of rubrics for assessing authentic student work because they are tools that communicate to students the expected elements to include in the completed assignment. Rubrics for setting criteria and determining student attainment of the target objectives have been suggested to uphold equity and excellence for all students (Montgomery, 2002; Peden et al., 2017). That said, an evidence-based approach for evaluating PBAs using validated rubrics was needed.

A campus-based assessment initiative, led by the Association of American Colleges and Universities (AAC&U), published 16 Valid Assessment of Learning in Undergraduate Education (VALUE) Rubrics (AAC&U, 2017b). Faculty and other educational professionals gathered from over 100 different institutions of higher education, under the direction of the AAC&U, to develop the rubrics. The VALUE rubrics were designed to be scoring guides that can be used by universities to evaluate authentic student work samples. Further, the AAC&U outlined four families of Essential Learning Outcomes in order to advance VALUE rubrics as relevant assessment tools across a wide range of disciplines, courses, and objectives (National Leadership Council for Liberal Education & America's Promise, 2008).

The VALUE rubrics were designed to be scoring guides that can be used by universities to evaluate authentic student work samples.

These VALUE rubrics serve as a scaffold to government policies that endeavor to guarantee the quality of education across the United States for all students (AAC&U, 2017b). Though policies vary by state, they broadly included six educational objectives: critical thinking, communication, empirical and quantitative skill (EQS), teamwork, social responsibility, and personal responsibility. In the southwestern United States, the Texas Higher Education Coordinating Board (THECB) adopted the six aforementioned objectives for implementation in the most recent revision of the Texas Core Curriculum (TCC; THECB, 2011). The THECB required that all two-year and four-year educational institutions submit regular reports detailing the assessment practices and results for student TCC objective attainment within general education courses that have been approved and designated for inclusion in the TCC (THECB, 2011). Decision making regarding methodologies for rating these performance-based student work samples was left to the discretion of each institution by the THECB.

Preliminary studies supported by SHEEO and AAC&U consisted of a collaborative effort by 60 institutions in nine states who agreed to test the utility of the VALUE Rubric to rate authentic student work (SHEEO, 2016). In 2014–2015, they examined faculty ratings of authentic student work to determine levels that indicate healthy thresholds for student mastery (Lederman, 2015). While the multi-state collaborative vetted two rubrics in the practice of evaluating student work during its initial phase and current studies expanded to include more institutions, to date, they have not explored rubric construct validity. Studies are needed to investigate the extent to which the measures within each VALUE Rubric accurately represent a single construct.

Importance of Studying the Quantitative Literacy VALUE Rubric

Studies are needed to investigate the extent to which the measures within each VALUE Rubric accurately represent a single construct.

Case studies document the use of the VALUE Rubrics nationwide (AAC&U, 2017a; Peden et al., 2017). This study examines the construct validity of the AAC&U Quantitative Literacy VALUE Rubric for evaluating EQS, a TCC objective. EQS allows an individual to understand information or raw data that is presented in tables, charts, graphs, or figures and evaluate it to draw accurate conclusions. Identifying applications of EQS across academic disciplines is straightforward. The ability to take information, analyze it, and predict outcomes is a common theme in the hard sciences such as engineering, physics, chemistry, and biology. In addition, EQS is utilized across disciplines, for instance, in nursing, business, and psychology.

Individuals possessing skills such as EQS are in high demand because they can use this expertise to find evidence-based solutions. EQS is typically described using action verbs including identify, extract, validate, and report (Georgesen, 2015). Further, the process often follows an ordered set of action steps. For example, Georgesen (2015) extended the list as a set of four steps: 1) define, scope, identify, document; 2) extract, aggregate, transform, create; 3) develop, analyze, simulate, validate; and 4) report, recommend, implement, monitor. The extent to which these verbs can be translated into observable measures is essential to evaluating student attainment of the TCC objective EQS.

The current study focused on the measures within the Quantitative Literacy VALUE Rubric and its utility for measuring EQS. The six skill indicators measured by this rubric are *Interpretation, Representation, Calculation, Application/Analysis, Assumptions, and Communication*. Explanations for each are contained in the rubric (see Appendix). Our hypothesis is that there is a single underlying trait or "latent variable" of which the six different skills are indicators. In short, we wish to validate that the six different skills being assessed, taken together, are reliable measures of something more general.

Method

The current study focused on the measures within the Quantitative Literacy VALUE Rubric and its utility for measuring EQS.

The skills within the Quantitative Literacy VALUE Rubric were assessed using written samples of undergraduate student work from approved Signature Assignments embedded in the existing undergraduate TCC courses at a four-year public institution in an urban setting. The institution met requirements to serve as a Hispanic Serving Institution by the U.S. Department of Education (2016) and, importantly, received the R-1 designation by the Carnegie Classification of Institutions of Higher Education (2015), the definitive list for top doctoral research. The measurement of student attainment of EQS is of extreme interest because of the institutional focus on research.

Signature Assignments were designed to be authentic performance-based work in which students responded to pedagogically relevant prompts. For example, some Signature Assignments consisted of written summaries of actual lab experiments conducted by students in life and physical sciences courses. These papers, illustrated by tables and figures, essentially included measurable elements of *Interpretation, Representation, Calculation, Application/Analysis, Assumptions, and Communication*. All the Signature Assignments in this sample were collected from courses related to science, technology, engineering and math (STEM). Trained faculty and staff who participated in calibration and training exercises (described in more detail to follow) performed the ratings.

Participants

Signature Assignments were obtained from 296 undergraduates enrolled in core curriculum courses in STEM areas at the university. The readability of a portion of the assignments (n = 51) was poor because they were scanned copies of handwritten summaries from lab books or "blue books." As such, these 51 Signature Assignments were dropped from the sample and not rated. Ratings were available for 245 of the student Signature Assignments. Over half of the participants were female (61%; n = 149), which closely matched the gender ratio at the university. The sample also reflected a rich diversity of students. About a third of the student participants identified as White (33%; n = 80), almost a third identified as Hispanic (27%; n = 67), and the balance was split between African American; Asian; foreign, nonresident alien; multiple ethnicity; and unknown, not specified. Students represented nine of ten colleges and schools at the university (see Table 1).

Table 1. *Student Characteristics for the Rated Sample of Signature Assignments*

Categorical Variables	*N*	%
Gender		
Female	149	60.8
Male	96	39.2
Ethnicity		
African American	33	13.5
Asian	49	20.0
Caucasian	80	32.7
Foreign, nonresident alien	6	2.4
Hispanic	67	27.3
Multiple	5	2.0
Unknown, not specified	5	2.0
College/School		
College of Architecture	2	0.8
College of Business	24	9.8
College of Education	13	5.3
College of Engineering	15	6.1
College of Liberal Arts	26	10.6
College of Nursing	61	24.9
College of Science	61	24.9
School of Social Work	15	6.1
Undeclared	26	10.6
[a]Missing college or school information	2	0.8
Level		
Freshman	67	27.3
Sophomore	85	34.7
Junior	49	20.0
Senior and above	42	17.1
[a]Missing level information	2	0.8

Note: N = 245 for each of the categorical variable. [a] Information was missing

Procedure

Faculty currently teaching undergraduate courses in STEM areas agreed to submit the course set of authentic student work deemed as the Signature Assignment for this study. The syllabus for each core curriculum class at the university describes the Signature Assignment and the students enrolled in these courses complete it as they would any other assignment or required course work. The samples submitted for this assessment process were ungraded, de-identified copies. Steps to redact personal and academic information were followed for two reasons. The first was to prevent any bias among rater scores in response to the grade the paper received from the professor. The second was to protect the confidentiality of student, faculty, and course information.

Assessment Instrument

The Signature Assignments were assessed using the VALUE Rubric for Quantitative Literacy (AAC&U, 2009), which categorizes EQS into six measures: *Interpretation, Representation, Calculation, Application/Analysis, Assumptions, and Communication*. The rubric describes each measure and uses a four-point Likert scale for determining scores (see Appendix). Higher values indicate more evidence of EQS. Using the rubric, raters assigned a score to each of the six skill measures.

Typically, in student samples, the six measures are adequately represented in the narrative of the Signature Assignment. It is important to note that visual communication in the form of charts, graphs, and figures enhanced the identification of the *Representation and Communication* measures. This is not unexpected because communication (written and visual) is required for fleshing out and articulating ideas in STEM areas. Visual communication is particularly important, and in many cases essential, for depicting information in STEM areas.

Raters, Rater Calibration, and Scoring

EQS allows an individual to understand information or raw data that is presented in tables, charts, graphs, or figures and evaluate it to draw accurate conclusions.

For the purposes of this study, the unit of analysis was an individual rater's score for a particular Signature Assignment. Raters scored the student writing samples during a scheduled scoring day so each paper was read and then rated by at least two separate raters working independently in a group setting. The rater group included ten faculty members and professional staff with advanced degrees. Scoring day began with an orientation and description of the rating process. Then, the entire group read one anchor paper chosen by the facilitator. Next, the facilitator led a discussion focused on reaching a common understanding of the EQS measures and finding exemplar indicators within the anchor paper for the rubric's levels of mastery. Then the rating process began and raters individually read their assigned papers to score each measure with the rubric (four-point Likert scale). Two raters independently rated each paper. Measure scores were calculated as the average of both scores. The facilitator checked each paper, after the completion of the two ratings, to review whether disagreement between measure ratings exceeded acceptable metrics. If so, the facilitator assigned a third rater as a separate, impartial mediator. In those cases (n = 4) the outlier of the three ratings was replaced.

Inter-rater Agreement

To examine the agreement between raters, an estimate of inter-rater reliability was calculated to see how frequently the rater pairs agreed on the score when rating the same paper. Conclusions about the consistent measurement of the six measures depend on this estimate. A calculation of the intraclass correlation coefficient (ICC) was used to determine the level of inter-rater agreement. High ICC values indicate more agreement between raters. A one-way random model was used to measure consistency within the mean measure values. ICC values for *Interpretation, Representation, Calculation, Application/Analysis, Assumptions, and Communication* indicated good inter-rater agreement (see Table 2) even though rater pairs varied across ratings, which typically results in lower ICC values (Landers, 2015).

Table 2. *ICC Values by Measure*

Measure	ICC Value
Interpretation	.52
Representation	.51
Calculation	.47
Application/Analysis	.56
Assumptions	.51
Communication	.60

Note: N = 245 for each measure.

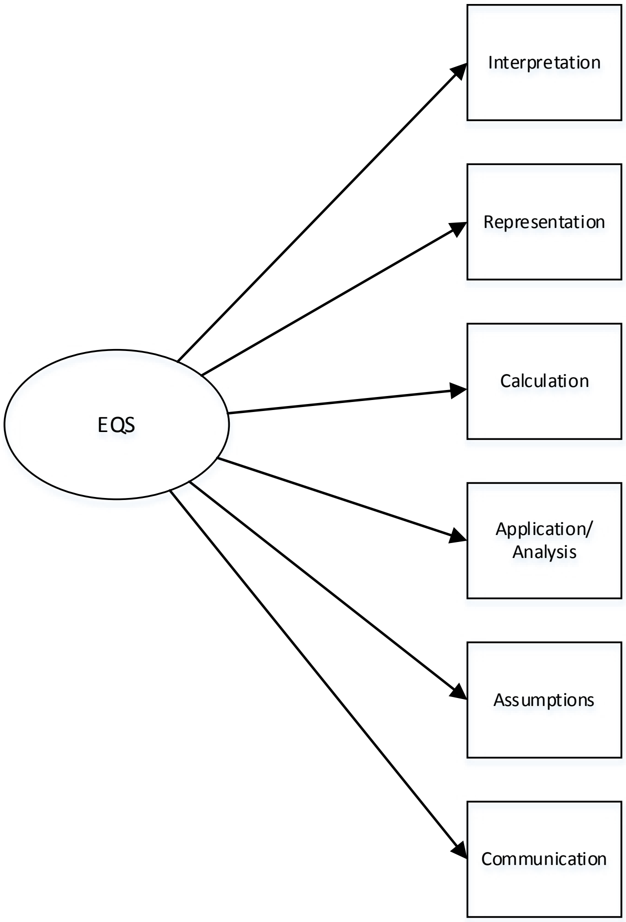


Figure 1. Conceptual Model of Underlying EQS Traits

Analysis Plan

We used confirmatory factor analysis to assess whether the six measured skills are reliable indicators of an underlying more general construct (Brown 2006). One key advantage of this approach is the ability to isolate the underlying construct from random error variance in the indicator measures. Further, correlations across the error components of each survey item can also be modeled to account for method effects that detract from the underlying construct, such as any tendency to rate two of the skills more similarly than the others. Figure 1 depicts the conceptual model (H_0).

Signature Assignments were designed to be authentic performance-based work in which students responded to pedagogically relevant prompts.

Because the measure ratings are in the form of a Likert scale, and therefore categorical, we used a mean- and variance-adjusted weighted least squares (WLSMV) estimator to estimate the loadings of each measure on the underlying EQS trait (Muthén & Muthén, 1998–2012).

The same estimator also yields fit statistics that provide information on the overall reliability of the model in terms of its ability to reproduce the variances and covariances of the indicator measures. Ideally, the model reports a nonsignificant chi square value indicating that imposing the hypothesized structure on the data does not amount to a substantial loss of information. However, since chi-square statistics are proportional to sample size other statistics are commonly used to assess model fit. In particular, a Root Mean Square Error of Approximation (RMSEA) statistic that is below 0.05 and a Comparative Fit Index (CFI) greater than 0.95 indicates a model that is a good fit to the data (Byrne, 2012).

Results

All the analyses were conducted in Mplus v.7.31 (Muthén & Muthén, 2012), which also reports ways of improving the model via modification indices. Analysis of the set of ratings from rater 1 and then the set from rater 2 (from the rater parings) indicated that

Current efforts toward the use of PBA to augment standardized testing with students present a challenge for educators because of the possible rater bias and other differences in scoring authentic student work; thus, there is a need to validate the rubrics that raters use.

significant model improvement would be obtained by allowing the random error variances in the *Representation and Calculation* measures to correlate. The fit statistics of the two models, i.e., the model with the specified error correlation (the H_1 model) and the model with no error correlations (the H_0 model), are summarized in Table 3. The H_1 model met all the criteria of a well-fitting model in terms of the key fit statistics: chi square, RMSEA, and CFI. The table also showed a significant loss of fit for the H_0 model in terms of a chi-square difference test.

The unstandardized loadings of each of the six skill measures on the underlying EQS latent variable are summarized in Table 4. The standardized estimates, along with associated standard errors, are shown in Figure 2. Also included in Figure 2 is the estimate for the error correlation between *Representation and Calculation*.

Table 3. *Model fit statistics for the H_1 and H_0 models with X^2 difference test*

	N	χ^2	*df*	P-Value	RMSEA	CFI
H_1 Model	245	9.31	8	0.317	0.03	0.99
H_0 Model	245	47.31	9	0.000	0.13	0.98
Difference Test		19.40	1	0.000		

The estimates in Table 4 are akin to regression estimates of the effect of the underlying EQS trait on the skill in question—all of which were statistically significant at the 0.01 alpha level. The three strongest indicators were *Communication, Application/Analysis, and Interpretation*, and the amount of variance in these indicators explained by EQS is 77%, 73%, and 71%, respectively. Weaker effects were found in the case of *Calculation* (53%), *Assumptions* (47%), and *Representation* (42%).

Table 4. *Weighted Least Squares estimates for the six skill measures*

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
Interpretation[a]	1.00	0.00		
Representation	0.77	0.07	11.20	0.000
Calculation	0.87	0.07	12.72	0.000
Application/Analysis	1.02	0.06	17.62	0.000
Assumptions	0.81	0.06	12.91	0.000
Communication	1.04	0.06	18.98	0.000

Note: [a]For the purpose of scaling the latent variable, *Interpretation* is treated as the marker indicator. As such, the associated loading of this indicator on EQS is set at a value of one (Brown, 2006, p.71).

Discussion

Current efforts toward the use of PBA to augment standardized testing with students present a challenge for educators because of the possible rater bias and other differences in scoring authentic student work; thus, there is a need to validate the rubrics that raters use. The goal of the current study was to examine the construct validity of the Quantitative Literacy VALUE Rubric, one of 16 rubrics developed by the AAC&U. Overall, the findings show that the six underlying skill measures tapped into a common underlying EQS trait. These results extend previous research that has primarily focused on the use of the rubrics to study trends in student attainment (SHEEO, 2016).

Summary of Findings

Our hypothesis-testing results suggested that the six measures each reflect EQS as an underlying trait and that raters using the rubric produced valid EQS scores. Significant consistency was confirmed by analyzing rubric ratings of authentic student work from

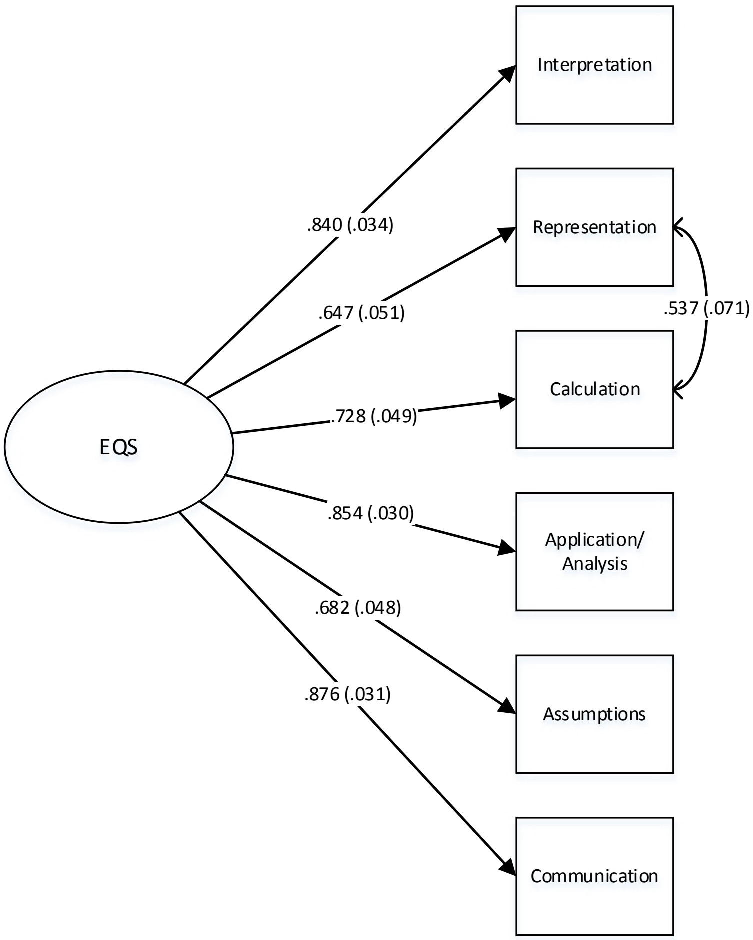


Figure 2. Standardized estimates for final solution (H_1)

STEM courses at a four-year public university. Each of the six measured skills loaded on the same construct and the model accounted for a large proportion of variance in each of the indicators. This validates that the six different skills assessed by the Quantitative Literacy VALUE Rubric are reliable measures of the general trait, EQS. Though the importance of nonsubjective measures of PBA has been well established, to our knowledge this is the first study that confirmed how well the measured skills in the Quantitative Literacy VALUE Rubric fit together as a model of EQS.

The goal of the current study was to examine the construct validity of the Quantitative Literacy VALUE Rubric, one of 16 rubrics developed by the AAC&U.

In addition, the pattern of results indicated three measures with very strong contributions to the model, *Interpretation, Application/Analysis, and Communication*. These three skill measures are widely used in statistical texts to describe the analytical process researchers use after research questions are posed, studies are designed, and data are collected. Without them, the research process is just a collection of numbers, and does not contribute answers to research questions that often have real consequence in many fields. Indeed, national surveys of employers repeatedly list skills involving *Interpretation, Application, and Communication* as essential qualities in job applicants (National Association of Colleges and Employers, 2016). The model confirmed the strength of the rubric in representing these highly marketable skills—those that are involved in quantitative literacy.

In further support for the model, analyses revealed inter-rater reliability estimates in the moderate to good range for the six measures. This suggests that rater calibration activities conducted on scoring day may have held a degree of utility in terms of promoting agreement among raters. The literature about VALUE rubrics contains many case studies of the use of calibration as a best practice (AAC&U, 2017a; Finley, 2011; Peden et al., 2017) yet, to our knowledge, it does not contain findings related to calibration activity effectiveness that directly compared a trained group of raters with a group that did not undergo any sort of training.

This validates that the six different skills assessed by the Quantitative Literacy VALUE Rubric are reliable measures of the general trait, EQS.

In addition, while inter-rater agreement may have differed with the introduction of more than two raters for all Signature Assignments, the study design accounted for the importance of good inter-rater agreement by planning the facilitator-led calibration activities and using a third rater to mediate unacceptable differences. Indeed, Stanny, Gonzalez, and McGowan (2015) mention improvement in rater agreement through the use of similar activities that operationalize rubric guidelines with "notes [added to the rubric] about difficult decisions, to build and maintain consensus for future decisions" (p. 905). Further, Finley (2011) recommends that rating sessions include the type of facilitator-led discussions that were used in this study before the application of the rubric to ensure adequate agreement. Though not a primary focus of the current study, findings suggested that the level of agreement for the ratings in the sample provided adequate justification for proceeding with the analysis of the rubric's construct validity.

In addition to strengths already mentioned, the model improved when the association between *Calculation* and *Representation* was allowed to covary. This makes sense because a single-minded focus on *Calculation* makes drawing conclusions hard to visualize and a skill such as *Representation* strengthens its meaning. In that way, *Calculation* and *Representation* dovetail together. In practice, calculation turns to representation to derive meaning and understanding as two parts of the same whole. In the process of problem solving, making a visual representation is a natural process for deriving meaning from computational problems (Van Garderen & Montague, 2003) and for enhancing the decision-making value of quantitative information (Tufte, 1997).

Continued efforts are needed to promote the use of authentic student work in educational assessment.

Limitations

The findings of the current study are promising but a few limitations should be noted. For instance, student samples only represented STEM courses in the life and physical sciences. This limited the ability to examine the independent effects of other types of courses and potential confounds. In future studies, course types should be extended to include all three of the foundational component areas required by the THECB (life and physical science, mathematics, and social and behavioral science). Though all students at the university were also required to take courses across eight foundational component areas as part of the TCC, conclusions would be strengthened through the incorporation of a wider range of courses. Additionally, performance-based work was gathered only from TCC-approved courses and the naturalistic design of the study did not allow for randomized assignment of papers from across all the STEM courses on campus regardless of level. Nonetheless, the student demographics suggest that the sample was consistent with the campus population as a whole.

Conclusion

Continued efforts are needed to promote the use of authentic student work in educational assessment. This study examined a widely utilized rubric using a relatively large sample of STEM assignments to capitalize on the strength of the AAC&U initiatives that measure student attainment of broadly accepted educational learning objectives. Results suggest that the six skill measures contained in the Quantitative Literacy VALUE Rubric fit together well to explain EQS. Consequently, efforts to promote VALUE rubrics have the potential to accurately measure student attainment of EQS. Further research is needed to confirm the construct validity of the full array of AAC&U VALUE Rubrics. Continuation of this line of inquiry is essential for maximizing the effectiveness of PBA.

Keywords: quantitative literacy, empirical and quantitative skill, VALUE rubric, STEM, EQS, performance-based assessment, Texas Core Curriculum, AAC&U

Appendix

Quantitative Literacy VALUE Rubric

for more information, please contact value@aacu.org

Association of American Colleges and Universities

Definition

Quantitative Literacy (QL) – also known as Numeracy or Quantitative Reasoning (QR) – is a "habit of mind," competency, and comfort in working with numerical data. Individuals with strong QL skills possess the ability to reason and solve quantitative problems from a wide array of authentic contexts and everyday life situations. They understand and can create sophisticated arguments supported by quantitative evidence and they can clearly communicate those arguments in a variety of formats (using words, tables, graphs, mathematical equations, etc., as appropriate).

Evaluators are encouraged to assign a zero to any work sample or collection of work that does not meet benchmark (cell one) level performance.

	Capstone 4	Milestones 3	2	1
Interpretation *Ability to explain information presented in mathematical forms (e.g., equations, graphs, diagrams, tables, words)*	Provides accurate explanations of information presented in mathematical forms. Makes appropriate inferences based on that information. *For example, accurately explains the trend data shown in a graph and makes reasonable predictions regarding what the data suggest about future events.*	Provides accurate explanations of information presented in mathematical forms. *For instance, accurately explains the trend data shown in a graph.*	Provides somewhat accurate explanations of information presented in mathematical forms, but occasionally makes minor errors related to computations or units. *For instance, accurately explains trend data shown in a graph, but may miscalculate the slope of the trend line.*	Attempts to explain information presented in mathematical forms, but draws incorrect conclusions about what the information means. *For example, attempts to explain the trend data shown in a graph, but will frequently misinterpret the nature of that trend, perhaps by confusing positive and negative trends.*
Representation *Ability to convert relevant information into various mathematical forms (e.g., equations, graphs, diagrams, tables, words)*	Skillfully converts relevant information into an insightful mathematical portrayal in a way that contributes to a further or deeper understanding.	Competently converts relevant information into an appropriate and desired mathematical portrayal.	Completes conversion of information but resulting mathematical portrayal is only partially appropriate or accurate.	Completes conversion of information but resulting mathematical portrayal is inappropriate or inaccurate.
Calculation	Calculations attempted are essentially all successful and sufficiently comprehensive to solve the problem. Calculations are also presented elegantly (clearly, concisely, etc.)	Calculations attempted are essentially all successful and sufficiently comprehensive to solve the problem.	Calculations attempted are either unsuccessful or represent only a portion of the calculations required to comprehensively solve the problem.	Calculations are attempted but are both unsuccessful and are not comprehensive.
Application / Analysis *Ability to make judgments and draw appropriate conclusions based on the quantitative analysis of data, while recognizing the limits of this analysis*	Uses the quantitative analysis of data as the basis for deep and thoughtful judgments, drawing insightful, carefully qualified conclusions from this work.	Uses the quantitative analysis of data as the basis for competent judgments, drawing reasonable and appropriately qualified conclusions from this work.	Uses the quantitative analysis of data as the basis for workmanlike (without inspiration or nuance, ordinary) judgments, drawing plausible conclusions from this work.	Uses the quantitative analysis of data as the basis for tentative, basic judgments, although is hesitant or uncertain about drawing conclusions from this work.
Assumptions *Ability to make and evaluate important assumptions in estimation, modeling, and data analysis*	Explicitly describes assumptions and provides compelling rationale for why each assumption is appropriate. Shows awareness that confidence in final conclusions is limited by the accuracy of the assumptions.	Explicitly describes assumptions and provides compelling rationale for why assumptions are appropriate.	Explicitly describes assumptions.	Attempts to describe assumptions.
Communication *Expressing quantitative evidence in support of the argument or purpose of the work (in terms of what evidence is used and how it is formatted, presented, and contextualized)*	Uses quantitative information in connection with the argument or purpose of the work, presents it in an effective format, and explicates it with consistently high quality.	Uses quantitative information in connection with the argument or purpose of the work, though data may be presented in a less than completely effective format or some parts of the explication may be uneven.	Uses quantitative information, but does not effectively connect it to the argument or purpose of the work.	Presents an argument for which quantitative evidence is pertinent, but does not provide adequate explicit numerical support. (May use quasi-quantitative words such as "many," "few," "increasing," "small," and the like in place of actual quantities.)

References

Association of American Colleges and Universities. (2009). *Quantitative literacy VALUE rubric*. Retrieved from https://www.aacu.org/value/rubrics/quantitative literacy

Association of American Colleges and Universities. (2017a). *Campus models and case studies*. Retrieved from https://www.aacu.org/campus-model/3305

Association of American Colleges and Universities. (2017b). *VALUE rubrics*. Retrieved from https://www.aacu.org/value-rubrics

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York and London: The Guilford Press.

Byrne, B. M. (2012). *Structural equation modeling with Mplus: Basic concepts, applications and programming*. New York and London: Routledge.

Carnegie Classification of Institutions of Higher Education. (2015). *Classification update: List of R-1 doctoral universities*. Retrieved from http://carnegieclassifications.iu.edu

Cobb, R. (2014). *The paradox of authenticity in a globalized world*. New York: Palgrave Macmillan.

Finley, A. P. (2011). How reliable are the VALUE rubrics? *Peer Review, (14)*1, 31–33.

Georgesen, J. (2015). *Evolving from big data to smart data: New ways CX researchers predict customer behavior*. Retrieved from http://mrweek.com/content

Gewertz, C. (2015). ESSA's flexibility on assessment elicits qualms from testing experts. *Education Week*. Retrieved from http://www.edweek.org/ew/articles/2015/12/21/essas-flexibility-on-assessment-elicits-qualms-from.html

Landers, R. N. (2015). Computing intraclass correlations (ICC) as estimates of interrater reliability in SPSS. *The Winnower*. doi:10.15200/winn.143518.81744

Lederman, D. (2015). New effort aims to standardize faculty-driven review of student work. *Inside Higher Ed*. Retrieved from https://www.insidehighered.com/news/2015/09/25/new-effort-aims-standardize-faculty-drivenreview-student-work

McGuinn, P. J. (2006). *No Child Left Behind and the transformation of federal education policy, 1965–2005*. Lawrence, KS: University Press of Kansas.

Montgomery, K. (2002). Authentic tasks and rubrics: going beyond traditional assessments in college teaching. *College Teaching, (50)*1, 34–40. doi:10.1080/87567550209595870

Muthén, L.K. & Muthén, B.O. (1998-2012). *Mplus user's guide*. (7th ed.). Los Angeles, CA: Author.

National Association of Colleges and Employers. (2016). *Job Outlook 2016*. Bethlehem, PA.

National Leadership Council for Liberal Education & America's Promise. (2008). *College learning for the new global century*. Washington, DC: Association of American Colleges and Universities.

Peden, W., Reed, S., & Wolfe, K. (2017). *Rising to the LEAP challenge: Case studies of integrative pathways to student work*. Washington, DC: Association of American Colleges and Universities.

Rhodes, T. (Ed.). (2010). *Assessing outcomes and improving achievement: Tips and tools for using rubrics*. Washington, DC: Association of American Colleges and Universities.

Rhodes, T. & Finley, A. (2014). *The VALUE rubrics: Frequently asked questions about development, interpretation, and use of rubrics on campuses*. Retrieved from http://www.sheeo.org/sites/default/files/project-files/VALUERubrics_Webinar %28R%29.pptx

Stanny, C., Gonzalez, M., & McGowan, B. (2015). Assessing the culture of teaching and learning through a syllabus review. *Assessment & Evaluation in Higher Education, 40*(7), 898–913. doi:10.1080/02602938.2014.956684

State Higher Education Executive Officers Association. (2016). *MSC: A multi-state collaborative to advance learning outcomes assessment*. Retrieved from http://www.sheeo.org/projects/msc-multi-state-collaborative-advance-learning-outcomes-assessment#

Supovitz, J. (2009). Can high stakes testing leverage educational improvement? Prospects from the last decade of testing and accountability reform. *Journal of Educational Change, 10*(2–3), 211–227. Retrieved from http://link.springer.com/journal/10833

Texas Higher Education Coordinating Board. (2011). *Texas Core Curriculum*. Retrieved from http://www.thecb.state.tx.us/index.cfm?objectid=417252EA-B240-62F79F6A1A125C83BE08

Tufte, E. R. (1997). *Visual explanations: images and quantities, evidence and narrative*. Cheshire, CT: Graphics Press.

U.S. Department of Education. (2006). *A test of leadership: Charting the future of U.S. higher education*. Washington, D. C.

U.S. Department of Education. (2016). *FY 2016 eligible Hispanic-serving institutions*. Retrieved from https://www2.ed.gov/about/offices/list/ope/idues/hsi-eligibles-2016.pdf

Van Garderen, D., & Montague, M. (2003). Visual-spatial representation, mathematical problem solving, and students of varying abilities. *Learning Disabilities Research & Practice, 18*(4), 246. doi:10.1111/1540-5826.00079

Abstract

Academic outcomes assessment in student affairs is integral for both service improvement and demonstrating the unit's value to the university's academic mission. However, identifying the right measures is challenging. We implemented three common measures (pre-post self-reported academic functioning, retrospective perceptions of service impact, and semester grades) within a single counseling center client sample (N = 368) and examined the impact of measure selection on the representativeness of client subsamples and the conclusions that might be drawn about service effectiveness. Students' perceptions of academic outcomes suggested greater impact than pre-post or grade measures overall but all three showed positive effects for clients identified as academically at-risk at baseline. No single measure captured a fully representative sample of clients. Rather than providing evidence for one "best" measure, results point to the importance of using multiple measures to assess academic outcomes. Implications for best practices in service outcomes assessment are discussed.

AUTHORS

Stacy J. Priniski, M.S.
University of Wisconsin-Madison

Erin Winterrowd, Ph.D.
Regis University

Proof in the Pudding: Implications of Measure Selection in Academic Outcomes Assessment

The American College Personnel Association's (1994) release of the Student Learning Imperative (re)sparked a dedication to improving assessment practices and a corresponding call to document the impact of student affairs services on student learning and development (Reynolds & Chris, 2008; Upcraft & Schuh, 1996). Pressure for student affairs divisions to demonstrate their value to the university's academic mission has increased simultaneously (Nafziger, Couillard, & Smith, 1999; Varlotta, 2012). However, with a plethora of assessment approaches and measures available it can be difficult to determine the best way to assess service impact. In this article we explore the ways in which practitioners from one unit (campus mental health services) have measured academic outcomes, and we investigate how measure selection influences conclusions about service effectiveness.

Outcomes Assessment in Campus Counseling Centers

Campus counseling centers (CCCs) provide a useful context for studying academic outcomes assessment for two reasons. First, the mechanisms by which CCC services might influence academic functioning are evident in the literature. CCCs improve students' psychological well-being (e.g., Minami et al., 2009), and psychological well-being is an important predictor of academic well-being (e.g., Miller & Markman, 2007; Stallman, 2010). Second, literature on the relationship between counseling and academics includes a variety of academic outcome measures with mixed results that give insight into the potential differences among them.

CORRESPONDENCE

Email
spriniski@wisc.edu

Lambert and Hawkins' (2004) conceptual model of CCC assessment provides a useful framework for measure selection. The model characterizes outcome measures by content (the construct of interest), source (e.g., client, therapist), method of data collection (e.g., self-report, behavioral), and time orientation (e.g., state vs. trait measures). Importantly, the model considers psychometric strength (reliability, validity, sensitivity to change), applicability, and practicality, emphasizing that not all measures are equally suited to capture a given outcome. In light of this model, we review three commonly used measures of service impact on academic functioning: two self-report measures (pre-post self-reports of academic functioning and retrospective self-reports of counseling's impact) and one institutional measure (grades), and then report a field test of those measures within a single CCC sample.

Assessing Academic Outcomes

Pre-post self-reports. A common practice for CCCs is to assess the impact of services on academic functioning using pre-post measurements of school-related "symptoms" (e.g., difficulty keeping up with schoolwork, thoughts of leaving college). These measures fit seamlessly into existing assessments of self-reported psychological symptoms at most CCCs, and a number of validated questionnaires that include academic functioning are available (e.g., Counseling Center Assessment of Psychological Symptoms (CCAPS); Locke et al., 2011). Studies that assess academic outcomes using pre-post measures generally find that clients' academic functioning improves over a set number of appointments (six, on average) whereas academic functioning remains unchanged over a similar time period among non-clients (DeStefano, Mellot, & Petersen, 2001; Lockard, Hayes, McAleavey, & Locke, 2012; Nafziger et al., 1999).

In this article we explore the ways in which practitioners from one unit (campus mental health services) have measured academic outcomes, and we investigate how measure selection influences conclusions about service effectiveness.

Retrospective self-reports. Another commonly used approach is to ask students directly whether they feel services helped them academically, typically after a set number of appointments or at the end of the semester. These retrospective self-report measures are often created in-house so most have not been validated in the literature. However, existing studies suggest that they are internally reliable and strongly correlated with other learning outcomes of counseling (Winterrowd, Priniski, Achter, & Abhold, 2016) and may be better tailored to specific student affairs units (Erwin & Sivo, 2001). National surveys of CCC directors indicate that over 60% of centers collect these measures and most find that clients report that counseling has a positive impact on their academic functioning (Gallagher, 2011) with the few published studies also supporting that conclusion (Winterrowd, et al., 2016; Reynolds & Chris, 2008; Turner & Berry, 2000).

Grade point average. Grade point average (GPA) information can be asked from students directly or taken from institutional records, with the latter being more common in the literature. Researchers typically measure changes in GPA from before service delivery to after, or test the relationship between extent of participation in services (e.g., number of appointments) and grades, controlling for prior performance (e.g., high school GPA). As an academic outcome of student services, GPA resonates with campus administration and students alike, although it is unclear whether GPA is sensitive to the kinds of changes that counseling is intended to create (Illovsky, 1997; Lockard et al., 2012). Most studies show null effects (Lee, Olson, Lock, Michelson, & Odes, 2009; Illovsky, 1997), although studies examining the impact of counseling for academically "at-risk" students (underprepared first-year students, Cholewa & Ramaswami, 2015; students on academic probation, Wlazelek & Coulter, 1999) find positive impacts of counseling on GPA.

Summary

Together, this body of literature highlights three key points about measure selection. First, researchers and practitioners can and do choose from a wide variety of academic measures to assess service impact. Second, the methods of data collection employed with each outcome measure differ in ways that impact the sample of clients evaluated. For example, pre-post self-reports can only be collected from clients who attend a certain number of appointments, and change in GPA can only be collected from students with continuous enrollment in credit-bearing courses. It is unclear whether different measures of academic outcomes capture a representative subsample of clients—which raises concerns about the validity of the conclusions that are drawn from them. Finally, it appears that different measures lead to different conclusions about the relationship between counseling and academics. Specifically, studies using pre-post self-report measures (e.g., Lockard et al., 2012) and retrospective self-reports (e.g., Winterrowd, et al., 2016) found positive effects of counseling on academic functioning whereas the results of studies that utilized GPA had mixed results (e.g., Cholewa & Ramaswami, 2015; Lee et al., 2009).

The Lambert and Hawkins (2004) model emphasizes the importance of measure selection and the dimensions upon which measures can vary. Implicit in their discussion is the idea that diverse measures can result in diverse conclusions about service effectiveness, highlighting the importance of using multiple measures of learning outcomes in counseling and in higher education generally (e.g., Astin & antonio, 2012; Schuh, 2011; Suskie, 2009). The mixed results of the previous studies appear to support that assertion. However, because pre-post self-reports, retrospective reports, and grades have never been compared within a single sample, it is unclear whether inconsistent results in the literature reviewed here are due to differences in the measures themselves or other factors (e.g., differences in samples, timeframe of assessment, quality of services provided).

Current Study

In the current study we investigated the impact of measure selection in service evaluation using a comprehensive framework for assessment of academic outcomes within a single counseling center client sample.

In the current study we investigated the impact of measure selection in service evaluation using a comprehensive framework for assessment of academic outcomes within a single counseling center client sample. We compared the impact of counseling services indicated by three commonly used measures of academic outcomes—pre-post self-reports of academic functioning, retrospective reports of counseling impact on academics, and semester grades—for all clients, generally, and for clients identified at baseline as academically at-risk in particular (see Table 1 for a summary of measure features). In line with the Lambert and Hawkins (2004) model, the validity of the measures and the practical implications of measure selection were also of interest. Therefore, we examined differences in subsample characteristics to determine whether each measure captured a representative sample of the client population.

Table 1. *Characteristics of Three Common Measures of Academic Outcomes on the Dimensions of the Lambert and Hawkins (2004) Model*.

	Pre-Post Self-Reports	Retrospective Self-Reports	Grade Point Average (GPA)
Content	changes in academic functioning over the course of counseling	perceptions of counseling's impact on academic functioning	course performance
Source	student	student	university records
Method of data collection	self-report	self-report	institutional
Time orientation	state measure; varies day to day	state measure; varies day to day	trait measure; varies semester to semester
Psychometrics	several validated measures available; sensitive to change	typically in-house measures (not validated), but existing data suggests reliability and validity; sensitive to change	highly externally valid; less sensitive to change
Applicability	high	high	high

We hypothesized that pre-post self-reports, retrospective self-reports, and grades would each yield unique results within a single sample of counseling center students. Such results would suggest that the mixed findings of prior studies might be due to differences in the measures themselves and demonstrate the importance of measure selection in assessing the precise aspects of academic functioning each student service intends to support.

Method

Participants

Data were collected from 368 undergraduate students who received counseling services during the fall semester at a midsize predominantly undergraduate institution in the Midwest. Participants identified as White (89%), African American/Black (4%), Asian American/Asian (2%), Multiracial (2%), American Indian or Alaskan Native (1%), Hispanic/Latino(a) (1%), or other self-identified ethnicities (1%; 0.3% unreported). Women were 69% of the sample, men 30%, transgender individuals 0.3%, and other self-identified genders 0.5%. Participants were 25% first-year students, 25% sophomores, 20% juniors, and 29% seniors (1% unreported), with a mean age of 21.12 (*SD* = 3.98; 2% unreported).

Measures

Demographics and Presenting Concerns. From the counseling center intake paperwork we collected information about participating clients' gender, race/ethnicity, age, and year in school. We also noted whether clients selected "school and grades" as one of their reasons for seeking counseling services (on a 29-item presenting concerns checklist). This was used as a baseline measure of academic functioning and one indicator of being academically at-risk.

Counseling Center Assessment of Psychological Symptoms- Academic Distress Scale. The counseling center administered the long form of the Counseling Center Assessment of Psychological Symptoms (CCAPS-62; Locke et al., 2011) at intake and the short form (CCAPS-34; Locke et al., 2012) at the fifth appointment. The CCAPS is a self-report questionnaire that measures changes in psychological well-being generally and across various mental health subscales. We used the four-item Academic Distress Subscale to provide a baseline measure of academic functioning as well as examine changes in academic distress ("It's hard to stay motivated for my classes," "I am not able to concentrate as well as usual," "I feel confident that I can succeed academically" (reversed), and "I am unable to keep up with my schoolwork"). Participants respond via a four-point Likert-type scale with subscale scores > 2.75 considered "elevated," an indicator of being academically at-risk. The measure had high internal consistency in this sample (α = .82 at intake, .83 at fifth appointment).

Learning Outcomes and Satisfaction Survey- Academic Outcomes Scale. The Learning Outcomes and Satisfaction Survey (LOS; Winterrowd, et al., 2016) measures client perceptions of counseling outcomes and satisfaction with services. The Academic Outcomes (AO) scale assesses the extent to which clients feel counseling helps their academics, with four items ("Counseling has helped with my academic performance," "Counseling has increased my academic motivation and/or attendance," "Counseling has helped me to focus better on my academics," and "Counseling has helped me stay at school") scored on a five-point Likert-type scale. The scale had high internal consistency in this sample (α = .83).

Grade Point Average. Participants' semester grade point averages (GPAs) were collected from the university's Institutional Research Office for the semester prior to counseling (baseline) and the end of the semester in which they received services. Prior-semester GPA—specifically whether students were below the cutoff for academic probation (< 2.0 GPA)—was also used as an indicator of being academically at-risk. Finally, we collected clients' high school GPA, which is commonly used to control for individual differences in academic performance in studies examining GPA (e.g., Lee et al. 2009).

Procedure

Questionnaire data were collected at the counseling center in two stages. All clients completed intake questionnaires (demographics, presenting concerns, CCAPS-62, research informed consent) prior to their first appointment and follow-up measures (CCAPS-34, LOS) at their fifth appointment (defined as intake plus four individual and/or group appointments). This allowed the counseling center to use the questionnaires for clinical purposes in addition to keeping the staff blind to which clients were participating in the study. The fifth appointment was chosen for outcome data collection to maximize both the potential for measurable change and the number of participants (Gallagher, 2011). Questionnaire data and the total number of individual counseling sessions attended during the semester were shared with the researchers for consenting clients only[1]. To protect confidentiality, counseling center and institutional data were linked by student identification number so that no client names were used. The study was approved by the university's Institutional Review Board.

[1]The counseling staff counted both individual and group counseling sessions toward the total number of appointments for the purpose of collecting outcome data after the fifth appointment. However, these two types of appointments are tracked with different systems, and the counseling center only released data on the number of individual counseling sessions attended by each client. Therefore, all analyses including number of appointments utilize the number of individual counseling sessions.

Results

We examined the impact of measure selection in student affairs assessment research and practice by analyzing differences among three measures of academic outcomes (pre-post self-reports, retrospective reports, and grades) in a single sample of students using mental health services. We compared (1) the representativeness of each subsample (an indicator of the validity of the measure for capturing overall client outcomes) and (2) the conclusions drawn from each measure about academic outcomes of all participants generally, and of participants who were academically at-risk in particular. Measure statistics and correlations are presented in Table 2, and a summary of results by measure is presented in Table 3.

Table 2. *Measure Statistics and Intercorrelations of Baseline Academic Functioning and Academic Outcomes.*

1. Intake CCAPS Academic Distress Scale[a]	-						
2. 5th Appointment CCAPS Academic Distress Scale[a]	.68***	-					
3. Presenting Concerns: School or Grades[b]	.50***	.18*	-				
4. LOS Academic Outcomes Scale[c]	-.01	-.16	.10	-			
5. High School GPA[d]	-.17**	-.29**	-.15*	.05	-		
6. Prior-Semester GPA[d]	-.34***	-.10	-.29***	.10	.29***	-	
7. Current-Semester GPA[d]	-.35***	-.53***	-.26***	.16	.44***	.63***	-
Cronbach's α	.82	.83		.83			
Mean	1.74	1.69	0.46	3.50	3.24	2.79	2.69
Standard Deviation	1.08	1.00	0.50	0.73	0.43	0.82	0.94
N	365	122	368	117	292	240	350

[a]Pre-post self-reports from Counseling Center Assessment of Psychological Symptoms (CCAPS) [b]"School or Grades" selected from presenting concerns checklist intake: 1 = selected, 0 = not selected [c]Retrospective self-reports from Learning Outcomes and Satisfaction Survey (LOS) at the 5th appointment [d]Grade point averages (GPA) from institutional records $*p<.05$, $**p<.01$, $***p<.001$

Table 3. *Comparison of Subsample Representativeness and Counseling Services Impact Across Measures of Academic Outcomes.*

	Pre-Post Self-Reports: CCAPS Academic Distress (AD)[a]	**Retrospective Self-Reports: LOS Academic Outcomes (AO)[b]**	**Grades: Change in Semester GPA**	**Grades: Relationship Between Number of Appointments and GPA**
Description	change in distress scores from the 1st (intake) to 5th appointment (average of 4 items on a 4-point Likert-type scale)	students' retrospective perceptions that services helped them academically (average of 4 items on a 5-point Likert-type scale)	change in GPA from the semester prior to services (spring) to the semester in which services were received (fall; on a 4.0 scale)	predicting semester GPA from number of appointments attended, controlling for prior academic performance (i.e., high school GPA; both on a 4.0 scale)
Subsample	students who attended five or more appointments; *n* = 121	students who attended five or more appointments; *n* = 117	sophomore through senior students with continuous enrollment; *n* = 226	students with available high school GPA data; *n* = 283
Criteria for identifying academically at-risk clients	elevated AD scores at intake (> 2.75); *n* = 22 (18%)	listed "school or grades" among their reasons for seeking counseling; *n* = 46 (39%)	academic probation (GPA < 2.0); *n* = 35 (15%)	academic probation (GPA < 2.0); *n* = 26 (14% of those enrolled in college the prior semester)
Under-represented in the subsample	first-year students	first-year students	Students without continuous enrollment (e.g., first-year students, transfer students)	Older students (e.g., non-traditional aged students, veterans, international students, transfer students)
Academic outcomes	on average: no change ($M_{pre} = 1.67$; $M_{post} = 1.68$) among academically at-risk students: significant reductions in academic distress ($M_{pre} = 3.19$; $M_{post} = 2.66$)	on average: students perceived that services helped academically ($M = 3.50$) among academically at-risk students: somewhat stronger perceptions that services helped academically ($M = 3.59$)	on average: no change ($M_{spring} = 2.81$; $M_{fall} = 2.84$) among academically at-risk students: significant increases in GPA ($M_{spring} = 1.37$; $M_{fall} = 1.93$)	on average: no relationship between number of appointments and GPA (each session associated with an increase of 0.03 grade points) among academically at-risk students: positive but non-significant relationship (each session associated with an increase of 0.14 grade points)

Note. CCAPS = Counseling Center Assessment of Psychological Symptoms; LOS = Learning Outcomes & Satisfaction Survey; GPA = grade point average. [a]Locke et al. (2011, 2012); [b]Winterrowd et al. (2016)

Representativeness of the Subsamples

We first investigated whether different measures captured academic outcomes for representative subsamples of clients. We began by comparing the subsample of clients that attended five or more appointments and completed outcome measures (i.e., those eligible for analyses of pre-post self-reports and retrospective self-reports: 123 clients, 33% of the total sample[2]) to the full client sample in terms of gender, race/ethnicity, age, and year in school. First-year students were underrepresented in this subsample (13% vs. 25% of the full sample), $\chi^2(3, N = 122) = 8.55$, $p = .04$. There were no differences by gender, race/ethnicity, or age, $p > .30$.

Together our results demonstrate some of the potential consequences of measure selection, highlighting the importance of these choices for best practice in service outcomes assessment.

We then considered the subsample of clients with available data for two common analyses of semester GPA: change in GPA from the semester prior to counseling (spring) to the semester in which counseling services were received (fall), and the relationship between number of appointments and semester GPA, controlling for high school GPA. Change in GPA was only available if the client was continuously enrolled in credit-bearing courses from spring to fall (226 clients, 61% of the total sample). This excluded 128 clients who were not enrolled in spring (most often because they were first-semester students in fall; n = 86, leaving only four first-year/second-semester students), 14 clients who were not enrolled in fall (often because they withdrew from all their courses; $n = 10$), four clients who were not enrolled in either semester, and 46 clients who were missing semester GPA data for other reasons (e.g., taking only noncredit-bearing courses, being a transfer student, or taking the semester off). Despite low representation among first-year students this subsample did not differ from the total sample on gender, race/ethnicity, or age, $p > .10$

The subsample with available data for examining the relationship between number of appointments and fall GPA, controlling for high school GPA ($n = 283$; 77% of the total sample), excluded 14 clients without fall-semester GPA data and 76 clients for whom the university did not collect high school GPA data (e.g., nontraditionally aged students, veterans, international students). Accordingly, this subsample was younger than the total sample, $t(278) = -9.79$, $p < .001$. There were no significant differences in gender, race/ethnicity, or year in school, $p > .15$.

In sum, subsamples varied considerably across academic measures, both in size (33-77% of the total sample) and representativeness in terms of age and year in school. First-year students were underrepresented in analyses involving measures collected at the fifth appointment (i.e., pre-post and retrospective self-reports) and systematically excluded from analyses involving change in GPA. Analyses of the relationship between the number of appointments and GPA, with high school GPA as a covariate, underrepresented older students. None of the measures appeared to exclude students of a particular gender or race/ethnicity.

Conclusions Regarding Service Outcomes

Next we investigated whether different measures of academic outcomes would point to the same conclusions about service impacts, both for all clients generally and for clients identified as academically at-risk in particular. For these analyses, we identified clients as academically at-risk using baseline measures that paralleled each outcome measure. For change in CCAPS Academic Distress we used the CCAPS cutoff score for elevated Academic Distress at baseline (> 2.75). For retrospective reports of whether counseling helped academically we used clients' baseline presenting concerns (i.e., whether they listed school and grades among their reasons for seeking counseling). For both analyses of semester GPA (i.e., change in GPA, relationship between number of appointments and GPA), we used clients' academic probation status (prior semester GPA < 2.0).

Pre-post assessments. We first examined academic impact using changes in clients' CCAPS Academic Distress (AD) scores from intake (baseline) to the fifth appointment ($n = 121$[3]). Intake (baseline) scores from the CCAPS Academic Distress (AD) scale revealed

[2]Thirteen clients attended five or more appointments but did not complete the outcome measures.

[3]Two clients who attended five or more appointments and completed the retrospective self-report measure did not complete the pre-post Academic Distress measure.

low to moderate levels of academic distress overall (M = 1.67, SD = 1.00); however 18% of these clients (n = 22) fell in the elevated range and were therefore identified as academically at-risk. Average AD scores at the fifth appointment (M = 1.68, SD = 0.99) did not differ from intake, $t(120) = -.09$, $p = .93$, despite improvements in CCAPS scores for overall (nonacademic) well-being, $t(120) = 2.19$, $p = .03$. However, the subset of academically at-risk clients (i.e., those who had elevated academic distress at intake and attended five or more appointments, did show a significant reduction in Academic Distress from intake (M = 3.19, SD = 0.41) to the fifth appointment (M = 2.66, SD = 0.91), $t(21) = 3.23$, $p = .004$. Importantly, this indicates an improvement from an average score above the 2.75 cutoff for elevated Academic Distress to an average below the clinical cutoff.

One finding was consistent across measures: academic outcomes weremost positive for clients identified as academically at-risk at baseline.

Retrospective self-reports. Learning Outcomes and Satisfaction Survey Academic Outcomes (AO) scores were collected at the fifth counseling session (n = 117[4]). Just under half of the clients who completed the AO reported that counseling helped increase their academic motivation and/or attendance (43%), academic focus (50%), and academic performance (49%), and helped them stay in school (48%). The resulting AO scale mean of 3.50 (SD = 0.73) was significantly higher than the scale's neutral midpoint, $t(116) = 7.31$, $p < .001$. In addition, the clients who listed school or grades among their reasons for seeking counseling (an indicator of being academically at-risk) were especially likely to report that counseling had a positive impact on their academics five sessions later (n = 46, M = 3.59, SD = 0.72).

Semester grades. We then examined academic impact using changes in clients' semester GPAs (n = 226). Clients' average prior (spring) GPA was 2.81 (SD = 0.83), with 35 clients (15% of this subsample) below the cutoff for academic probation (< 2.0 GPA). On average, current semester (fall) GPAs (M = 2.84, SD = 0.87) were not significantly higher than the prior semester, $t(225) = 0.46$, $p = .65$. The subset of clients who were on academic probation did make significant improvements in GPA, however (M_{spring} =1.37, SD = 0.57; M_{fall} = 1.93, SD = 0.91), $t(34) = 3.43$, $p = .002$, with 16 clients moving off of academic probation.

We also examined fall GPA as a function of the number of individual counseling appointments attended, controlling for high school GPA (n = 283). The relationship between number of individual counseling appointments and fall GPA was positive but small, b = 0.03, SE = 0.02, $t(280) = 1.25$, $p = .21$. Among clients on academic probation (n = 26) each additional individual counseling session was associated with an increase of 0.14 grade points in GPA, an effect that did not reach statistical significance, b = 0.14, SE = 0.08, $t(23) = 1.62$, $p = .12$, but may be clinically significant for these students.

These three measures of academic functioning worked differently, even within a single subsample of clients, highlighting the importance of measure selection not just for documenting academic outcomes but also for identifying clients who are most in need of academic support.

Ancillary Analyses: Impact of Baseline Measure Selection

We assessed academic outcomes of counseling services using pre-post self-reports, retrospective self-reports, and grades in one CCC sample and replicated the pattern of mixed results found in prior research using disparate samples. Measure selection appeared to influence both the conclusions that could be drawn about counseling impact and the validity of those conclusions (due to the non-representativeness of the samples). However, one finding was consistent across measures: academic outcomes were most positive for clients identified as academically at-risk at baseline. This underscores the importance of selection of baseline measures in addition to outcome measures. Therefore, we conducted ancillary analyses with clients who had data from all three baseline measures (intake CCAPS Academic Distress, presenting concerns, prior-semester GPA) to examine the impact of measure selection on identification of academically at-risk clients.

There were 240 clients (65% of the total sample) with available data on all three baseline measures. Of these, 110 unique clients were identified as academically at-risk by at least one measure; 62 clients were in the dysfunctional range for CCAPS Academic Distress scores, 76 clients listed school and grades among their presenting concerns, and 38 clients were on academic probation (prior-semester GPA < 2.0). However, only 15 clients (13.6% of academically at-risk clients) were identified as struggling by all three measures, and only 49

[4]Six clients who attended five or more appointments and completed the pre-post Academic Distress measure did not complete the LOS-AO

clients (44.5%) were identified as struggling by two measures. In other words, many clients with low GPAs were not distressed about academics, and many clients with higher levels of academic distress or low GPAs did not report school and grades as a primary reason for seeking counseling. These three measures of academic functioning worked differently, even within a single subsample of clients, highlighting the importance of measure selection not just for documenting academic outcomes but also for identifying clients who are most in need of academic support.

Discussion

We compared three types of academic outcome measures (i.e., pre-post self-reports, retrospective self-reports, and grades) within a single counseling center client sample and found that measure selection impacted both the representativeness of the subsample and the conclusions that might be drawn about the effectiveness of services. No one measure captured a fully representative sample on its own: subsamples differed in size (33–77% of the total sample) and in representativeness in terms of age and year in school. Retrospective self-report measures demonstrated positive academic impacts for all clients, on average, and particularly for academically at-risk students. In contrast, pre-post self-report and institutional (GPA) measures showed positive impacts for academically at-risk students only. Interestingly, ancillary analyses revealed that diverse baseline measures resulted in unique groups of students being identified as academically at-risk in the first place. Together our results demonstrate some of the potential consequences of measure selection, highlighting the importance of these choices for best practice in service outcomes assessment.

This demonstrates the dramatic impact of timing of assessment: collecting outcomes at the fifth appointment excluded two-thirds of the students receiving services and also underrepresented first-year student clients.

Representativeness of Measure Subsamples

Inherent in choosing an assessment method is selecting the subsample of students with available data. Consistent with previous literature (e.g., DeStefano et al., 2001), pre-post and retrospective self-reports in this study captured academic outcomes for students who attended a minimum number of appointments (e.g., five) but excluded those who attended fewer (a majority in this study). This demonstrates the dramatic impact of timing of assessment: collecting outcomes at the fifth appointment excluded two-thirds of the students receiving services and also underrepresented first-year student clients. A shorter time frame minimizes attrition (and potentially increases first-year student representation) but longer time frames may maximize opportunities for academic impact.

For grades, using change in semester GPA from before counseling to after (similar to Illovsky, 1997; Wlazelek & Coulter, 1999) may be ideal for students with continuous enrollment but it underrepresents students in their first semester (i.e., first-year and transfer students). These exclusions are particularly problematic because first-year and transfer students may be more likely to struggle academically than their peers (Berger & Malaney, 2003; Lee et al., 2009). Assessing the relationship between number of individual counseling appointments and semester grades, controlling for high school GPA (similar to Lee et al., 2009), provided the largest subsample of students. However, this assessment still underrepresented nontraditional students without high school GPA information (e.g., older students, veterans)—a group with noted differences in academic needs (e.g., Spitzer, 2000). Together, these results call into question the extent to which any single measure can be used to capture academic outcomes representative of the whole client population.

Together, these results call into question the extent to which any single measure can be used to capture academic outcomes representative of the whole client population.

Academic Outcomes of Counseling Services

The existing research on academic outcomes of counseling services is mixed (e.g., Lee et al., 2009; Lockard et al., 2012; Turner & Berry, 2000) and this study provides some insights into why that might be the case. We implemented three common measures of academic outcomes—pre-post self reports, retrospective reports, and grades—and replicated the mixed results of prior research within a single sample, suggesting that the apparent inconsistency in the literature may be due, at least in part, to differences in academic outcome measures. The Lambert and Hawkins (2004) model illuminates some of the important differences among these measures, including variation in content (changes in academic functioning vs. perceptions of being helped vs. course performance), source (client vs. university records), method (self-report vs. institutional data), and time orientation (self-reports vary day to day, whereas GPA varies semester to semester; Table 1). This is consistent with other research in higher education that highlights variation in outcomes assessment with different sources (e.g., Sexton, 1996) and methods (e.g., Bowman, 2013). Our results suggest that variability in sample characteristics may also contribute to the mixed findings of prior research—a result not explicitly addressed in the Lambert and Hawkins model (2004).

Implications for Best Practice in Service Outcomes Assessment

Our results demonstrate empirically what many practitioners might have guessed intuitively—that differences among pre-post self-reports, retrospective reports, and grades are more profound than simple variations in operationalization; they capture discrete academic outcomes.

This study has important implications for research and practice. Our results suggest that measure selection plays a fundamental role in demonstrating service effectiveness. Specifically, the "best" measure for capturing academic impact appears to depend on which subsample of clients and which aspect of academic functioning researchers or practitioners most want to assess. Practitioners should therefore (1) determine their specific service goals and which aspects of academic functioning they intend to support, (2) identify academic outcome measures consistent with those goals, (3) choose appropriate baseline measures, and (4) determine the best data collection time frame to capture the intended outcomes for the majority of clients and/or targeted client groups.

Our results demonstrate empirically what many practitioners might have guessed intuitively—that differences among pre-post self-reports, retrospective reports, and grades are more profound than simple variations in operationalization; they capture discrete academic outcomes. For example, as depicted in Table 2, responses to retrospective self-reports of service perceptions were unrelated to pre-post symptom questionnaires and GPA. Using multiple measures in combination could therefore help researchers and student affairs practitioners alike to better understand the students' academic experiences and to document the impact of student services on many different aspects of academic functioning (Astin & antonio, 2012; Schuh, 2011; Suskie, 2009). In addition, including both self-report and institutional or observational outcomes increases confidence in conclusions about service outcomes (Sexton, 1996) and protects against the risk of using self-reports solely as a proxy for student learning or growth (e.g., Bowman, 2013). We encourage researchers and practitioners to consider a variety of academic outcomes that might be consistent with their service goals, including those examined here as well as others (e.g., academic self-efficacy, engagement, satisfaction; see Kuh, Kinzie, Buckley, Bridges, & Hayek, 2006). Many of these variables have been considered predictors of academic achievement (i.e., grades) but can be important outcomes in and of themselves.

Assessment of academic outcomes continues to be of paramount importance in student affairs—as best practice and as a means of demonstrating each unit's value in supporting the academic mission of the university.

In this study we analyzed service outcome data statistically. We hope our analyses give practitioners some ideas of ways they can look at their own outcome data. However, we recognize that many counseling centers (and other student affairs services) have small client populations and/or limited staff and resources for statistical analyses. Certainly practitioners could examine their data descriptively. In fact, some assessments (such as the CCAPS assessment we used in this study) include in their user manuals guidance on how to detect and interpret change over time, without statistical analyses. Even if a campus or center is too small to collect meaningful data in any given semester or year, intentional and systematic measure selection will allow for examination of trends in service utilization and outcomes across multiple years or in collaboration with multiple centers (e.g., Winterrowd et al., 2016).

As the Lambert and Hawkins (2004) model emphasizes, it is important that assessments are applicable and practical. Ultimately, outcomes assessments will only lead to service improvement if they are useful to practitioners. Therefore, the "best" measures and methods can and should vary unit to unit and campus to campus. We hope that our study highlights some of the considerations practitioners might take into account when selecting outcome measures and that our suggestions will help student affairs units maximize their opportunities to demonstrate their value and further improve their services.

Limitations

The current study provides a direct comparison of several commonly used measures of academic outcomes (pre-post self-reports, retrospective reports, and grades, including both change in GPA and the relationship between number of appointments and GPA). However, this study is by no means a comprehensive comparison of all assessment designs and measures. We examined self-reported academic outcomes at the fifth counseling appointment whereas previous research on mental health services has typically examined pre-post self-report outcomes after six appointments (e.g., DeStefano et al., 2001; Lockard et al., 2012; Nafziger et al., 1999) and retrospective reports at the end of the semester (e.g., Winterrowd, et al., 2016; Reynolds & Chris, 2008; Turner & Berry, 2000). Furthermore, we counted both individual and group counseling sessions toward our shorter timeframe for assessment (five appointments) whereas many studies count only individual counseling sessions. For example, Lockard and colleagues (2012) found positive effects of counseling services on the CCAPS Academic Distress scale after six individual counseling sessions. Thus our timeframe for assessment may have limited our ability to detect the academic benefits of counseling.

We hope that our study highlights some of the considerations practitioners might take into account when selecting outcome measures and that our suggestions will help student affairs units maximize their opportunities to demonstrate their value and further improve their services.

In terms of institutional measures, we examined changes in semester GPA from the semester prior to counseling to the semester in which counseling services were received as well as the relationship between number of appointments and semester GPA. Other studies have considered cumulative GPA (Lee et al., 2009) or semester GPA from semesters after counseling was received (Illovsky, 1997). In addition, the current study did not examine retention, a variable of interest among many in student affairs. Although retention is argued to be an inappropriate outcome for counseling services (e.g., Heitzmann & Nafziger, 2001; Lockard et al., 2012), it may be more appropriate for other student services and its relationship to diverse measures of academic outcomes should be explored in future research.

Conclusion

Assessment of academic outcomes continues to be of paramount importance in student affairs—as best practice and as a means of demonstrating each unit's value in supporting the academic mission of the university. However, it can be difficult to determine which academic outcome measures to use to best capture the impact of student services. By considering the characteristics of a given academic outcomes measure, including the subsample of clients who will have available data, student affairs practitioners can select the appropriate measures for the particular population they are trying to serve and evaluate the specific aspects of academic functioning their services are designed to promote. In the end, it may be best to utilize multiple measures in combination in order to fully examine academic outcomes across students.

References

American College Personnel Association. (1994). *The student learning imperative: Implications for student affairs.* Washington, D.C.: American College Personnel Association.

Astin, A. W., & antonio, a. l. (2012). *Assessment for excellence: The philosophy and practice of assessment and evaluation in higher education* (2nd ed.). New York, NY: Rowman & Littlefield/American Council on Education.

Berger, J. B., & Malaney, G. D. (2003). Assessing the transition of transfer students from community colleges to a university. *NASPA Journal, 40*(4), 1–23.

Bowman, N. A. (2013). Understanding and addressing the challenges of assessing college student growth in student affairs. *Research & Practice in Assessment, 8*, 5–14.

Center for Collegiate Mental Health (2012). *CCAPS 2012 User Manual.* University Park, PA.

Cholewa, B., & Ramaswami, S. (2015). The effects of counseling on the retention and academic performance of underprepared freshmen. *Journal of College Student Retention, 17*, 204–225. doi: 10.1177/1521025115578233

DeStefano, T. J., Mellot, R. N., & Petersen, J. D. (2001). A preliminary assessment of the impact of counseling on student adjustment to college. *Journal of College Counseling, 4*, 113–121. doi: 10.1002/j.2161-1882.2001.tb00191.x

Erwin, T. D., & Sivo, S.A. (2001). Assessing student learning and development in student affairs: A nuts and bolts introduction. In R. B. Winson, Jr., D. G. Creamer, & T. K. Miller (Eds.), *The professional student affairs administrator: Educator, leader, and manager*. (pp. 357–375). New York: Brunner-Routledge.

Gallagher, R. P. (2011). *National Survey of Counseling Center Directors 2011* (Monograph No. 8T). The International Association of Counseling Services, Inc.

Heitzmann, D., & Nafziger, K. L. (2001). Assessing counseling services. In J. H. Schuh & M. L. Upcraft and Associates (Eds.), *Assessment practice in student affairs: An applications manual*. (pp. 390–412). San Francisco: Jossey-Bass.

Illovsky, M. E. (1997). Effects of counseling on grades and retention. *Journal of College Student Psychotherapy, 12*, 29–44. doi:10.1300/J035v12n01_04

Kuh, G. D., Kinzie, J., Buckley, J. A., Bridges, B. K., & Hayek, J. C. (2006). What matters to student success: A review of the literature. Commissioned Report for the National Symposium on Postsecondary Student Success: Spearheading a Dialog on Student Success. Washington, DC: National Postsecondary Education Cooperative. Retrieved from: https://www.ue.ucsc.edu/sites/default/files/WhatMattersStudentSuccess(Kuh, July2006).pdf

Lambert, M. J., & Hawkins, E. J. (2004). Measuring outcome in professional practice: Considerations in selecting and using brief outcome instruments. *Professional Psychology: Research and Practice, 35*, 492–499. doi: 10.1037/0735-7028.35.5.492

Lee, D., Olson, E. A., Locke, B., Michelson, S. T., & Odes, E. (2009). The effects of college counseling services on academic performance and retention. *Journal of College Student Development, 50*, 305-319. doi: 10.1353/csd.0.0071

Lockard, A. J., Hayes, J. A., McAleavey, A. A., & Locke, B. D. (2012). Change in academic distress: Examining differences between a clinical and nonclinical sample of college students. *Journal of College Counseling, 15*, 233–246. doi: 10.1002/j.2161-1882.2012.00018.x

Locke, B. D., Buzolitz, J. S., Lei, P.-W., Boswell, J. F., McAleavey, A. A., Sevig, T. D., & Hayes, J. A. (2011). Development of the counseling center assessment of psychological symptoms-62 (CCAPS-62). *Journal of Counseling Psychology, 58*, 97–109. doi: 10.1037/a0021282

Locke, B. D., McAleavey, A. A., Zhao, Y., Lei, P.-W., Hayes, J. A., Castonguay, L. G., Li, H., Tate, R., & Lin, Y.-C. (2012). Development and initial validation of the counseling center assessment of psychological symptoms-34. *Measurement and Evaluation in Counseling and Development, 45*, 151–169. doi:10.1177/07481756114326242

Miller, A. & Markman, K. D. (2007). Depression, regulatory focus, and motivation. *Personality and Individual Differences, 43*, 427–436. doi: 10.1016/j.paid.2006.12.006

Minami, T., Davies, D. R., Tierney, S. C., Bettmann, J. E., McAward, S. M., Averill, L. A., Huebner, L. A., Weitzman, L. M., Benbrook, A. R., Serlin, R. C., & Wampold, B. E. (2009). Preliminary evidence on the effectiveness of psychological treatments delivered at a university counseling center. *Journal of Counseling Psychology, 56*, 309–320. doi: 10.1037/a0015398

Nafziger, M. A., Couillard, G. C., & Smith, T. B. (1999). Evaluating therapy outcome at a university counseling center with the College Adjustment Scales. *Journal of College Counseling, 2*, 3–13.

Reynolds, A. L., & Chris, S. (2008). Improving practice through outcomes based planning and assessment: A counseling center case study. *Journal of College Student Development, 49*, 374–387. doi: 0.1353/csd.0.0017

Schuh, J. H. (2011). *Assessment methods for student affairs*. San Francisco, CA: Jossey-Bass, Inc.

Sexton, T. L. (1996). The relevance of counseling outcome research: Current trends and practical implications. *Journal of Counseling & Development, 74*, 590–600. doi: 10.1002/j.1556-6676.1996.tb02298.x

Spitzer, T. M. (2000). Predictors of college success: A comparison of traditional and non-traditional age students. *NASPA Journal, 38*, 82–98. doi: 10.2202/1949-6605.1130

Stallman, H. M. (2010). Psychological distress in university students: A comparison with general population data. *Australian Psychologist, 45*, 249–257. doi: 10.1080/00050067.2010.482109

Suskie, L. (2009). *Assessing student learning: A common sense guide* (2nd ed.). San Francisco, CA: Jossey-Bass.

Turner, A. L., & Berry, T. R. (2000). Counseling center contributions to student retention and graduation: A longitudinal assessment. *Journal of College Student Development, 41*, 627–636.

Upcraft, M. L., & Schuh, J. H. (1996). *Assessment in student affairs: A guide for practitioners*. San Francisco: Jossey-Bass, Inc.

Varlotta, L. E. (2012). Toward a more data-drive supervision of collegiate counseling centers. *Journal of American College Health, 60*, 336–339. doi: 10.1080/07448481.2012.663843

Winterrowd, E., Priniski, S. J., Achter, J., & Abhold, J. (2016). Correlates of satisfaction, intrapersonal learning, and academic outcomes at counseling centers in a university system. *College Student Journal, 50*, 288–301.

Wlazelek, B. G., & Coulter, L. P. (1999). The role of counseling services for students in academic jeopardy: A preliminary study. *Journal of College Counseling, 2*, 33–41.

Book Review

Learning Assessment Techniques:
A Handbook College Faculty.
Elizabeth F. Barkley & Claire H. Major.
San Francisco: Jossey Bass, 2016. 480 pp.
ISBN 1119050898. Paperback. $40.00.

REVIEWED BY:
Monica Stitt-Bergh, Ph.D.
***University* of Hawaii at Manoa**

Elizabeth F. Barkley and Claire Howell Major's book, *Learning Assessment Techniques: A Handbook for College Faculty* (Wiley & Sons, 2016), strives to take a fresh look at course-level learning assessment techniques. The admirable aim of the book is to integrate teaching, learning, and assessment to serve multiple purposes: improve student learning, enhance pedagogy, use faculty time efficiently, and fulfill (external) demands for learning evidence. Certainly Barkley and Major tackle an important topic that will interest educators, assessment practitioners, and support personnel. The worthy goals of *Learning Assessment Techniques*, explained in the preface and introduction, create lofty expectations for readers of this latest handbook for college faculty and staff. They situate their book in (a) the scholarship of teaching and learning and (b) classroom assessment techniques by Patricia Cross and Thomas Angelo. Many of us have the *Classroom Assessment Techniques: A Handbook for College Teachers* (Angelo & Cross, 1993) on our bookshelves and it is not gathering dust in my office. That is quite an accomplishment for an academic book to maintain relevance and usefulness for decades. Will Barkley and Major's book experience the same fate? I'm not sure.

The authors want to help faculty, assessment practitioners, and instructional designers effectively and efficiently "draw teaching and assessment together to create a seamless and unified process" (p. xiv) and, just as important in today's competitive higher education context, help them "document, interpret, and report student learning to a variety of stakeholders" (p. xv). Thus the authors address a need that did not exist when Angelo and Cross published their handbook. Although individual elements of Barkley and Major's book are valuable, the book as a whole could be more carefully presented to maximize use for readers.

The authors' qualifications and experiences give them credibility on the topic of teaching and learning, which is evident in their accessible, easy to understand introductory chapter. Barkley is a pianist and music educator who has also worked with faculty at many higher education institutions. Major specializes in instructional design and technology and qualitative research. She has taught at several types of institutions. The two have co-authored, along with Cross, another book for college faculty, *Collaborative Learning Techniques: A Resource for College Faculty* (Barkley, Cross, & Major, 2005). Readers should be aware that some of the techniques in this book appear in the previous books or have been modified from the previous books in the Handbook for College Faculty series.

Learning Assessment Techniques has two main parts. First, an overview of why they promote learning assessment techniques (LATs) and how to implement, report results, and improve student learning. Second, they describe 50 specific LATs divided into six learning domains: knowledge, application, integration, human dimension, caring, and learning how to learn. In the overview, Barkley and Major describe why the LATs promote learning. First, "in order to effectively guide students in their own acquisition of knowledge, a college teacher also needs knowledge of pedagogy" (p. 2). Second, they explain that the LATs employ elements of effective pedagogy: "1. Identifying and communicating clear learning goals 2. Helping students achieve these goals through activities that promote active, engaged learning 3. Analyzing, reporting, and reflecting upon results in ways that lead to continued improvement" (p. 3). Third, Barkley and Major illustrate how LATs intertwine learning goals, learning activity, and outcomes assessment in a unified whole and how "it is impossible to tell where one begins and the other ends" (p. 4). In other words, by using LATs, the faculty member is teaching, engaging, and assessing students all at the same time. This is an important point because it places the assessment-for-improvement concept as foundational to an effective educational experience. I applaud the authors for their stance.

In other words, by using LATs, the faculty member is teaching, engaging, and assessing students all at the same time. This is an important point because it places the assessment-for-improvement concept as foundational to an effective educational experience.

The authors draw from Suskie (2009) to differentiate assessment and grading and from Wiggins (1998) (embedded and authentic assessment) to clarify assessment for readers, which is appropriate and supports their overarching goals for the book. The parts on selecting and implementing LATs will likely be useful to readers because the authors give sufficient details, examples, and practical steps. The authors also describe basic ways to analyze results from the LATs—from descriptive statistics to cross-case comparisons—which support the authors' goal of helping faculty report results to multiple stakeholders. The last chapter in the overview ('Closing the Loop') addresses a particularly important question: how can faculty improve student learning after the results are in? They provide five recommendations: modify the goals/objectives/outcomes, assessment purpose, LAT, implementation, or analysis of findings. Given that the authors themselves state the importance of this chapter because their primary goal is student learning improvement, a more in-depth discussion was needed than this two-page chapter. "Closing the loop" has been notoriously difficult; this chapter could benefit from a richer analysis of what results might be best addressed by which of their five recommended changes or from a discussion of how to choose one of these five solutions so that improved student learning is likely.

The authors also describe basic ways to analyze results from the LATs—from descriptive statistics to cross-case comparisons—which support the authors' goal of helping faculty report results to multiple stakeholders.

The second and final section of the book has 50 LATs. Each LAT includes examples from different academic disciplines, lists the amount of time involved, the steps to implement, a consideration for use in an online course, a description of how to report to external audiences, and variations. LATs range from quick (e.g., entry tickets, sequence chains) to involved (e.g., think-aloud protocols, editorial reviews, e-portfolios). The authors' inclusion of rubrics, tables, and charts that illustrate how to report aggregate results is good, although, in some cases I found myself disagreeing with the table/chart format or rubric content. For example, the detailed oral presentation rubric (p. 326) seems mismatched to the LAT's three-minute, one-slide presentation. I encourage using the tables, charts, and rubrics as starting points for faculty and professional staff to modify, not as the ideal models.

The 50 LATs provide evidence of learning because students produce written documentation or an observable behavior (such as a debate). Most of the LATs are very good in providing formative information and developing student knowledge, skill, or values but not all of the techniques are designed for summative evaluation. More important, faculty/staff may need an additional evaluation tool to provide information on whether learning in the specified domain actually occurred. My primary criticism of this work is highlighted in a brief description of the book's organizing framework using Fink's (2013) taxonomy and examples from the chapter on the caring domain.

The authors use Fink's (2013) significant learning taxonomy to organize the 50 LATs, but it does not provide readers with practical insight. Fink's work is a fresh departure from the psychometrically-influenced taxonomies of educational objectives for cognitive and affective domains, popularly called "Bloom's taxonomy" (Bloom, Engelhart, Furst, Hill, & Krathwohl, 1956; Krathwohl, Bloom, & Masia, 1964). These two taxonomies have categories that are hierarchical, developmental, and non-overlapping. However, unlike Bloom et al., Fink's taxonomy is intentionally non-hierarchical; the six domains of learning overlap and are interactive and synergistic. Half of Fink's categories specify what content makes the learning significant. Fink's human dimension involves learning about the self and others; the caring domain involves caring primarily about learning; and the learning how to learn domain involves, well, the subject of learning. These and the other three domains—foundational knowledge, application, and integration—are intertwined and meaningful and that is their strength in Fink's taxonomy. The synergistic nature of the taxonomy makes assigning each LAT to a single domain difficult and probably impossible but Barkley and Major insist on doing so. I think the book would have been more successful if they did not use Fink's taxonomy.

The authors try to address unfamiliarity or potential confusion with Fink's taxonomy and terminology by starting each LAT chapter with a definition and description of the domain. Fink's terms such as human dimension, integration, and caring have particular meaning so readers might benefit from reviewing the opening pages of each domain chapter. For example, Fink (2013) describes the human dimension as "important relationships and interactions we all have with ourselves and with others" (p. 50) and the caring dimension as caring more deeply about something— that is, to "value something differently" (p. 55).

Barkley and Major give us "action verbs" (p. 19) for Fink's six categories. Their suggested use of this verb list is inappropriate because it does not correspond to Fink's taxonomy nor to the spirit of the taxonomy. For example, they list adapt, evaluate, and propose as verbs in the human dimension category. But if faculty create learning objectives such as the student adapts mathematical models, evaluates geographic regions, or proposes a feeding schedule for fish, that learning does not fall into the human development category because it does not directly honor and advance relationships with the self or others as Fink's taxonomy specifies. Appropriate objectives using these verbs and the human dimension category could be that the student adapts one's self, evaluates interactions with others, or proposes ways to develop better relationships among people. Fink's domains of significant learning do not hinge upon verbs or generic behaviors, as is more the case in the cognitive taxonomy by Bloom et al. (1956). Fink's domains intentionally involve what is being learned and thus a verb list as proposed by Barkley and Major is not a useful match.

I had particular problems with the "Teaching and Assessing for the Caring Domain" chapter. The LATs themselves are useful and some have the potential to develop caring for the subject at hand but the LATs do not help to adequately evaluate students' levels of caring as a result of the educational experience. Readers who are in fields that explicitly value caring—e.g., medical education, nursing, social work, teacher education—will likely find these LATs not at all useful for figuring out if they have succeeded in developing caring students.

The disciplinary examples in the caring domain chapter include tasks for the student such as communication of original research results, editorial writing, and problem identification and solution development. I remain unconvinced that evaluating these tasks using the rubrics provided would allow faculty to infer that caring occurred. In my experience, in order to evaluate caring it should be an explicit part of the teaching, the task, and the rubric. On the oral communication rubric (p. 326) the enthusiasm dimension might be a proxy for caring but the advertisement (p. 332), editorial (p. 341), and debate (p. 348) rubrics do not evaluate students' degree of caring. We cannot automatically conclude that changes in caring occur when we compel a student via grades and credits to argue one side of an issue. For readers interested in evaluating students' caring, I recommend adding an explicit caring dimension to a rubric or using an additional evaluation tool (e.g., a self report) to connect the task to the caring domain. Despite finding the LATs in this section to be useful as classroom teaching, learning, and assessment tools, I do not see their direct connection to caring.

Faculty, instructional designers, assessment practitioners, and others who want to use this book to implement changes in pedagogy or learning measurement need to think carefully about the LATs and what learning claims can be made from their application and results. As I describe in the paragraph above, the LATs may not provide evidence related to their chapter title/learning domains. Readers may also benefit by considering which LATs give students sufficient time and guidance in order to produce their best work. If they do not, the learning artifact is likely best used for formative assessment, not summative. If the authors would have fully immersed the reader in Fink's overlapping domains and the implications for teaching, learning, and assessment, Barkley and Major's book would be more helpful to the academic and assessment community who are actively engaged in student learning improvement. Although I found the book to fall short in this area, the book's description of how to implement the LATs and the LATs themselves are useful.

Faculty, instructional designers, assessment practitioners, and others who want to use this book to implement changes in pedagogy or learning measurement need to think carefully about the LATs and what learning claims can be made from their application and results.

References

Angelo, T. A., & Cross, K. P. (1993). *Classroom assessment techniques: A handbook for college teachers*. San Francisco: Jossey-Bass.

Barkley, E. F., & Major, C. H. (2016). *Learning assessment techniques: A handbook for college faculty*. San Francisco: Jossey Bass.

Barkley, E. F., Cross, K. P., & Major, C. H. (2005). *Collaborative learning techniques: A resource for college faculty*. San Francisco: Jossey-Bass.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain*. New York: Longmans, Green and Co.

Fink, L. D. (2013). *Creating significant learning experiences: An integrated approach to designing college courses* (2nd Ed.). San Francisco: Jossey-Bass.

Krathwohl, D. R., Bloom, B. S., & Masia, B. B. (1964). *Taxonomy of educational objectives: The classification of education goals. Handbook II: Affective domain*. New York: David McKay.

Suskie, L. (2009). *Assessing student learning: A common sense guide* (2nd Ed.). San Francisco: Jossey Bass.

Wiggins, G. P. (1998). *Educative assessment: Designing assessments to inform and improve student performance*. San Francisco: Jossey Bass.

Notes in Brief

There are likely as many approaches to teaching assessment as there are people teaching assessment. Graduate courses on assessment can be structured with a singular focus, such as learning outcomes assessment, or along a competencies-based framework. Such frameworks include the Assessment Skills and Knowledge (ASK) Standards developed by College Student Educators International (ACPA) in 2006 and the ACPA/NASPA Professional Competencies (Bresciani & Todd) introduced in 2010 and revised in 2015, which include the Assessment, Evaluation, and Research competency area. The purpose of this article is to share reflections on an approach to teaching assessment through the use of a CAS Standards (Council for the Advancement of Standards in Higher Education, 2012) self-study in a master's level assessment course during the Fall 2015 semester.

AUTHORS

Brian Bourke, Ph.D.
Mruuay State University

Using a CAS Self-Study to Teach Assessment Practice

There are likely as many approaches to teaching assessment as there are people teaching assessment. Since first developing an assessment course in 2009 I have approached teaching assessment in a few different ways. I first used the Assessment Skills and Knowledge (ASK) Standards developed by College Student Educators International (ACPA) (Mitchell, 2006). I have tweaked, added, and subtracted instructional approaches, activities, and assignments in the course with the introduction of the ACPA/NASPA Professional Competencies (Bresciani & Todd) in 2010, which included the Assessment, Evaluation, and Research competency area. The purpose of this article is to share reflections on a new approach I took using the CAS Standards (Council for the Advancement of Standards in Higher Education, 2012) in a master's level assessment course during the Fall 2015 semester.

One of the required courses in the master's graduate preparation program at Murray State University is institutional research and assessment. The purpose of the course is to help students develop core competencies related to assessment work in postsecondary institutions. The learning outcomes from the course include understanding the development and evolution of the institutional research and assessment functions, identifying research needs within an institution, preparing assessment plans, and gathering, analyzing, and synthesizing data from multiple sources.

CORRESPONDENCE

Email
bbourke@murraystate.edu

While on the faculty at a previous institution, I designed the assessment course so that students would work with campus partners to carry out assessment projects—but that came with mixed results. At times, some campus partners had not previously engaged in assessment efforts and others had expectations beyond what students learning the process themselves could realistically provide. With those past experiences in mind, I set out to develop a different hands-on project for students to learn how to gather, analyze, and interpret data, and then synthesize and report findings.

The hands-on project that students completed during the Fall 2015 semester was an analysis of the master's graduate preparation program at Murray State University. The program analysis project was based on the CAS Standards for Master's Level Student Affairs Administration Preparation Programs (Council for the Advancement of Standards in Higher Education, 2012; referred to as Master's Program CAS Standards throughout the remainder of this article). Students enrolled in the course served as the review team conducting the analysis of the master's program. The Master's Program CAS Standards are divided into nine parts: program mission and objectives; recruitment and admission; curriculum policies; pedagogy; curriculum; equity and access; academic and student support; professional ethics and legal responsibilities; and program evaluation.

CAS Standards

With those past experiences in mind, I set out to develop a different hands-on project for students to learn how to gather, analyze, and interpret data, and then synthesize and report findings.

The Council for the Advancement of Standards in Higher Education (CAS) is a consortium of 42 professional associations with the aim of advancing the use of professional standards for the purpose of continuous quality improvement of programs and services (Council for the Advancement of Standards in Higher Education, 2015). Continuous quality improvement is addressed through the following goal: "to promote the assessment and improvement of higher education services and programs through self-study, evaluation, and the use of CAS standards" (Council for the Advancement of Standards in Higher Education, 2015, "CAS Purpose" p.2). The Master's Program CAS Standards reflect one of the 45 sets of standards.

The precursor to the current version of the CAS Standards for master's programs was first published in 1979 as "Standards for the Preparation of Counselors and College Student Affairs Specialists at the Master's Degree Level" (CAS, 2013, p. 2). The Master's Program CAS Standards, last revised in 2012 (CAS, 2012), consist of nine parts: mission and objectives, recruitment and admission, curriculum policies, pedagogy, curriculum, equity and access, academic and student support, professional ethics and legal responsibilities, and program evaluation. As with the other CAS Standards, the master's program standards are supplemented by a Self-Assessment Guide (SAG). The purpose of the SAG is to provide a systematic and standardized approach to identifying strengths and weakness through a self-study (CAS, 2013). Each SAG includes a section titled "Formulating an Action Plan," which consists of guiding questions provided to facilitate discussions aimed at enacting plans for improvement (CAS, 2013).

Steps in the self-study

While CAS does not prescribe a static procedure for conducting a self-study, there is a set of five recommended steps listed in the SAG (CAS 2013):

1. Form the review team. The recommendation is to form a team with diverse perspectives, including a chair and other members from outside of the unit under study.
2. Prepare and train the review team. To prepare, team members should familiarize themselves with the standards appropriate for the review and come to a consensus for interpreting information and generating ratings.
3. Compile and review documents and other sources of evidence. In addition to evaluating documents and other data, review team members might seek out other sources of data following sharing scale ratings with staff from the unit under study (CAS, 2016). For the master's program review additional sources included conducting interviews with various stakeholders, including faculty, current students, and administrators.
4. Review documents and other evidence of program performance. The fourth step consists of conducting the review, through which the review team will use the standards criteria statements and assign a rating to each one, using the scale provided to reflect degree of compliance (from Does Not Meet to Exceed). Generally, team members do this individually and then meet to

compare ratings, discuss and resolve discrepancies, and finalize their collective evaluation. (CAS, 2016, p. 7)

5. Write up review results and recommendations. The recommendations may be as specific as setting "a timetable for addressing deficiencies" (CAS, 2013).

Assessment Course Project

To reflect the course-level learning outcome of learning how to gather, analyze, and synthesize data, I provide an account through the rest of this article of the procedures that the review team (students enrolled in the course) utilized toward this aim. Note that the steps we took as a learning exercise varied slightly from those suggested for a self-study as described by in the Master's Program CAS Standards SAG (2013). Because the self-study was adapted as an instructional activity, significant emphasis was placed on the second and third steps in the self-study process.

One of the themes I stressed in the course was that good assessment work relies on multimodal data collection and analysis.

The students were provided the SAG developed for the Master's Program CAS Standards. We spent time reviewing and discussing the steps of what is termed as the self-assessment process during each weekly class session. The time spent each week served as the first two steps in the process—which are to establish and prepare the review team and to understand the CAS standards and guidelines of the self-study. We discussed and developed plans for completing the third step, which involves compiling and reviewing documentary evidence. As I share in the section on analyzing data, we used class meeting time to demystify the fourth step of judging performance, based on compiled evidence. Throughout the semester we discussed the fifth and final step of completing the CAS self-assessment process. This last step involves examining individual and group ratings assessed in the fourth step and synthesizing the review team's evaluation of the extent to which the master's program meets each CAS Standard.

Gathering data

One of the themes I stressed in the course was that good assessment work relies on multimodal data collection and analysis. I avoided calling this mixed methods research, as utilizing varied approaches to organizational effectiveness, student learning, or other common assessment aims does not reflect a cogent mixed methods research design (Creswell, 2015).

As a class, the students identified key stakeholders who could offer perspectives and provide data via interview. Each student assumed responsibility for interviewing a key stakeholder and then transcribing the interview. The transcripts were posted to a shared online file-sharing space for easy access. The transcripts were used to answer the questions associated with each SAG part. Students were also expected to identify data needs prior to conducting the interviews so that they could ask for further documentation from each stakeholder. All documents and other data gathered in this manner were also posted to the shared online file-sharing space.

Analyzing data

Data analysis presented a bit of a challenge. Although an introduction to research methods is a prerequisite for the assessment course students did not feel confident in their abilities to analyze data. As we engaged in the CAS review process, the students saw the process as being more qualitative and subjective. Their concerns and trepidations were with the prospect of manipulating SPSS or other statistical software, but the students became much more comfortable with the idea of data analysis for the CAS review as the course progressed. Students' concerns were further eased as I guided them through analyzing data for Part 1: mission and objectives. As noted in the previous section, data from various sources, including interview transcripts, were used to answer the questions from each part of the SAG.

As we engaged in the CAS review process the students saw the process as being more qualitative and subjective.

The primary component of the CAS self-study is the use of the rating scales for each of the nine parts. In order to complete the ratings for the items listed for each part in the SAG students gathered and evaluated evidence prior to determining rankings on individual items.

For example, when addressing Part 5: the curriculum, students examined program documents that described the curriculum, reviewed course syllabi, and read transcripts from interviews with stakeholders.

Synthesizing findings

To synthesize the findings, the students created a format based on the nine parts of the SAG that was used for the semester-long project. In class, we discussed pulling together data from multiple sources related to the same points of inquiry: the numeric rating items and the summary questions associated with each part within the SAG. The students experienced the challenge of synthesizing the basis for each numeric rating along with addressing information from transcripts of interviews with stakeholders.

The student-written report has led program faculty to develop a more extensive and comprehensive assessment plan that goes beyond learning outcomes assessment as mandated by the institution.

During class discussions, we talked about the challenge of acknowledging subjective biases when attempting to report in a seemingly objective manner. The challenge rested in evaluating something that they were in the process of experiencing: their own graduate preparation program. Through our class discussions I emphasized the importance of looking at multiple data points in order to arrive at numeric ratings. As we addressed each of the parts of the SAG I asked students to mark their ratings based on their own experiences and perceptions. Then students put those ratings aside and attempted to make their ratings based on the data that had been collected. More often than not students' data-based ratings varied from those recorded from their personal experiences and perceptions.

Reporting findings/results

The students synthesized their findings and produced a 30-page report. This extensive report was structured based on the parts of the SAG and each section consisted of an item-by-item breakdown of the numeric ratings with a summary of the analysis that led to the rating of each item.

The final report that the students generated was shared with program faculty, and the department chair. The students expressed concern about being identified in the event there were items in the report that were (or perceived to be) negative. I addressed this by only including the course number and semester on the report. With this step taken the students indicated that they felt they could be honest in writing up the report— in the event of any negative findings. However, students remained anxious about the possibility of backlash in the event of negative findings due to the small number of students in a single section of the course. Their worries stemmed from their position as students and the power differential between themselves and the stakeholders interviewed as part of the data collection process.

Lessons Learned

The lessons learned from the use of the CAS Standards for a program self-study that are addressed in this section are focused on programmatic efforts and not directly on student learning. The self-study was the graduate program's first foray into formalized assessment and helped to establish a foundation for a culture of assessment. The institution requires student learning outcomes assessment but overall program evaluation is not required. The faculty wanted to capitalize on opportunities to assess the graduate program because it is new but also wanted some form of a baseline to guide future assessment efforts. The student-written report has led program faculty to develop a more extensive and comprehensive assessment plan that goes beyond learning outcomes assessment as mandated by the institution. The plan includes the continuation of the self-study as part of the assessment course, alumni surveys, benchmarking of comparable graduate programs at peer institutions, and data on internship and job placements.

In its initial offering, six students were enrolled in the course. With a small number of students in the course I was able to divide the nine parts of the SAG among students and also had students collaborate on some of the parts of the SAG that were more labor intensive. I was concerned about workload and did not have students extend institutional comparisons beyond what was available through peer-program websites and graduate program directories. Through the CAS self-study students demonstrated learning on multiple fronts. Students learned a hands-on approach to assessment via self-study. The biggest learning takeaway students demonstrated was the collection, evaluation, and reporting based on multiple sources of evidence. As demonstrated in their written reports, students analyzed data from documents, interviews, and institutional data to draw conclusions and make recommendations.

In the future, with more students enrolled, I will divide the parts differently so that students are gaining experience in gathering, analyzing, and interpreting data from multiple sources, across multiple parts of the SAG, as well as synthesizing and reporting findings. One area I did not address in this first attempt was to report findings in varied formats and for varied audiences. While we discussed various reporting formats in the course, students did not gain direct experience.

Conclusion

When I began writing this article my initial intent was to reflect on the use of the CAS Standards as a tool for teaching an aspect of assessment. By shifting my reflection to the form of a publication I engage in a key aspect of the scholarship of teaching and learning that Shulman (2001) labels as professionalism. As a member of professional and scholarly communities I have a responsibility to share what I learn through teaching (Shulman, 2001). By sharing my reflections from aspects of a course that I teach not only do I share what I have learned but I am also making my teaching available for public view and critique (Ginsberg & Bernstein, 2011).

The approaches I took in using the CAS Standards in a graduate-level course do not have to be exclusive to formal courses. Similar approaches can be taken in concert with efforts to build and sustain cultures of assessment (see Culp & Dungy, 2012). A CAS self-study can serve as a great tool for staff within a department to learn aspects of assessment and evidence-based decision making. I have seen a CAS self-study process modified to be conducted completely by within-unit staff as a precursor to a review by an external team.

References

Bresciani, M. J., & Todd, D. K. (Committee co-chairs). (2010). ACPA/NASPA professional competency areas for student affairs practitioners. Washington, D.C.: A Joint Publication of College Student Educators International and Student Affairs Administrators in Higher Education.

Council for the Advancement of Standards in Higher Education (2012). *CAS professional standards for higher education* (8th ed.). Washington, DC: Author.

Council for the Advancement of Standards in Higher Education (2013). *CAS Self-Assessment Guide for Masters-Level Student Affairs Preparation Programs*. Washington, DC: Author.

Council for the Advancement of Standards in Higher Education (2015). CAS purpose. Retrieved from http://www.cas.edu/mission

Council for the Advancement of Standards in Higher Education (2016). CAS program review. Retrieved from http://www.cas.edu/programreview

Creswell, J. W. (2015). A concise introduction to mixed methods research. Los Angeles, CA: SAGE.

Culp, M. M., & Dungy, G. J. (Eds.) (2012). *Building a culture of evidence in student affairs: A guide for leaders and practitioners*. Washington, DC: National Association of Student Personnel Administrators.

Ginsberg, S. M., & Bernstein, J. L. (2011). Growing the scholarship of teaching and learning through institutional culture change. *Journal of the Scholarship of Teaching and Learning, 11*(1), 1–12.

Mitchell, A. A. (Committee chair). (2006). The ACPA ASK Project: Assessment Skills and Knowledge Content Standards for Student Affairs Practitioners and Scholars. Washington, D.C.: American College Personnel Association.

Shulman, L. (2001). From Minsk to Pinsk: Why a scholarship of teaching and learning? *Journal of the Scholarship of Teaching and Learning, 1*(1), 48–53.