RESEARCH & PRACTICE IN ASSESSMENT

VOLUME THIRTEEN | SUMMER/FALL 2018
www.RPAjournal.com
ISSN# 2161-4210

RPA

A PUBLICATION OF THE VIRGINIA ASSESSMENT GROUP

RESEARCH & PRACTICE IN ASSESSMENT

CALL FOR PAPERS

Research & Practice in Assessment is currently soliciting articles and reviews for its Summer 2019 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 March 1 will receive consideration for the summer 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 | Copyright © 2018

TABLE OF CONTENTS

4 **FROM THE EDITOR**

Sharpening the Ax
- Katie Busby

5 **ARTICLES**

Categorizing College Students Based on Their Perceptions of Civic Engagement Activities: A Latent Class Analysis Using the Social Agency Scale
- Dena A. Pastor, Thai Q. Ong, and Christopher D. Orem

22 **Contextualizing Effect Sizes in the National Survey of Student Engagement: An Empirical Analysis**
- Louis M. Rocconi and Robert M. Gonyea

39 **Learning Assessment in Student Affairs Through Service-Learning**
- Blanca Rincón and Milagros Castillo-Montoya

51 **Five Years of Video-Based Assessment Data: Lessons from a Teacher Education Program**
- Peter D. Wiens and Matthew D. Gromlich

62 **The Dependability of the Updated NSSE: A Generalizability Study**
- Kevin Fosnacht and Robert M. Gonyea

75 **BOOK REVIEW**

Book Review of: Demonstrating Results: Using Outcome Measurement in Your Library
- Beyza Aksu Dunya

78 **NOTES IN BRIEF**

Two Underused, Best Practices for Improvement-Focused Assessments
- Phyllis Blumberg

FROM THE EDITOR

Sharpening the Ax

Abraham Lincoln said, "Give me six hours to chop down a tree, and I will spend the first four sharpening the ax." His wise words apply to many endeavors, including higher education assessment. Faculty, assessment practitioners and scholars, student affairs educators, and administrative leaders alike should dedicate time and effort to ensuring their assessment tools are "sharp". By using a sharp ax - or in this case appropriate methods and measures - decision makers will have the necessary and accurate information to improve student learning. The contributions presented in this issue of *Research & Practice in Assessment* represent a sharpening of assessment axes and demonstrate important advancements in the practice and scholarship of assessment.

The Summer/Fall 2018 issue of RPA includes five peer-reviewed articles that exhibit the development and use of assessment measures. Pastor, Ong, and Orem use latent class analysis to categorize students' civic engagement activities as part of a general education curriculum. The National Survey of Student Engagement (NSSE) is the focus of two articles in this issue. First, Rocconi and Gonyea examine effect sizes within the context of NSSE offering users of this instrument recommendations for interpreting effect sizes in the context of the survey. Later in this issue, Fosnacht and Gonyea use Generalizability Theory to examine the dependability of NSSE Engagement Indicators and consider the sample size needed to draw appropriate conclusions. Rincon and Castillo-Montoya offer a qualitative study demonstrating how student affairs graduate students learn assessment best practices and apply those skills through a service-learning course. Weins and Gromlich examine the use of the Video Assessment of Interactions and Learning (VAIL) to assess learning in teacher preparation programs and offer recommendations for utilizing this instrument.

Beyza Asku Dunya reviews *Demonstrating Results: Using Outcome Measurement in Your Library*, by Rhea Joyce Rubin, a text that provides guidance and best practices for assessment student learning outcomes foster by programs with university libraries. This issue also includes a Notes in Brief by Blumberg highlighting two assessment practices – anticipating use of assessment results and identifying academic bottlenecks – that are often underused, but can be very effective.

I hope this issue of *Research & Practice in Assessment* will serve to sharpen your approach to assessment practice and scholarship.

Regards,

Katie Busby

University of Mississippi

RESEARCH & PRACTICE IN ASSESSMENT

Abstract

A common approach to assessing one facet of civic engagement (CE) is through administering the Cooperative Institutional Research Program's (CIRP) social agency scale, which captures the extent to which respondents feel personally responsible to be involved in addressing various social and political issues. To summarize the scale's results in a manner that conveys the type of CE activities college students consider important, the current study used latent class analysis (LCA) with responses from 2,591 students. A 4-class solution was favored with one class considering all activities important, another class considering few activities important, and two other classes differing in the extent to which they preferred political to nonpolitical activities. Validity analyses partially supported the 4-class solution. Implications of the results for the development of CE programming are discussed, with particular attention paid to the relative emphasis of nonpolitical and political CE on college campuses.

AUTHORS

Dena A. Pastor, Ph.D.
James Madison University

Thai Q. Ong, M.A.
James Madison University

Christopher D. Orem, Ph.D.
James Madison University

Categorizing College Students Based on Their Perceptions of Civic Engagement Activities: A Latent Class Analysis Using the Social Agency Scale

The State Council for Higher Education in Virginia (SCHEV) recently added civic engagement (CE) as a core competency, which is an area of knowledge and/or skills considered essential to the success of all undergraduates regardless of their discipline or institution (State Council for Higher Education in Virginia, 2017). CE now shares the same status as critical thinking, written communication, and quantitative reasoning in being one of the required areas for assessment by all Virginia institutions. SCHEV's move to elevate the status of CE corresponds with recent calls to reinvigorate higher education's civic mission across the nation. For instance, arguments for a renewed focus on CE were made in "A Crucible Moment," a 2012 report commissioned by the U.S. Department of Education (National Task Force on Civic Learning and Democratic Engagement, 2012).

Defining CE

Given the attention institutions are encouraged to devote to this competency, it is important to provide a definition. A popular definition is provided by Ehrlich (2000):

> Civic engagement means working to make a difference in the civic life of our communities and developing the combination of knowledge, skills, values, and motivation to make that difference. It means promoting the quality of life in a community, through both political and nonpolitical processes. (p. vi)

A notable feature of this definition is the inclusion of both political and nonpolitical processes. These two types of processes align with two areas from which much of our understanding of CE is derived: community service-learning, which is largely nonpolitical

CORRESPONDENCE

Email
pastorda@jmu.edu

in nature, and political engagement (Finley, 2011; Reason & Hemer, 2015). Community service-learning programs are often characterized by the pairing of learning with community service, with the programs providing an experiential learning experience for the student while at the same time addressing a community need. In contrast, political engagement programs emphasize the systems, policies, and societal structures that contribute to the community need. To clarify the distinction[1], consider a student in a leadership class who works with an area food bank to organize a food drive. This is an example of non-political community service or non-political civic engagement (NPCE). If instead the student investigates and takes action to affect the systems, policies, and structures that contribute or cause people in the community to go hungry in the first place, the activity is an example of political civic engagement (PCE). If the student organizes the food drive and also investigates and takes action to affect the causes of hunger, the CE activity has both political and non-political elements and is best classified as PCE.

Recognizing that CE activities can be classified as being NPCE, PCE, or both NPCE and PCE, leads to the question of what kinds of activities *should* be promoted at an institution.

Recognizing that CE activities can be classified as being NPCE, PCE, or both NPCE and PCE, leads to the question of what kinds of activities *should be* promoted at an institution. One factor to consider when answering this question is the kind of training students need, which can be understood through assessment. If assessment reveals that students are well-prepared for one kind of CE but not the other, a university might decide to devote more resources to the area in need of development.

Assessing Social Agency: Different Approaches to Summarizing and Presenting Results

A comprehensive CE assessment approach would address a wide array of knowledge, skills, values, attitudes, and behaviors. In this paper, we focus on only one aspect of the value component, which is social agency, described by Eagan et al. (2017) as "the extent to which students value political and social involvement as a personal goal" (p. 56). A popular approach to the assessment of social agency includes a collection of items that have been used for over 40 years by the Higher Educational Research Institute (HERI) in the CIRP surveys. Various civic activities are presented to students (e.g., helping others who are in difficulty, promoting the political structure) who rate the importance of each activity to them personally. The same or similar items appear on the civic action subscale of the Civic Attitudes and Skills Questionnaire (Moely, Mercer, Ilustre, Miron, & McFarland, 2002) and the Political and Social Involvement scale (Center of Inquiry in the Liberal Arts, 2013), which is used in the Wabash National Study, a longitudinal study of college student learning and developmental outcomes.

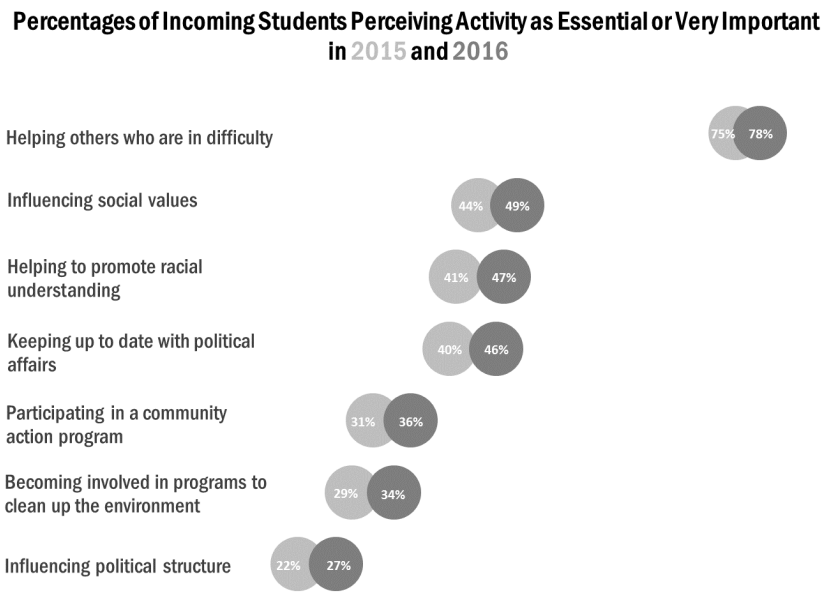
A comprehensive CE assessment approach would address a wide array of knowledge, skills, values, attitudes, and behaviors.

An important consideration when using the social agency scale to inform the development or effectiveness of programming is how to summarize and present the results. There are three possibilities. The first approach is to summarize and present the results for each item. That is, the frequencies of responses for each item are calculated and compared across items. To illustrate, Figure 1 provides results from the administration of the American Freshman Survey to entering college students at four-year U.S. colleges and universities in 2015 and 2016 (Eagan et al., 2015; Eagan et al., 2017). This presentation of results is useful for conveying the typical response to each item, with results indicating the majority of students believe it is important to help others and far fewer believe it is important to influence the political structure.

A second approach to presenting the results is to compute a single score from the items, either by summing the item responses or using item response theory to estimate a theta value for each student. Several researchers have used a single score for the items in their studies (e.g., O'Neill, 2012; Pascarella, Ethington, & Smart, 1988; Rhee & Dey, 1996). Although there is some support for the unidimensionality of the items (Lott & Eagan, 2011),

[1] Example adopted from Westheimer and Kahne (2004). However, they used this example to make the distinction between participatory citizens and justice-oriented citizens, not between NPCE and PCE.

Figure 1



a single score is not useful if the purpose in using the scale is to understand the types of CE activities students deem important. For instance, if the items in Figure 1 were summed to produce a single score and a student received a score of five, we would know the student considers five of the seven activities important but we would not know which activities they consider important.

Use of a classification technique with the social agency items is useful over a single score because it conveys the types of activities different classes of students deem important.

A third approach to presenting the results involves classifying students into classes based on their patterns of responses to the items. The term "classes" instead of "groups" is used with classification techniques to distinguish categorizations of persons created by the analysis (classes) from existing categorizations of persons (groups). Using classification techniques such as cluster analysis or latent class analysis (LCA), the number and nature of different classes with different profiles of responses across items can be captured. For instance, use of these techniques might indicate there is a class of students who value all activities and another class that favors only nonpolitical activities. Use of a classification technique with the social agency items is useful over a single score because it conveys the types of activities different classes of students deem important. Classification techniques are also advantageous over the overall results provided for each item (as in Figure 1) in being better able to capture the variability among students in their civic preferences as well as covariability among item responses. Thus, applying classification techniques to students' responses can reveal an abundance of new information about students and their perceptions of CE activities that are unobtainable when the responses are summarized using the previous two approaches.

To date, classification techniques have not been used with these items exclusively[2], but these techniques have been used with other CE measures to classify people into different categories based on their CE preferences or behaviors (see Table 1). The studies in Table 1 differ greatly from one another in the variables and analyses used to classify individuals and in the individuals classified. Despite these differences some common classes were identified. Almost all studies found a small-to-medium-sized class of what Weerts, Cabrera, and Meijas

[2] Rios-Aguilar and Mars (2011) used the social agency items in a classification study employing cluster analysis with data from CIRP's 2005 Continuing Senior Survey. The social agency items were separated into two different subscales (i.e., Community Action and Political Action) along with other items. These subscales were used along with six other subscales and demographic variables to classify students into classes. Because demographic variables were used to create classes and more importantly, because the resulting classes only differed meaningfully in their demographies, the results are not included in Table 1.

Table 1
Summary of Previous CE Classification Studies

Study	Sample	Indicators	Classification Technique	Findings
Lopez et al. (2006)	1,700 young adults, ages 15-25	19 questions on the Civic and Political Health of a Nation Survey about participation in NPCE and PCE activities	Classified by number and type of activity	4 classes: *Electoral specialists* participated in at least two PCE activities (17%); *civic specialists* participated in at least two NPCE activities (12%); *disengaged* did not meet the criteria for either class (58%); and *dual activists* met the criteria for both classes (13%)
Moely, Furco, & Reed (2008)	2,000+ college students enrolled in service learning courses across various institutions	Questions about preference of engagement in service-learning activities aligned with the charity paradigm (similar to NPCE) and social change paradigm (similar to PCE)	Median split	4 classes: the *social change preference* class (16%) preferred only social change paradigm activities; the *charity preference* class (20%) preferred only charity paradigm activities; *low value undifferentiated preference* class (29%) did not prefer either kind of activity; and the *high value undifferentiated* class (35%) preferred activities in both paradigms
Weerts, Cabrera, & Meijas (2014)	1,876 recent graduates from bachelor degree programs between 1999 and 2003	Items on the ACT Alumni Outcomes Survey asking about level of involvement in various kinds of organizations (e.g., environmental, political, social)	LCA	4 classes: *apolitical engagers* (39%) were characterized by involvement in professional, service, social and community organizations but low involvement in political or environmental groups; *social-cultural engagers* (6%) were characterized by a high involvement in social and cultural organizations; *non-engagers* (25%) were characterized by low involvement in all organizations; and *super engagers* (30%) were characterized by high involvement in all organizations
Brunton-Smith (2011)	Survey data collected from adults in several countries in the European Union	Variables capturing participation in different kinds of civic activities: voting in the national election, conventional political participation beyond voting (e.g., campaigning or donating money), nonconventional political participation (e.g., boycotting, signing a petition, protesting), and involvement in nonpolitical organizations	LCA	4 classes: the *voters only* class (41%) voted, but were not involved in other ways; the *non-conventional participation* class (9%) participated in politics in non-conventional ways and in nonpolitical organizations; the *not politically active* class (13%) were not involved; and the *highly politically active* class (38%) were involved in all areas

Table 1, continued
Summary of Previous CE Classification Studies

Study	Sample	Indicators	Technique	Findings
Torney-Purta (2009)	30,000 14-year olds in 10 European countries during 1999	12 social and political attitudinal scales administered as part of the Civics Education Study by the Institute of Educational Sciences	Cluster analysis	5 classes: the *social justice* class (17%) characterized by "I believe in rights for everyone but do not feel obligated to do much about it" (p. 829); the *conventionally political* class (33%) characterized by "I believe in my country and will support the status quo with positive political and civic actions that are expected of me" (p. 829); the *indifferent* class (9%) and *disaffected* class (35%) both characterized by "I have better ways to spend my time than thinking about being active in politics, but I won't do anything rash" (p. 830) with the indifferent class having more negative beliefs about minorities' rights and norms of citizenship; the *alienated* class (7%) characterized by: "I'm angry about the immigrants and minority groups in my country, and I don't trust the government; I have the right to do what I want" (p. 830)

(2014) call *super engagers*, or individuals who prefer or engage in both NCPE and PCE activities (Lopez et al., 2006; Moely et al., 2008) All studies also identified a class of *non-engagers*, or individuals who do not prefer or engage in either NCPE and PCE activities (Lopez et al., 2006; Moely et al., Torney-Purta, 2009; Weerts et al., 2014). Some studies also found a relatively small class of *political engagers* who preferred or engaged in PCE activities over NCPE activities (Lopez et al., 2006; Moely et al., 2008). *Non-political engagers*, or those who prefer NCPE activities over PCE activities, were also identified as a small-to-medium-sized class by some studies (Lopez et al., 2006; Moely et al., 2008; Weerts et al., 2014).

Purpose of the Study

To date, classification techniques have not been used to categorize college students according to the importance they assign to various CE activities. Because the *kinds* of CE activities students value may be more informative than the *number* of CE activities they value, the present study performs LCA using the social agency items in Figure 1 to classify students into classes according to the kind of activities they deem important. Understanding what kinds of classes exist is useful for two primary reasons. First, the results can be informative to the development of CE initiatives on campus. For instance, if a large class of non-political engagers is identified a campus might decide to place more emphasis on helping students connect politics with their NCPE experiences or create and promote PCE initiatives. Second, the results are also useful for assessment purposes. For example, if action is taken on a campus to promote PCE activities, the percentage of students in classes that value both NCPE and PCE can be compared before and after the promotion. The membership of the same student in various classes can also be tracked over time. For instance, it would be favorable to find a student who started college as a non-engager transition during their academic career to a class that valued one or both types of CE.

It is important for researchers to validate the identified classes because classes that emerge in LCA may be an artifact of the data and not true qualitatively different groups of students.

Given the potential utility of a social agency typology, we conducted LCA on the social agency items to address the following research questions:

1. In how many different ways might students be categorized with respect to their CE preferences? In other words, how many different classes exist?
2. What is the nature of the classes? How might the classes be characterized with respect to the importance they assign to various CE activities?
3. What percentage of students belong to each class and how accurately can students be classified?
4. In what other ways do the classes differ?

Based on the results of other classification studies in the CE literature we anticipated we might find one or more of the following classes: super engagers, who value both NPCE and PCE activities, non-engagers, who find little value in CE activities, non-political engagers, who value NPCE activities more than PCE activities, and political engagers, who value PCE activities more than NPCE activities.

The first three research questions were pursued to describe the number and nature of social agency classes at our university. Based on the results of other classification studies in the CE literature we anticipated we might find one or more of the following classes: super engagers, who value both NPCE and PCE activities, non-engagers, who find little value in CE activities, non-political engagers, who value NPCE activities more than PCE activities, and political engagers, who value PCE activities more than NPCE activities.

The purpose in pursuing the last research question was to provide validity evidence for our LCA solution. Validity evidence for our typology can be obtained by considering how classes differ on variables beyond those used in their classification (i.e., auxiliary variables). Auxiliary variables had to be chosen from those collected at the same time as the social agency items because the data in this study were not collected specifically for this research. Of these variables, those that are used often in CE research were selected with the resulting auxiliary variables including gender, race, and student academic classification (e.g., freshman, sophomore, junior, senior). We also used cohort (i.e., academic year of the response) as an auxiliary variable since data from multiple cohorts were used in our study.

Prior research and knowledge of our campus' practices informed the hypotheses guiding our validity analyses. For instance, because O'Neill (2012) and Lott and Eagan (2011) found that seniors assigned higher levels of importance to social agency items than incoming students, we hypothesized that classes emerging from the analysis characterized by endorsement of more activities would consist of more upperclassmen. We also hypothesized that students in more recent cohorts would be represented in classes where more activities were valued because of our campus' recent heightened emphasis on CE. Because prior classification studies found more females in classes preferring NPCE over PCE (Lopez et al., 2006; Moely et al., 2008), we anticipated the same gender discrepancy in our own study if such a class emerged. We also anticipated more males in classes preferring PCE over NPCE based on findings from other classification studies (Brunton-Smith, 2011; Lopez et al., 2006). Findings regarding racial differences in class membership were mixed across studies. Support for the hypothesis that a larger number of minorities would be found in classes that value both NPCE and PCE or PCE over NPCE is based on Moely et al. (2008), who found non-Whites more likely to be in the class endorsing both types of engagement, and Lopez et al. (2006) and Eagan et al. (2015) who both found that minorities value political involvement more than Whites. In summary, to provide supportive validity evidence for our LCA solution we expected the following hypotheses to be supported:

1. More upperclassmen and students from recent cohorts represented in classes valuing a larger number of civic activities
2. If such classes emerge, more females in classes valuing NPCE over PCE and more males in classes valuing PCE over NPCE
3. A larger percentage of minorities in classes where PCE activities are valued

Methods

The social agency items and auxiliary variables were all collected as part of an annual survey at our university for institutional research and assessment purposes. In the following sections we first describe the general procedures for the survey, participants, and variables used in our study. Then, we describe the details of the LCA and validity analyses.

The majority of our students believe it is important to help others (88%) and far fewer believe it is important to influence the political structure (42%).

Procedure

The survey is administered to a sample of students during the middle of the fall semester each year. A sample of roughly 30% of the 20,000 undergraduate student body is selected, resulting in an overall sample of 6,000 undergraduate students. Because the survey is administered via paper and pencil, only on-campus course sections are selected, resulting in a possible population of 19,000 students. A random sample of on-campus undergraduate course sections is compiled and then manually adjusted to ensure that the sample is representative of the university population concerning important demographic features such as gender, race, and student academic classification (e.g., freshman, sophomore, junior, senior). To maximize the number of survey items while also minimizing survey fatigue, five different versions of the survey are used. All students answer a common set of demographic questions followed by one of five sets of items. The different versions of the survey are distributed randomly throughout each sampled course section such that all versions might be answered by different students in a single section. The items used for this research all came from one version of the survey.

Participants

Data collected in three different years were combined to create the data set used in the analyses[3]. The final sample consisted of 2,591 students with 27%, 47%, and 27% from the 2013/2014, 2015/2016, and 2016/2017 administrations[4], respectively. The distribution of gender and race aligns with the overall distribution at our university, with 62% of the sample identifying as females and 81% of the sample identifying as White. Students were fairly evenly distributed across credit-hour categories, with 20% having completed fewer than 28 credit hours (freshman), 25% having completed between 28 and 59 credit hours (sophomores), 27% having completed between 60 and 89 credit hours (juniors) and 29% having completed more than 89 credit hours (seniors).

As hypothesized, there were significant differences among classes in gender composition, with females more likely to be classified as non-political engagers and males more equally dispersed across classes, including the political engagers class.

Variables

Latent class analysis variables. The CIRP social agency items[5] were used to classify students into categories using LCA. Students originally responded to these items using a four-point Likert scale (1 = *Essential*; 2 = *Very Important*; 3 = *Somewhat Important*; 4 = *Not Important*). Due to the skewed distributions of responses, with most reporting either Essential (1) or Very Important (2), we decided to collapse the four response categories into two response categories to avoid estimation issues and simplify the interpretations of the results. Thus, the two response categories included in our analyses were Important (1), which

[3] Because data collected across different years were combined, it is possible for a single student to be represented multiple times in our final data set. For example, if a student were randomly selected to complete the survey in both 2013/2014 and 2015/2016 they would be represented twice in the data. Because no identifying information was collected from students we cannot ascertain the extent to which this occurred, although we suspect it is rare. To clarify, consider a student attending the university during all three years of data collection, where the probability of being selected for the survey is .30 (because we are obtaining a random sample of 30% of the student population). The probability of this student being randomly selected to complete the survey twice is .09 ($.30^2$) and three times is .03 ($.30^3$). Therefore, it is possible but unlikely for the same student to be surveyed multiple times. Given the infrequency with which this is likely occurring the impact on our results is suspected to be negligible.

[4] For reasons unrelated to this research, the survey was not administered in 2014/15.

[5] Items are from the 2017 Cooperative Institutional Research Program (CIRP) Freshman Survey (Eagan et al., 2017). These items were used with permission from the Higher Education Research Institute.

included Essential and Very Important, and Not Important (0), which included Somewhat Important and Not Important. The same approach to collapsing response categories is used in the reporting of the results for these items by CIRP (Eagan et al., 2015; Eagan et al., 2017).

Although we hypothesized that minorities would have a stronger representation in classes favoring PCE activities, our results indicate that minorities have a stronger representation in the super engager class favoring both NCPE and PCE activities.

Auxiliary variables. Once the final LCA solution was obtained (i.e., the best fitting LCA was determined), we conducted validity analyses to ascertain whether the resulting categorizations of students aligned with prior research. As mentioned above, we used gender (female; male), race (White; non-White), student academic classification (freshman; sophomore; junior; senior), and cohort (2013/2014; 2015/2016; 2016/2017) as auxiliary variables.

Data Analysis

Latent class analysis. We conducted a series of LCAs on the social agency items to explore if different types (classes) of students exist who differ in how much they value involvement in various civic activities. We initially fit a one-class model to the data and in subsequent analyses we increased the number of classes (C) by one. We followed this model-building procedure until estimation issues were encountered. The equation for the general C-class LCA model with binary indicators is presented below, where j is used to refer to item j, with there being $j = 1$ to J items, and c is used to refer to a specific class, with there being $c = 1$ to C classes:

$$P(x_j = 1) = \sum_{c=1}^{C} \rho_c \, P(x_j = 1 | \, c)$$

The general C-class LCA equation specifies the marginal probability of endorsing Important on item j, $P(x_j=1)$, as equal to the weighted sum of the conditional probability of endorsing Important on item j in each class, $P(x_j=1| \, c)$. The weights, ρ_c, represent the proportion of students in each class c. The number of estimated parameters in the general C-class LCA model depends on the number of items (J) and classes (C). For example, in a 2-class LCA model with seven dichotomous items, a total of 15 parameters are estimated: one class weight[6] and 14 conditional probabilities (7 items x 2 classes).

We estimated all LCA models using full information maximum likelihood (FIML) estimation via the Expectation Maximization (EM) algorithm in Mplus version 7.3 (Muthén & Muthén, 1998–2012). A common concern when estimating LCA models is converging on a local maxima. To avoid this issue Mplus implements a two-stage estimation procedure in which multiple sets of random start values are first generated and optimized up to 10 iterations (initial stage). Then, the best sets of random start values (i.e., the ones with the highest likelihood of producing the data) are used as starting values in the subsequent step and optimized to completion (final stage). We specified a random start value of 1,000 and final stage optimization value of 500 for our study. Thus, for each LCA model, Mplus generated 1,000 sets of random start values and optimized them to 10 iterations. Then, Mplus used the best 500 sets of random start as starting values in the subsequent step and optimized them to completion to obtain the final model solution.

Model fit. We examined model-data fit via the log-likelihood (*LL*), Bayesian information criterion (BIC; Schwarz, 1978), and sample-size adjusted *BIC* (*SSABIC*; Sclove, 1987). The *LL* for each model represents the likelihood of the data given the specified estimated model parameters. *LL* values closer to zero indicate a higher likelihood of the data and thus, better model-data fit. Because *LL* values will always be closer to zero for more complex models (e.g., models with more classes), we also examined model-data fit via two information criteria measures: *BIC* and *SSABIC*. The *BIC* and *SSABIC* penalize the *LL* for

[6] Only C-1 weights are estimated because the weights, ρ_c, are constrained to be positive and to sum to one across classes.

model complexity in different ways, with smaller values indicating more superior model-data fit. The *BIC* and *SSABIC* have been shown to perform well in simulation studies (Henson, Reise, & Kim, 2007; Tofighi & Enders, 2008). We championed the model with the lowest *BIC* and *SSABIC* values as the best-fitting model in our study.

Model comparison. We compared models differing in the number of classes using the Lo-Mendell-Rubin likelihood test (*LMRT*; Lo, Mendell, & Rubin, 2001), bootstrap likelihood ratio test (*BLRT*; McLachlan & Peel, 2000), and the approximate Bayes factor (*BF*). The *LMRT* and *BLRT* compare a *C* class model to a *C*-1 class model. A significant *LMRT* or *BLRT* would indicate that the model with *C* classes fits the data significantly better than the model with *C*-1 classes. The approximate *BF* compares the *BIC* values between two models ($BF_{1,2}$),

$$BF_{1,2} = \exp[(-0.5BIC_1) - (-0.5BIC_2)]$$

where BIC_1 and BIC_2 represent the *BIC* values associated with model one and model two (e.g., one-class model and two-class model). A *BF* value greater than one would imply that model one is more strongly supported by the data than model two (Wasserman, 2000).

Validity analysis. It is important for researchers to validate the identified classes because classes that emerge in LCA may be an artifact of the data and not true qualitatively different groups of students. A variety of methods have been developed to obtain validity evidence in LCA. One simple method is to modally assign students to classes based on their highest posterior probability and use the new class membership variable in subsequent traditional analyses (e.g., ANOVA, regression). To clarify, consider a 2-class model. Each individual in a 2-class model has two posterior probabilities: one conveying their probability of membership in Class 1 and another conveying their probability of membership in Class 2. Thus, a fictitious individual might have posterior probabilities of .85 and .15 for Classes 1 and 2, respectively. The new class membership variable captures the class for which the posterior probability is the highest, which would be Class 1 for our fictitious individual. Once the new class membership variable is created traditional analyses can be used to relate it to other variables. This method, however, assumes perfect classification accuracy (i.e., all posterior probabilities are one or zero). For this reason, other methods that account for classification accuracy have been developed (e.g., 3-step method, Lanza, and BCH). The choice among the latter methods is dependent on whether (a) the auxiliary variables are treated as predictors or outcomes of class membership and (b) the auxiliary variables are continuous or categorical. In our study we treated gender, race, student academic classification and cohort as categorical predictors of class membership. Given these criteria, we chose to use the 3-step method (Asparouhov & Muthén, 2014; Vermunt, 2010) to conduct our validity analyses, running the analysis separately for each auxiliary variable. In the 3-step method, multinomial regression is used to regress the new class membership variable on auxiliary variable(s) while taking into account the classification accuracy of the model.

One facet of CE that is commonly assessed is social agency, or the extent to which one considers involvement in civic or political activities as a personal goal.

Results

Descriptive Statistics

The percentage of students considering each CE activity important is reported in Table 2. Compared to the percentages based on the dichotomized responses obtained by Eagan et al. (2015) and Eagan et al. (2017) from entering college students shown in Figure 1, a larger percentage of our students perceived the CE activities as being important (see Table 2). Note, however, that Eagan et al. (2015) and Eagan et al. (2017) surveyed only entering college students whereas our sample consisted of a wide range of students at our university. Thus, this may be one reason for the discrepancy. Despite this, the trend of responses was similar. The majority of our students believe it is important to help others (88%) and far fewer believe it is important to influence the political structure (42%).

Table 2
Percentages of Students Considering Activity as "Essential" or "Very Important"

Item	*N*	%
1. Helping others who are in difficulty	2586	88
2. Influencing social values	2586	73
3. Helping to promote racial understanding	2585	54
4. Participating in a community action program	2570	70
5. Becoming involved in programs to clean up the environment	2583	55
6. Keeping up to date with political affairs	2584	57
7. Influencing political structure	2586	42

Note. The sample sizes reported in this table are slightly lower than the final sample size of 2,591 because of missing data. All 2,591 cases were used in the LCA, even those with missing data on one or more items. The LCA estimation procedure, full information maximum likelihood (FIML), accommodates missing data by estimating parameters using all available data. Although this method makes certain assumptions about the missing data mechanism, these assumptions are easier to satisfy than the assumptions made by more traditional missing data techniques (e.g., listwise or pairwise deletion). For further information see Enders (2010).

Latent Class Analysis

Although more research is needed to support the 4-class solution, the validity evidence was mainly supportive; importantly, the nature and number of classes aligned with classes found in other CE classification studies.

We estimated a total of five LCA models. When estimating the 5-class model, we encountered estimation issues. Specifically, the 5-class solution had estimated conditional probabilities that were at the boundary of the parameter space (0 or 1.0). We chose not to interpret the results from the 5-class model and only consider the results from the remaining models because such solutions are typically deemed as untrustworthy (Geiser, 2013).

Model fit. The fit indices for the models are presented in Table 3. The 4-class model, overall, provided better fit to the data compared to the other three models. The *BIC* and *SSABIC* fit indices were lowest for the 4-class model. The *LMRT* and *BLRT* were both statistically significant, which indicated the 4-class model fit significantly better than the 3-class model. Lastly, the *BF* was greater than 10, which suggested the 4-class model is more strongly supported by the data than the 3-class model. The entropy statistic for the 4-class model is .66, which indicates only moderate certainty about classifying individual students into classes.

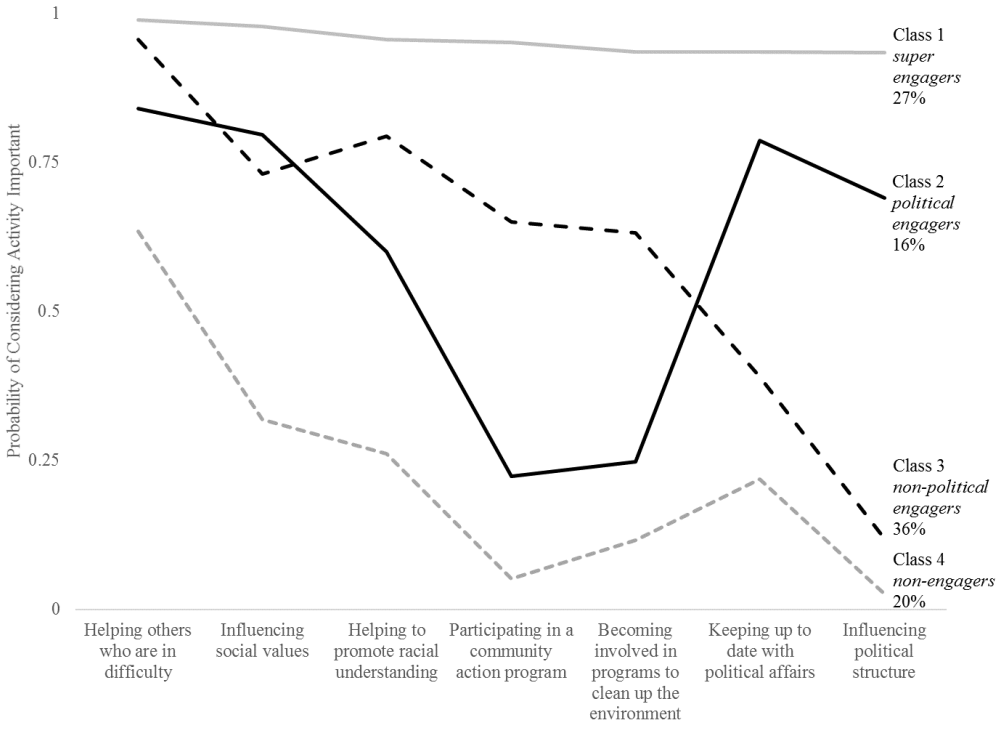
Table 3
Fit Indices and Entropy for the 1-Class, 2-Class, 3-Class, and 4-Class Models

# of classes	# of paras.	*LL*	*BIC*	*SSABIC*	Entropy	*LMRT p*	*BLRT p*	*BF*[a]
1-class	7	-11124	22303	22281	1	---	---	---
2-class	15	-10026	20170	20122	.69	< .01	< .01	> 10
3-class	23	-9839	19860	19787	.69	< .01	< .01	> 10
4-class	31	-9733	19710	19611	.66	< .01	< .01	> 10

Note. # of classes = number of classes; # of paras. = number of parameters estimated;
LL = log-likelihood; *BIC* = Bayesian information criterion; *SSABIC* = sample size adjusted Bayesian information criterion; *LMRT p* = Lo-Mendell-Rubin likelihood ratio *p*-value;
BLRT p = bootstrap likelihood ratio *p*-value; *BF* = Bayes factor
[a] The Bayes factor compared the *C* class model to the *C*-1 class model.

Four-Class Solution. Figure 2 illustrates the probability of considering each CE activity as important based on the 4-class model. The four classes found in our study closely align with those identified by previous researchers. Class 1, which contained 27% of students, was characterized by having high probabilities of considering all CE activities as important. Students in this class resemble individuals previously identified as super engagers (Lopez et al., 2006; Moely et al., 2008; Weerts et al., 2014). Class 2, which contained 16% of students, was characterized by having high probabilities of considering PCE activities as important and low to moderate probabilities of considering NPCE activities as important. Students in

Figure 2



this class resemble individuals previously identified as political engagers (Lopez et al., 2006; Moely et al., 2008). Class 3, which contained 36% of students, was characterized by having low probabilities of considering PCE activities as important and high probabilities of considering NPCE activities as important. Students in this class resemble individuals previously identified as non-political engagers (Lopez et al., 2006; Moely et al., 2008; Weerts et al., 2014). Lastly, Class 4, which contained 20% of students, was characterized by having low to moderate probabilities of considering all CE activities as important. Students in this class resemble individuals previously identified as non-engagers (Lopez et al., 2006; Moely et al., 2008; al., 2006; Moely et al., 2008; Weerts et al., 2014).

Validity Evidence

The validity results are presented in Table 4, which contains the parameter estimates of the multinomial logistic regression models used in the 3-step method for each auxiliary variable (gender, cohort, race, and student academic classification). To aid in the interpretation of the significant results[7], the estimates were used to obtain the predicted probabilities of class membership, also shown in Table 4 along with a detailed interpretation of the findings. Statistically significant differences in class membership that aligned with our hypotheses were found for gender, race, and cohort but not for student academic classification. As hypothesized, there were significant differences among classes in gender composition, with females more likely to be classified as non-political engagers and males more equally dispersed across classes, including the political engagers class.

To promote transition of political engagers to super engagers, programming would need to increase the value these students place in environmental stewardship activities (which may not be seen by some as relevant to CE), and participation in community action programs.

The distribution of class membership also differed across race. Although we hypothesized that minorities would have a stronger representation in classes favoring PCE activities, our results indicate that minorities have a stronger representation in the super engager class favoring both NCPE and PCE activities. Our hypothesis regarding class differences in cohort membership was also supported, with members of the most recent cohort more likely to be classified as super engagers than members in earlier cohorts. The remaining hypothesis was not supported. Latent classes did not significantly differ from one another in student academic classification (e.g., freshman, sophomore, junior, senior).

Discussion

When considering how non-political engagers compare to the super engagers, the largest differences are in the importance placed on political activities, with non-political engagers unlikely to consider these activities important.

Although there may be disagreement on the precise definition of CE researchers agree that the construct is multidimensional and is characterized by a wide array of knowledge, skills, attitudes, values, and behavior. One facet of CE that is commonly assessed is social agency, or the extent to which one considers involvement in civic or political activities as a personal goal. For decades, CIRP surveys have included social agency items, with the same or similar items appearing on other scales. Given the popularity of these items and the potential for their results to inform CE programming and assessment, this study utilized a classification technique to explore if the results could be summarized and presented in a manner more informative than use of a single score or descriptive statistics based on the individual item scores. LCA was used with the responses from students at our university to identify four classes of students who differed in the kinds of CE activities they valued. In the sections below we consider the results of validity analyses (which were mainly supportive of the 4-class solution), the implications of our results for CE programming, limitations of our study, directions for future research, and implications of our results for broad definitions of CE.

Table 4
Validity Results

Auxiliary Varliable	Multinomial Logistic Regression Parameter Estimate and Standard Erros from the 3-step Method							Predicted probabilities of class membership conditional on auxiliary variable				
		Class 2/Class 1		Class 3/Class 1		Class 4/Class 1			Class 1	Class 2	Class 3	Class 4
	Parameter	Value	*SE*	Value	*SE*	Value	*SE*	Category	*super engagers*	*political engagers*	*non-political engagers*	*non-engagers*
Gender	Intercept	**-0.788**	0.130	**0.443**	0.076	**-0.441**	0.092	Female	0.27	0.12	0.43	0.18
	Gender	0.600	0.180	**-0.510**	0.140	**0.355**	0.137	Male	0.27	0.22	0.25	0.25
Race	Intercept	**-0.421**	0.098	**0.347**	0.071	**-0.209**	0.008	White	0.26	0.17	0.36	0.21
	Race	-0.442	0.233	**-0.324**	0.157	**-0.419**	0.175	Non-White	0.34	0.14	0.34	0.18
Student Academic Classification	Intercept	**-0.447**	0.194	0.272	0.143	**-0.304**	0.154	Freshman	---	---	---	---
	Sophomore	0.081	0.255	0.042	0.190	-0.247	0.216	Sophomore	---	---	---	---
	Junior	-0.190	0.265	-0.076	0.189	0.128	0.198	Junior	---	---	---	---
	Senior	-0.119	0.259	0.031	0.186	0.113	0.197	Senior	---	---	---	---
Cohort	Intercept	0.011	0.204	**0.870**	0.150	**0.835**	0.137	2013/2014	0.15	0.15	0.36	0.34
	2015/2016	-0.290	0.238	**-0.404**	0.176	**-1.301**	0.178	2015/2016	0.25	0.19	0.40	0.16
	2016/2017	**-1.210**	0269	**-1.261**	0.191	**-1.950**	0.097	2016/2017	0.43	0.13	0.29	0.14

Note. Class 1 served as the baseline category in all models. Each predictor was represented by one or more dummy coded variables in the model, with females, whites, Freshman, and 2013/2014 serving as the reference categories in the models including gender, race, student academic classification, and cohort, respectively. Coefficients significant at $p < .05$ are shown in bold.

Interpretation. When considering the classes two at a time (e.g., Class 2 versus Class 1), gender was a statistically significant predictor in the vast majority of comparisons. The largest gender discrepancies indicated that females are more likely to be classified as non-political engagers than as political engagers and non-engagers, while males are equally likely to be classified in these three groups. Race was a statistically significant predictor of class membership for only some comparisons. The probability of classification in Class 1 (super engagers) versus Class 4 (non-engagers) was significant, with whites only slightly more likely to be classified as super engagers than non-engagers, and non-whites far more likley to be classified as super engagers. The probability of classification in Class 1 (super engagers) relative to Class 3 (non-political engagers) also was significant, with whites more likely to be classified as non-political engagers than super engagers, and non-whites equally likely to be classified in these two groups. Student Academic Classification was not a statistically siginifcant predictor of class membership; that is, the probability of class membehsip was the same across academic clalssification levels. Because of the lack of statistical significance, predicted probabilities are not reported. Cohort was a statistically significant predictor of class membership the vast majority of the time. For the 2016/2017 cohort, the probability of membership in the super engagers class was more likely than membership in the other classes. The same is not true of the 2013/2014 cohort, who are more likely to be in Classes 3 (non-political engagers) and 4 (non-engagers) relative to Class 1 (super engagers) and equally likely to be in Class 2 (political engagers) relative to Class 1 (super engagers).

Validity Results

Our validity hypotheses were supported for three of the four variables. Classes differed as hypothesized based on gender, race, and cohort membership. Although we suspect the increase in CE programming at our university might explain why members of more recent cohorts were likely to be classified as super engagers, our study does not allow for the exploration of whether the increase in the number of activities valued in recent

[7] In addition to the information in Table 4 we also considered the multinomial logistic regression results using each class as the baseline category in the model. Table 4 provides the results using Class 1 as the baseline category; the results using every other class as the baseline category are provided in the Mplus output and available to readers upon request.

years is a function of CE programming at our university or other factors (e.g., 2016 general election). Although we hypothesized for more upperclassmen to be in classes where a larger number of CE activities is valued (Class 1) our validity results did not support this hypothesis. Therefore, it is reasonable to question both the meaningfulness of our 4-class solution and our hypothesis. For instance, we based our hypothesis about student academic classification on two studies (Lott & Eagan, 2011; O'Neill, 2012) indicating that seniors assigned higher levels of importance to social agency items than incoming students. However, other research studies did not find student academic classification differences among college students when grouped according to their CE activity preferences (Moely et al., 2008). More research is certainly needed to explore these competing explanations. In the meantime, our results offer a first step in understanding the validity of the 4-class solution on which future research can build.

Implication of Results

Although more research is needed to support the 4-class solution, the validity evidence was mainly supportive; importantly, the nature and number of classes aligned with classes found in other CE classification studies. For these reasons, we proceed below in considering the results and their implications for CE programming.

First, we found it encouraging that only 1/5 of the student population in this study was classified as non-engagers (Class 4) and that more recent cohorts had a smaller probability of membership in this class. Of course, the presence of any non-engagers is not ideal. Therefore, an important next step is to consider the characteristics of students in this class. For instance, if particular majors are heavily represented in this class, CE programming might be targeted to such majors. We also found it encouraging that although the probabilities in Figure 2 are low for most activities for non-engagers, the probability is equal to .63 for the item "helping others who are in difficulty." Thus, perhaps an important way to increase the value these students place in CE activities is to convey to them how such activities help others who are in difficulty.

Second, we were encouraged to find nearly 1/3 of students in the super engagers class (Class 1) and a higher probability of membership in this class for more recent cohorts. This class is the most ideal class because all kinds of CE activities—political, environmental, community-oriented—are considered important. Because this class is ideal, it is important to consider how the political engagers (Class 2) and non-political engagers (Class 3) differ from super engagers. The political engagers are similar to the super engagers in having high probabilities on the items with the exception of low probabilities on two items: one asking about participation in community action programs and another asking about involvement in environmental programs. To promote transition of political engagers to super engagers, programming would need to increase the value these students place in environmental stewardship activities (which may not be seen by some as relevant to CE), and participation in community action programs. With respect to the latter, it is possible that some students, including those in the political engagers class, have a low endorsement of this item[8] because they do not understand what is meant by "community action programs". We personally consider this description vague and suspect that is why it does not appear on the Political and Social Involvement scale (Center of Inquiry in the Liberal Arts, 2013).

Although it is tempting to classify the individual students in our sample into the four classes so that we might be better able to direct them to suitable CE programs on campus, the moderate classification accuracy of our model prohibits us from doing so.

When considering how non-political engagers compare to the super engagers, the largest differences are in the importance placed on political activities, with non-political engagers unlikely to consider these activities important. It is encouraging that non-political engagers value many activities, but the low endorsement of PCE activities is troubling, particularly given the size of this class. Universities can help non-political engagers transition to super engagers by providing and promoting PCE programming and helping students consider their NPCE activities through a political lens.

Limitations of Study & Directions for Future Research

In the above section we considered different actions that might be taken to help develop students in various classes. Although it is tempting to classify the individual students in our sample into the four classes so that we might be better able to direct them to suitable CE programs on campus, the moderate classification accuracy of our model prohibits us from doing so. To clarify, it is important to understand how individual students would be assigned to classes. Assignment of individuals to classes involves the use of the posterior probabilities of class membership for each student, which here would be four values capturing the probability of the student's membership in each of the four classes. In an ideal situation, the probability would be one for single class and zero for the remaining classes. As indicated by our entropy value of .66, the classification accuracy of our model is not perfect, so use of the posterior probabilities to assign individuals to classes is not straightforward. Although a less than perfect entropy value does not affect our use of the LCA result to understand the number and nature of latent classes it does affect our use of the results to classify individual students. Thus, the moderate entropy value does not discount our results; it just cautions the use of results for the classification of individual students. To use LCA with these items in this population to classify individuals, steps would need to be taken to increase its classification accuracy. This can be accomplished by using more items or better quality items (i.e., those useful for discriminating among classes) or by including predictors of latent class membership in the analysis.

Exploring the extent to which the results replicate across institutions is needed, with the CIRP surveys or the Wabash National Study being ideal data sources for such an investigation.

One of the largest limitations in our study is the sample, which includes students at only one university. Exploring the extent to which the results replicate across institutions is needed, with the CIRP surveys or the Wabash National Study being ideal data sources for such an investigation. The variables included in our analyses were also not ideal. Because the data were collected for another purpose, we were limited in what auxiliary variables could be used and based our hypotheses on research that sometimes was not strongly aligned with the present research. Future research should consider other auxiliary variables, such as student's major or their actual civic engagement behaviors, that may yield stronger hypotheses with respect to class differences.

Another suggestion for future research is to consider the extent to which socially desirable response behavior (Spector, 2004) is influencing the results. Although our validity results suggest that most *super engagers* are students who value multiple civic engagement activities, it is possible that this class is also capturing students who are prone to socially desirable response behavior. Exploring the extent to which members in this class are prone to such behavior is warranted. If socially desirable responding is considered an issue, the use of different item types less susceptible to socially desirable responses (e.g., forced-choice items) should be pursued (Christiansen, Burns, & Montgomery, 2005).

Future research should consider other auxiliary variables, such as student's major or their actual civic engagement behaviors, that may yield stronger hypotheses with respect to class differences.

We also have concerns about the social agency items used to classify students in the present study. Having students verbalize their thoughts while reading and responding to items would be useful to ensure that respondents understand the items and are interpreting them in the same way because the language used in some of the items is vague. The results of Sequiera, Holzman, Horst, and Ghant (2017) underscore the need to ensure respondents understand the terms used in CE assessments. When Sequeira et al. (2017) asked college students to describe the ways in which their community service experience related to a current social justice issue several students reported that they did not know what was meant by "social justice." A study examining respondents' understanding of items is worth pursuing if the items continue to be used. But should these items continue to be used? Is this list of activities current and comprehensive if we are trying to capture the kind of civic and political activities students value? We believe these are important questions to address before moving forward in this line of research.

[8] Interestingly, this item also has the largest amount of missing data (see Table 2). It is possible that students did not respond to this item because they did not understand it.

Other limitations in our study are more methodological. We recognize that we engaged in the frowned-upon practice of dichotomizing variables, which results in a loss of information (MacCallum, Zhang, Preacher, & Richer, 2002). We did try LCA using responses on their original 4-point scale but quickly encountered computational issues. Researchers with larger data sets from multiple institutions may not encounter these issues and are encouraged to explore LCAs with the original responses if possible and dichotomized responses if not.

Our final limitation has to do with the narrow aspect of CE assessed by the social agency items included in our study. Our study only provided information on how different classes of students valued different kinds of civic activities; it did not characterize student differences with respect to the many other aspects of CE (e.g., knowledge, skills, motivations, attitudes, behaviors) that exist. To do so, a measure addressing multiple facets is needed, with the Civic Competency and Engagement assessment (Torney-Purta, Cabrera, Roohr, Liu, & Rios, 2015) being a promising assessment for such research.

Although our validity results suggest that most super engagers are students who value multiple civic engagement activities, it is possible that this class is also capturing students who are prone to socially desirable response behavior. Exploring the extent to which members in this class are prone to such behavior is warranted.

Implications of a broad definition of CE

In the beginning of this paper we provided commonly used definitions of CE that encompassed both political and non-political processes. Advantages to adopting a broad definition of CE are its inclusiveness, allowing many activities to be subsumed under single heading, and its flexibility, allowing universities to focus on those aspects of CE that best align with their unique strengths. There are disadvantages, however, to including NPCE and PCE within the larger umbrella of CE. One potential disadvantage is the risk of PCE getting lost within the broader CE initiative. As highlighted by the results of this study and several others, many students value NPCE activities over PCE activities. Use of a broad definition therefore runs the risk of PCE, which needs to be emphasized on campuses, not receiving enough attention if it is subsumed under the larger CE umbrella. PCE initiatives on campus should be highlighted, and the link between NPCE and politics made explicit, in order to increase students' political involvement and help them see political action as an avenue for helping others.

References

Asparouhov, T., & Muthén, B. (2014). Auxiliary variables in mixture modeling: Three-step approaches using Mplus, *Structural Equation Modeling: A Multidisciplinary Journal, 21*, 329–341. doi: 10.1080/10705511.2014.915181

Brunton-Smith, I. (2011). *Modelling existing survey data: Full technical report of PIDOP* work package 5. Department of Sociology, University of Surrey.

Center of Inquiry in the Liberal Arts. (2013). *Wabash national study 2006-2012: Outcomes and experiences measures*.

Christiansen, N. D., Burns, G. N., & Montgomery, G. E. (2005). Reconsidering forced-choice item formats for applicant personality assessment. *Human Performance, 18*(3), 267–307. doi: 10.1207/s15327043hup1803_4

Geiser, C. (2013). *Data analysis with Mplus*. New York, NY: Guilford Press.

Eagan, M. K., Stolzenberg, E. B., Bates, A. K., Aragon, M. C., Suchard, M. R., & Rios-Aguilar, C. (2015). *The American freshman: National norms fall 2015*. Los Angeles: Higher Education Research Institute, UCLA.

Eagan, M. K., Stolzenberg, E. B., Zimmerman, H. B., Aragon, M. C., Whang Sayson, H., & Rios-Aguilar, C. (2017). *The American freshman: National norms fall 2016*. Los Angeles: Higher Education Research Institute, UCLA.

Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford Press.

Ehrlich, T. (2000). *Civic responsibility and higher education*. Phoenix, AZ: Oryx Press.

Finley, A. (2011). *Civic learning and democratic engagement: A review of the literature on civic engagement in post-secondary education*. Washington, DC: Association of American Colleges and Universities.

Henson, J. M., Reise, S. P., & Kim, K. H. (2007). Detecting mixtures from structural model differences using latent variable mixture modeling: A comparison of relative model fit statistics. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(2), 202–226. doi: 10.1080/10705510709336744

Lo, Y., Mendell, N. R., & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika*, 88(3), 767–778.

Lopez, M. H., Levine, P., Both, D., Kiesa, A., Kirby, E., & Marcelo, K. (2006). The 2006 civic and political health of a nation: A detailed look at how youth participate in politics and communities. *College Park, MD: Center for Information and Research on Civic Learning and Engagement*.

Lott, J. L., III, & Eagan, M. K., Jr. (2011). Assessing the psychometric properties of civic values. *Journal of Students Affairs Research and Practice, 48*(3), 333–327. doi: 10.2202/1949-6605.6288

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods, 7*(1), 19–40. doi: 10.1037/1082-989X.7.1.19

McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. New York, NY: Wiley.

Moely, B. E., Furco, A., & Reed, J. (2008). Charity and social change: The impact of individual preferences on service-learning outcomes. *Michigan Journal of Community Service Learning, 15*, 37–48

Moely, B. E., Mercer, S. H., Ilustre, V., Miron, D., & McFarland, M. (2002). Psychometric properties and correlates of the Civic Attitudes and Skills Questionnaire (CASQ): A measure of students' attitudes related to service-learning. Michigan *Journal of Community Service Learning, 8*(2), 15–26.

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

National Task Force on Civic Learning and Democratic Engagement (2012). *A crucible moment: College learning and democracy's future*. Washington, DC: Association of American College and Universities.

O'Neill, N. (2012). *Promising practices for personal and social responsibility: Findings from a national research collaborative*. Washington, DC: Association of American Colleges and Universities.

Pascarella, E. T., Ethington, C. A., & Smart, J. C. (1988). The influence of college on humanitarian/civic involvement values. *The Journal of Higher Education, 59*(4), 412–427. doi: 10.1080/00221546.1988.1178

Reason, R. D., & Hemer, K. M. (2015). *Civic learning and engagement: A review of the literature on civic learning, assessment and instruments*.

Rhee, B. S., & Dey, E. L. (1996, October). *Collegiate influences on the civic values of students*. Paper presented at the Annual Meeting of the Association of Study of Higher Education.

Rios-Aguilar, C. & Mars, M. M. (2011). Integration or fragmentation? College student citizenship in the global society. *Education, Knowledge, & Economy, 5* (1–2), 29–44.

Sequiera, S. N., Holzman, M. A., Horst, S. J., & Ghant, W. A. (2017). Developing college students' civic-mindedness through service-learning experiences: A mixed-methods study. *The Journal of Student Affairs Inquiry, 2*(1), 1–32.

State Council of Higher Educational in Virginia (2017). *Policy on student learning assessment and quality in undergraduate education*.

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika, 52*(3), 333–343. doi: 10.1007/BF0229436

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*(2), 461–464.

Spector, P. E. (2004). Social desirability bias. In M. S. Lewis-Beck, A. Bryman, & T. Futing Liao, (Eds), *The SAGE encyclopedia of social science research methods*. Thousand Oaks, CA: SAGE Publications Ltd. doi: 10.4135/9781412950589.n932

Tofighi, D., & Enders, C. K. (2008). Identifying the correct number of classes in growth mixture models. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 317–341). Greenwich, CT: Information Age Publishing, Inc.

Torney-Purta, J. V. (2009). International psychological research that matters for policy and practice, *American Psychologist, 64*(8), 825–837. doi: 10.1037/0003-066X.64.8.825

Torney-Purta, J., Cabrera, J. C., Roohr, K. C., Liu, O. L., & Rios, J. A. (2015). *Assessing civic competency and engagement in higher education: Research background, frameworks, and directions for next-generation assessment* (Research Report No. RR-15-34). Princeton, NJ: Educational Testing Service.

Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis, 18*(4), 450–469. doi: 10.1093/pan/mpq025

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, *44*, 92–107. doi: 10.1006/jmps.1999.1278

Weerts, D. J., Cabrera, A. F., & Meijas, P. P. (2014). Uncovering categories of civically engaged college students: A latent class analysis. *The Review of Higher Education, 37*(2), 141–168. doi: 10.1353/rhe.2014.0008

Westheimer, J., & Kahne, J. (2004). What kind of citizen? Politics of educating for citizenship. *American Educational Research Journal, 41*, 237–269. doi: 10.3102/00028312041002237

Abstract

The concept of effect size plays a crucial role in assessment, institutional research, and scholarly inquiry, where it is common with large sample sizes to find small relationships that are statistically significant. This study examines the distribution of effect sizes from institutions that participated in the National Survey of Student Engagement (NSSE) and empirically derives recommendations for their interpretation. The aim is to provide guidelines for researchers, policymakers, and assessment professionals to judge the importance of an effect from student engagement results. The authors argue for the adoption of the recommendations for interpreting effect sizes from statistical comparisons of NSSE data.

AUTHORS

Louis M. Rocconi, Ph.D.
University of Tennessee, Knoxville

Robert M. Gonyea, Ed.D.
Indiana University, Bloomington

Contextualizing Effect Sizes in the National Survey of Student Engagement: An Empirical Analysis

The concept of effect size plays a crucial role in higher education assessment. Assessment professionals tasked with gauging the success of campus policies and interventions often use effect sizes of their most important outcome measures (e.g., Springer, 2006). Many of these efforts rely on statistical comparisons where stakeholders not only want to know whether an intervention or policy has an effect, but also *how large the effect is*. Simply knowing that one score is statistically different from another is not particularly helpful. Especially in research that involves large data sets, it is common to find very small relationships or differences that are statistically significant at even the most stringent alpha levels (e.g., $\alpha = .001$). This could lead decision-makers to redistribute precious resources based on matters that are immaterial. On the other hand, decisions may be better informed if based on the relative magnitude of the effect. Thus, estimates of effect size provide researchers and practitioners essential information on the practical or theoretical importance of research findings. However, to better interpret the substantive value of an effect, effect sizes need to be grounded within a meaningful context.

The aim of this article is to examine the distribution of effect sizes from institutional comparisons reported by the National Survey of Student Engagement (NSSE) and make recommendations for their interpretation. We begin with an introduction to NSSE and its use in higher education assessment. Next, we provide a definition of effect size and a review of the limitations of hypothesis testing. We then discuss different types of effect sizes and the challenges involved in interpreting them in different contexts. Then, after considering Cohen's (1988) rationale for interpreting the size of an effect, we use the distribution of NSSE effect sizes from nearly a thousand participating institutions as a normative context to interpret the "natural" or relative variation in magnitudes of institution-to-peer-group comparisons. Ultimately, our aim is to provide helpful guidelines for assessment professionals, policymakers, and researchers to judge the importance of their student engagement results.

CORRESPONDENCE

Email
lrocconi@utk.edu

Background: The National Survey of Student Engagement

NSSE is an annual survey administered to first-year and senior students at bachelor's degree-granting colleges and universities across the United States and Canada. NSSE is used to assess the extent to which undergraduate students are exposed to and participate in a variety of effective educational practices (McCormick, Kinzie, & Gonyea, 2013). Decades of research on undergraduate students (see Astin, 1993; McCormick et al., 2013; Pace, 1979; Pascarella & Terenzini, 1991, 2005) show that students benefit from college when their efforts are directed at learning-centered activities both inside and outside of the classroom. In an effort to leverage these ideas to inform the assessment and improvement of undergraduate education, the National Survey of Student Engagement was launched in 2000. Standardized sampling and administration procedures ensure the comparability of results among participating institutions.

Thus, estimates of effect size provide researchers and practitioners essential information on the practical or theoretical importance of research findings. However, to better interpret the substantive value of an effect, effect sizes need to be grounded within a meaningful context.

Since its launch in 2000, NSSE has been used in institutional assessment as a valid source of evidence, whether by itself or linked with other school records (see McCormick et al., 2013 for a review). Colleges and universities participate in NSSE for a variety of reasons but mainly to assess the quality of their curricular and co-curricular undergraduate learning programs. As such, NSSE provides a suite of student engagement measures—including 10 Engagement Indicators, six High-Impact Practices, and items about the amount time spent preparing for classes, the quantity of reading and writing, perceived course challenge, and more. NSSE content can be mapped to department, institution, or accreditation goals and can be used to evaluate key performance indicators or to track progress on a strategic plan. NSSE also provides comparative data on these measures from other participating campuses (in aggregate). Such comparisons are valuable to know where to direct institutional improvement efforts. Effect sizes from these comparisons are used to identify dimensions of student learning where the institution is doing well, and areas where improvement is warranted (for a discussion of using effect sizes in NSSE reporting see Springer, 2006). The NSSE website (nsse.indiana.edu) and their *Lessons from the Field* series (NSSE, 2015, 2017) catalog hundreds of examples of how colleges and universities employ engagement data in this way. In many of these examples, effect sizes provide a way not only to identify meaningful differences between the institution and comparison group but also to track the magnitude of changes across multiple years of NSSE administrations on the same campus.

Definition of Effect Size

While Jacob Cohen (1988, 1992) is credited with popularizing the use of effect sizes, the idea of supplementing significance tests with an effect size statistic can be traced back to the early 1900s and the works of Karl Pearson and Ronald Fisher (Fisher, 1925; Pearson, 1900). Cohen (1988) defines an effect size as "the degree to which the phenomenon is present in the population" (p. 9). Effect sizes have also been described as the degree to which results differ from the null hypothesis (Grissom & Kim, 2005, 2012), the degree to which study results should be considered important regardless of sample size (Hojat & Xu, 2004), and the degree to which sample results diverge from expectations in the null hypothesis (Vacha-Haase & Thompson, 2004). Kelley and Preacher (2012) summarize these various conceptualizations of effect size and offer a more inclusive definition of effect sizes as a "quantitative reflection of the magnitude of some phenomenon that is used for the purpose of addressing a question of interest" (p.140).

Lakens (2013) describes effect sizes as among the most important outcomes to report in empirical studies. Effect sizes are important because they provide evidence of practical significance by representing the magnitude and direction of a relationship or difference, often in standardized metrics which can be understood regardless of the scale used (Kirk, 1996). Standardized effect sizes are particularly useful with abstract measurement indices, like those often found in survey research (e.g., NSSE's Engagement Indicators), because they convert raw differences to a standardized metric that can be compared across studies.

This is not to say that standardized effect sizes are always the most appropriate or useful expression of results. Indeed, when the underlying metric is meaningful in terms of its unit of measurement (enrollments, expenditures, hours, etc.), raw difference effect sizes can be more useful and easier to interpret than a standardized effect size (Lipsey et al., 2012). Too often, higher education research does not generate concrete measurement indices so we rely on standardized effect sizes, which are the focus of this article.

Criticisms of null-hypothesis significance testing

Criticisms of null-hypothesis significance testing (NHST) are not new (e.g., Cohen, 1994; Ferguson, 2009; Hill & Thompson, 2004; Kirk, 1996; Kline, 2013; Wasserstein & Lazar, 2016). Scholars have long regarded NHST as imperfect for examining data, yet the discussion on the meaning and use of statistical significance continues to this day. Recently, the American Statistical Association (ASA, Wasserstein & Lazar, 2016) published a set of guidelines regarding the use and misuse of *p*-values. These critiques of NHST can be summarized in three main criticisms. The first concerns a misunderstanding of *p*-values. In NHST, the *p*-value gives the mathematical likelihood or probability of obtaining these data or more extreme data (D) given that the null hypothesis (H_0) is true—that is, $P(D|H_0)$. However, researchers sometimes misinterpret the *p*-value from statistical tests to mean the probability the null hypothesis is true given that we have observed these data—that is, $P(H_0|D)$ (Cohen, 1994; Kirk, 1996; Kline, 2013; Wasserstein & Lazar, 2016). Unfortunately for researchers $P(D|H_0) \neq P(H_0|D)$; nor does obtaining data with a small $P(D|H_0)$ imply that $P(H_0|D)$ is also small (Cohen, 1994; Kirk, 1996). The main criticism here is that NHST does not tell us what we really want to know, whether or not the null hypothesis is true (Ferguson, 2009).

A second criticism is that NHST is very sensitive to sample size. Given a large enough sample, nearly any statistic can be found to be statistically significant. Because sample size is part of the calculation of the standard error, as the number of cases increases the standard error becomes smaller and the test statistic becomes larger, thus making it easier to find statistical significance. As Thompson (1998) quipped, "If we fail to reject, it is only because we've been too lazy to drag in enough participants" (p. 799). This feature is not necessarily a flaw of the hypothesis testing but rather is how the hypothesis test was designed to work.

NSSE also provides comparative data on these measures from other participating campuses (in aggregate). Such comparisons are valuable to know where to direct institutional improvement efforts.

This brings us to our third criticism of NHST—statistical significance does not equal practical significance. People often trumpet a small *p*-value (e.g., $p<.001$) as if it indicates a particularly large effect (Kirk, 1996; Lipsey et al., 2012; Wasserstein & Lazar, 2016). Statistical significance evaluates the probability of sample results but it does not tell us whether the effects are substantively important—an issue of greater interest to assessment professionals and policymakers. Statistical significance merely represents statistical rareness, but unlikely events can be completely meaningless or trivial, and conversely, likely events may be quite noteworthy. Unfortunately, *p*-values are confounded by the joint influences of sample results and sample size. Therefore, we use effect sizes to gauge the practical importance of results.

Types of effect sizes

Effect sizes are generally classified into three broad categories, generally understood as (a) measures of difference, (b) measures of strength of association, and (c) other measures (e.g., Fritz, Morris, & Richler, 2012; Kirk,1996; Rosnow & Rosenthal, 2003; Vacha-Haase & Thompson, 2004). Measures of difference are sometimes referred to as the *d*-type family of effect sizes, after Cohen's popular *d* statistic. These effect sizes measure the magnitude of the distance between group scores, and include raw differences (e.g., $Mean_1 - Mean_2$), standardized differences (e.g., Cohen's *d*, Hedges' *g*, Glass's *g*), and transformed differences (e.g., Cohen's *h*, Cohen's *q*, probit *d*). Measures of strength of association are also known as the *r*-type family of effect sizes after Pearson's *r*, the popular Pearson product-moment correlation coefficient. This family of measures is concerned with measures of correlation and variance explained and includes such statistics as Pearson's *r*, r^2, eta-squared (η^2), partial

eta square (η_p^2), and omega-squared (ω^2). The third category often serves as a catchall and includes other measures of effect such as risk estimates like the odds ratio, relative risk, or risk difference.

Results from student engagement comparisons are generally measures of difference, so we focus in this article on two *d*-type effect sizes, Cohen's *d* and Cohen's *h*. Cohen's *d* is used to describe the standardized mean difference between the scores of two groups of independent observations. It is calculated by dividing the mean difference by the pooled standard deviation. While it was Hedges (1982) who first proposed using the pooled sample standard deviation to standardize the mean difference, we will continue to refer to this effect size by its more common name of Cohen's *d* (Fritz et al., 2012). The formula to compute Cohen's *d* is as follows:

$$d = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{(n_1-1)\,s_1^2 + (n_2-1)\,s_2^2}{n_1 + n_2 - 2}}}$$

where ($\overline{X}_j$) is the sample mean for the j^{th} group, s_j^2 is the sample variance for the j^{th} group, and n_j is the sample size for the j^{th} group. The denominator is often referred to as the pooled estimate of standard deviation (s_{pooled}) and is the square root of the unbiased estimate of the within-group variance (Kelley & Preacher, 2012).

Standardized effect sizes are particularly useful with abstract measurement indices, like those often found in survey research (e.g., NSSE's Engagement Indicators), because they convert raw differences to a standardized metric that can be compared across studies. This is not to say that standardized effect sizes are always the most appropriate or useful expression of results.

Cohen's *h* effect size is the difference between two independent proportions (e.g., the percentage of students who participated in a particular activity such as study abroad or an internship) after each proportion has been transformed using an arcsine transformation. Specifically, it is calculated as follows:

$$h = (2\sin^{-1})\sqrt{P_1} - (2\sin^{-1}\sqrt{P_2})$$

where P_j is the sample proportion for the j^{th} group. The reason for employing the arcsine transformation is to make the proportions comparable in the sense of having variances independent of the parameter (Cohen, 1988; Hojat & Xu, 2004; Rosnow & Rosenthal, 2003). This type of transformation is known as a variance stabilizing transformation. Since the variance of a proportion is equal to the proportion multiplied by one minus the proportion divided by the sample size [$VAR\,(p) = \frac{(p)\,(1-p)}{n}$ where p represents the proportion and n represents the sample size], the variance of a proportion is dependent upon the value of the proportion. The fact that the variance of the proportion depends on its particular value prevents the simple difference between proportions to be used in power calculations because constant differences between two proportions cannot always be considered equal on the scale of proportions (Cohen, 1988). It is easier to detect differences between proportions that fall on the ends of the proportion scale than it is to detect differences between proportions that fall in the middle of the proportion scale. Thus, a transformation must be made to the proportions such that differences between the transformed parameters are equally detectable. Values for Cohen's *h* range from $-\pi$ to π, or around -3.14 to 3.14; this is because values of the arcsine function range between $-\pi/2$ and $\pi/2$.

Interpreting effect sizes

The purpose of reporting effect sizes is for a reader to better judge the importance of the findings. However, in order to understand the importance of results for abstract measurement indices such as the NSSE Engagement Indicators, the effect size must be contextualized against some frame of reference. The most popular frame of reference—a set of benchmarks offered by Cohen (1988, 1992) — is also common in educational research (see, McMillan & Foley, 2011; Peng, Chen, Chiang, & Chiang, 2013 for a review of effect size reporting in major journals). Cohen described *small* effects as those that are hardly visible, *medium* effects as

observable and noticeable to the eye of the beholder, and *large* effects as plainly evident or obvious. He then reluctantly suggested that *d* and *h* values of .2, .5, and .8, and *r* values of .1, .3, and .5, would represent small, medium, and large effects respectively. Yet, Cohen (1988) cautioned that "there is a certain risk inherent in offering conventional operational definitions for those terms for use in power analysis in as diverse a field of inquiry as behavioral science" (p. 25) and urged researchers to interpret effect sizes within the *context of the data*, even suggesting to researchers to "avoid the use of these conventions, if he can, in favor of exact values provided by theory or experience in the specific area in which he is working" (p. 184). Further complicating the interpretation of effect sizes, Cohen's own recommendations are not even consistent across different effect size types. For example, Cohen suggested that both $d = .5$ and $r = .3$ indicate a medium effect size. Yet, converting *r* to *d* using the formula provided by Cohen (1988, p. 23), $r = (d / \sqrt{d^2+4})$, we see that $d = .5$ is the equivalent of $r = .24$, which would be considered a small effect by *r* standards. Similarly, a large *d* effect of .8 corresponds to $r = .37$, just over the medium threshold for an *r* effect. Nevertheless, Cohen's recommendation has been incorporated into many educational, behavioral, and social science studies.

While discussing interpretations of effect sizes, Cohen (1988) cautioned that when a construct cannot be brought into the laboratory to be studied, which is the case in the vast majority of higher education assessments, extraneous or uncontrollable factors could lead to smaller or more difficult-to-detect effect sizes. In the realm of educational research, Cohen was right. For example, Hill, Bloom, Black, and Lipsey (2008) summarized estimates of achievement effect sizes from studies of K-12 educational interventions and noted that the standardized mean differences (Cohen's *d*) typically ranged from .20 to .30. Similarly, investigating K-12 students' academic performance on standardized reading and mathematics achievement tests, Lipsey et al. (2012) found standardized mean differences as large as .30 to be rare. When investigating school-level performance gaps, Bloom, Hill, Black, and Lipsey (2008) found standardized mean differences between "weak" (i.e., 10th percentile) and "average" (i.e., 50th percentile) schools to be in the .20 to .40 range.

Statistical significance evaluates the probability of sample results but it does not tell us whether the effects are substantively important—an issue of greater interest to assessment professionals and policymakers.

Researchers in other social and behavioral sciences have also noted that study effects were often small by Cohen's standards. Ellis (2010a) investigated the average effect size in international business research from 1995 to 2009 and found typically small effect sizes ($r < .10$) by Cohen's standards. Rosnow and Rosenthal (1989, 2003) note that small effect sizes are not that unusual in biomedical research. They illustrate how a seemingly trivial or very small effect can have important real-life consequences. For example, in a study to examine the effects of aspirin on incidence of heart attacks, an effect size of $r = 0.034$ was used to end the study prematurely because it had become clear that aspirin prevents heart attacks and it would have been unethical to continue to give half the participants a placebo. Rosnow and Rosenthal (1989, 2003) argue that this is not to suggest that all small effects are noteworthy; rather, that small effects can have practical consequences in life and death situations. They conclude that in research involving hard-to-change outcomes, such as the incidence of heart attacks, small effects can have profound practical significance.

Few of the effects mentioned above would be described as anything other than small by Cohen's (1988, 1992) standards. What can be taken from these examples is that the interpretation of effect sizes is context dependent. In fact, many scholars (e.g., Cohen, 1988; Hill & Thompson, 2004; Kelley & Preacher, 2012; Kirk, 1996; Thompson, 2001; Vacha-Haase & Thompson, 2004) criticize the use of universally accepted guidelines, like Cohen's benchmarks, for interpreting effect sizes. As Thompson (2001) points out, "if people interpreted effect sizes with the same rigidity that $\alpha = .05$ has been used in statistical testing, we would merely be being stupid in another metric" (p. 82-83).

The American Psychological Association's (APA) publication manual is clear about the importance of reporting effect sizes: "For the reader to appreciate the magnitude or importance of a study's findings, it is almost always necessary to include some measure of effect size" (APA, 2010, p. 34). Additionally, the APA Task Force emphasized that reporting and interpreting effect sizes with consideration to effects from previous studies are

essential to good research (Wilkinson & APA Task Force on Statistical Inference, 1999). Similarly, the American Educational Research Association (AERA, 2006) recommended in its standards for reporting research that statistical results be accompanied by an effect size and a "qualitative interpretation" of the effect. These recommendations have been endorsed by journal editors in higher education (e.g., Smart, 2005) and other behavioral and social science disciplines (e.g., Lopéz, Valenzuela, Nussbaum, & Tsai, 2015; Vacha-Haase & Thompson, 2004) who have also called on distinguishing between the statistical and practical significance of a study's findings. Unfortunately, most research in education utilizes Cohen's recommendations of small, medium, and large effects rather than interpreting the effect size within the context of previous findings or research (McMillan & Foley, 2011; Peng et al., 2013). Given the importance of contextualizing an effect within a specific research area, assessment professionals, researchers, and policymakers assessing student engagement need the ability to interpret effect sizes of their results within the context of other student engagement results.

The purpose of reporting effect sizes is for a reader to better judge the importance of the findings. However, in order to understand the importance of results for abstract measurement indices such as the NSSE Engagement Indicators, the effect size must be contextualized against some frame of reference.

Purpose and Research Questions

The purpose of this study is to examine the distribution of effect sizes derived from institutional comparisons from the National Survey of Student Engagement (NSSE) and to make recommendations for their interpretation. The following research questions guided our study:

1. How do the effect sizes from NSSE institutional comparisons distribute within Cohen's small, medium, and large ranges?
2. Is it possible to derive more useful effect size interpretations that fit the context of institutional engagement results?

Method

Data Source

The NSSE data used in this study were obtained and used with permission from The Indiana University Center for Postsecondary Research. As mentioned previously, NSSE is an annual survey administered to first-year and senior students at baccalaureate degree-granting colleges and universities and is used to assess the extent to which students are exposed to and participate in effective educational practices (McCormick et al., 2013). The analytic sample consisted of 984 U.S. institutions that participated in the 2013 or 2014 administration of NSSE. For institutions that participated both years, we only included the 2014 data. Participating institutions represented a broad cross-section of the national profile of U.S. bachelor's degree-granting institutions (Table 1).

Measures

Effect sizes for the study were based on comparisons of two primary sets of variables generated from the NSSE questionnaire: Engagement Indicators (EIs) and High-Impact Practices (HIPs). NSSE's 10 EIs represent the multi-dimensional nature of student engagement, organized within four engagement themes. They include four measures of academic challenge: *Higher-Order Learning, Reflective & Integrative Learning, Learning Strategies*, and *Quantitative Reasoning*; two measures about learning with peers: *Collaborative Learning* and *Discussions with Diverse Others*; two measures describing experiences with faculty: *Student-Faculty Interaction* and *Effective Teaching Practices*; and two measures of the campus environment: *Quality of Interactions* and *Supportive Environment*. Each EI is a reliable scale that measures a distinct aspect of student engagement by summarizing students' responses to a set of related survey questions. The psychometric properties of these measures have been described in detail elsewhere (BrckaLorenz & Gonyea, 2014; Miller, Sarraf, Dumford, & Rocconi, 2016).

Table 1
Characteristics of Participating Institutions (N=984)

		%
Carnegie Classification	Research Universities (very high research activity)	5
	Research Universities (high research activity)	7
	Doctoral/Research Universities	6
	Master's Colleges and Universities (larger programs)	27
	Master's Colleges and Universities (medium programs)	11
	Master's Colleges and Universities (smaller programs)	6
	Baccalaureate Colleges—Arts & Sciences	16
	Baccalaureate Colleges—Diverse Fields	17
	Other types	6
Control	Public	40
	Private	60
Barron's Selectivity	Noncompetitive	4
	Less Competitive	10
	Competitive	46
	Very Competitive	19
	Highly Competitive	8
	Most Competitive	3
	Not available/Special	10

Given the importance of contextualizing an effect within a specific research area, assessment professionals, researchers, and policymakers assessing student engagement need the ability to interpret effect sizes of their results within the context of other student engagement results.

HIPs encompass several co-curricular educational experiences that have been recognized as "high-impact" due to their positive associations with student learning and development in college (Kuh, 2008; Kuh & O'Donnell, 2013). NSSE asks students if they have participated in six HIPs: *learning community, service-learning, research with a faculty member, internship or field experience, study abroad*, and *culminating senior experience*. We excluded comparisons for internships, study abroad, and culminating senior experiences for first-year students because these opportunities are typically not available until later in the undergraduate years.

Analysis

To answer the first research question, we generated a dataset by calculating effect sizes for each EI and HIP, separately for first-year and senior students, for comparisons of respondents attending each of the 984 institutions with respondents from all other institutions as a single group. Although institutional users of NSSE are allowed to customize comparison groups, we compared results to students enrolled at all other institutions in order to have a common comparison group for analytic consistency. Results were weighted by sex, enrollment status, and institution size (consistent with NSSE reports delivered to institutions).

To answer the second research question, we considered Cohen's (1988) rationale for observing a small effect (i.e., an effect that is hardly noticeable), a medium effect (i.e., an effect that is observable), and a large effect (i.e., an effect that is plainly evident) and considered ways in which such institutional differences would be observable in the data. To accomplish this, we derived a technique to model comparisons that would resemble effect sizes of increasing magnitude (illustrated in Figure 1). We conceptualized that a *small* effect would resemble the difference between the scores of students attending institutions in the third quartile (i.e., between the 50th and 75th percentiles) and those attending institutions in the second quartile (i.e., between the 25th and 50th percentile). These two sets of institutions are labeled groups A and B in Figure 1a. Because groups A and B are fairly close within the distribution, the difference between the average scores of the students attending those institutions is expected to be small. In a similar way, a *medium* effect would resemble the difference between the average scores of students attending institutions in the upper and lower halves of the distribution (Figure 1b), and a *large* effect would resemble the difference between the average scores of students attending institutions in the top and bottom quartiles

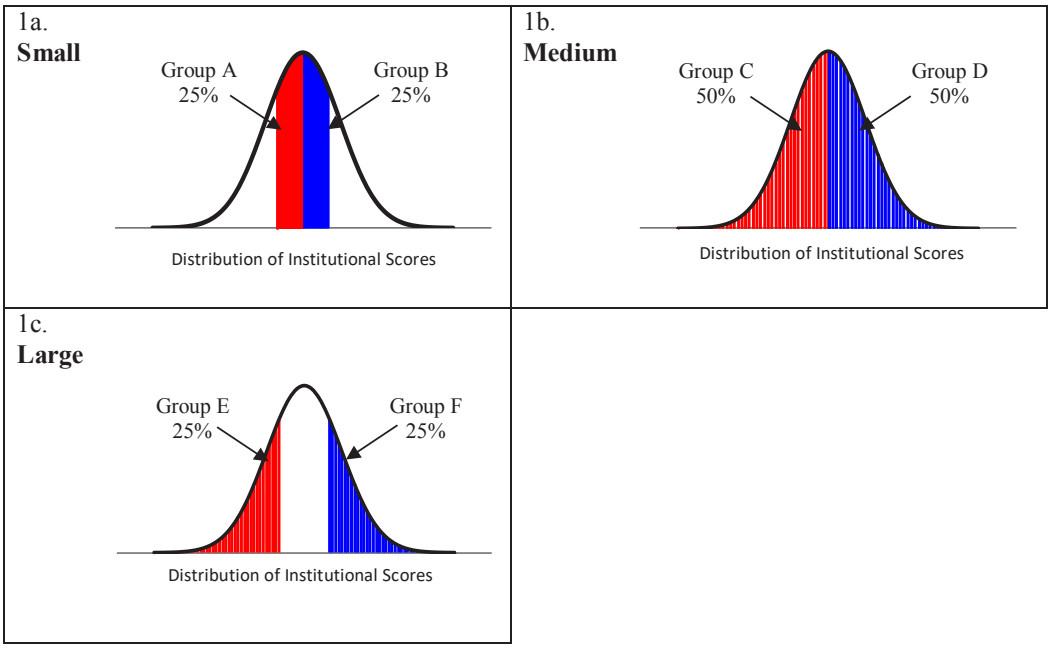


Figure 1
Illustration of Three Model Comparison Groups for Determining Empirically-Based Effect Size Thresholds Based on the Distribution of Student Engagement Measures

(Figure 1c). Our analytic approach is similar to a technique used by Bloom et al. (2008) and Konstantopoulos and Hedges (2008) to contextualize effect size estimates for K-12 school achievement in which they estimated differences in achievement for students at "average" (i.e., 50th percentile) and "weak" schools (10th percentile).

The first step in this process was assigning percentile rankings to each of the 984 institution's EI and HIP scores, separately for first-year and senior students. The percentile rankings were based on an institution's precision-weighted score. The precision-weighting process involved adjusting institutional mean scores using Empirical Bayes methods in order to account for lower reliability in institutional means due to small sample sizes and distance from the overall estimate (Hox, 2010). The objective of the precision-weighting adjustment was to avoid over-interpretation of statistical noise in ranking institutions. The precision-weighted means were only used to derive the percentile rankings; unadjusted student-level data were used in the effect size calculations. Once percentile rankings were obtained for institutions' EI and HIP scores, we used these percentile rankings to model effect size comparisons of increasing magnitude (Figure 1). Cohen's *d* and *h* effect sizes were computed according to the formulas presented earlier. For example, to calculate the "small" effect in our proposed scheme, students attending institutions that had percentile ranks between the 50th and 75th percentiles were compared with students attending institutions that had percentile ranks between the 25th and 50th percentiles (Figure 1a). Finally, we calculated confidence intervals for the effect sizes by bootstrapping 1,000 samples for each comparison that was used in each effect size calculation (Kelley & Preacher, 2012).

These results suggest that new criteria for the interpretation of Cohen's d effect sizes for EIs within the context of NSSE results are necessary.

Table 2
Frequency of NSSE Effect Sizes[a] by Cohen's Suggested Ranges[b]

	Trivial		Small		Medium		Large	
	ES < |.2|		|.2| ≤ ES < |.5|		|.5| ≤ ES < |.8|		ES ≥ |.8|	
Engagement Indicator	First-year	Senior	First-year	Senior	First-year	Senior	First-year	Senior
Higher-Order Learning	72%	75%	26%	23%	1%	1%	<1%	<1%
Reflective & Integrative Learning	71%	68%	26%	28%	2%	3%	<1%	1%
Learning Strategies	75%	66%	22%	33%	2%	1%	<1%	<1%
Quantitative Reasoning	76%	79%	20%	18%	2%	2%	1%	<1%
Collaborative Learning	64%	58%	30%	35%	4%	5%	2%	2%
Discussions with Diverse Others	61%	63%	34%	33%	4%	3%	<1%	1%
Student-Faculty Interaction	60%	41%	33%	39%	6%	16%	1%	4%
Effective Teaching Practices	68%	71%	30%	27%	1%	2%	<1%	<1%
Quality of Interactions	59%	59%	37%	37%	2%	4%	<1%	0%
Supportive Environment	61%	55%	34%	38%	4%	6%	<1%	<1%
High-Impact Practice								
Learning Community	57%	69%	38%	26%	3%	3%	1%	1%
Service-Learning	47%	46%	36%	36%	11%	13%	6%	5%
Research with Faculty	84%	55%	15%	32%	1%	11%	0%	2%
Internship[c]	--	43%	--	38%	--	15%	--	4%
Study Abroad[c]	--	40%	--	43%	--	10%	--	7%
Culminating Senior Experience[c]	--	36%	--	36%	--	17%	--	10%

[a]Effect sizes were derived from each institution's comparison with the other 983 institutions in the data, separately by class level for each EI and HIP.

[b]Cohen's suggestions of small (d & h = .2), medium (d & h = .5), and large (d & h = .8).

[c]Effect sizes for Internship, Study Abroad, and Culminating Senior Experience are not calculated for first-year students since these opportunities are typically not available until later in the undergraduate years.

Results

Research Question 1: How do the effect sizes from NSSE institutional comparisons distribute within Cohen's small, medium, and large ranges?

Table 2 shows the percentage of institutions that had effect sizes within each of Cohen's ranges on the EIs and HIPs for first-year and senior students. For most EIs, over 60% of the effect sizes were *trivial* (ES < |.2| in magnitude) and 20% to 30% were *small* (|.2| ≤ ES < |.5|). Only around 1% to 6% of comparisons were within the *medium* range and typically less than 2% met Cohen's criteria of a *large* effect. An exception was Student-Faculty Interaction for seniors, where fewer effect sizes were classified as trivial (41%), and more were classified as medium (16%) and large (4%).

HIP comparisons showed somewhat different patterns. While the largest number of HIP effect sizes were trivial in magnitude, they ranged widely between 36% and 84%. Compared to the EIs, more HIP effect sizes were in the medium and large range, particularly among seniors. For example, for service-learning, 17% of first-year effect sizes and 18% of senior effect sizes were at least medium in magnitude. Similar totals were tallied for senior

Table 3
Effect Sizes from NSSE EI Percentile Group Comparisons (95% confidence intervals given in parentheses)

	First-year			Senior		
	Small	Medium	Large	Small	Medium	Large
Higher-Order Learning	.087 (.074, .098)	.223 (.214, .232)	.372 (.359, .385)	.096 (.085, .106)	.246 (.239, .253)	.356 (.346, .365)
Reflective & Integrative Learning	.109 (.098, .121)	.260 (.251, .268)	.394 (.381, .407)	.103 (.094, .113)	.266 (.260, .272)	.414 (.404, .424)
Learning Strategies	.088 (.076, .099)	.227 (.218, .235)	.355 (.342, .368)	.078 (.068, .087)	.203 (.196, .209)	.312 (.302, .322)
Quantitative Reasoning	.092 (.079, .105)	.237 (.229, .246)	.354 (.341, .366)	.113 (.104, .123)	.304 (.298, .312)	.466 (.456, .476)
Collaborative Learning	.129 (.117, .141)	.363 (.354, .371)	.549 (.537, .561)	.125 (.116, .134)	.381 (.375, .388)	.594 (.584, .604)
Discussions with Diverse Others	.133 (.121, .146)	.330 (.321, .339)	.501 (.488, .515)	.120 (.110, .130)	.321 (.314, .329)	.510 (.500, .520)
Student-Faculty Interaction	.121 (.110, .133)	.335 (.326, .344)	.545 (.530, .560)	.194 (.183, .205)	.491 (.483, .498)	.744 (.732, .756)
Effective Teaching Practices	.100 (.087, .112)	.276 (.266, .285)	.414 (.401, .428)	.086 (.076, .096)	.245 (.238, .252)	.373 (.363, .383)
Quality of Interactions	.139 (.127, .152)	.317 (.308, .326)	.461 (.449, .472)	.135 (.124, .146)	.360 (.353, .367)	.515 (.505, .525)
Supportive Environment	.116 (.104, .130)	.310 (.301, .319)	.488 (.475, .501)	.136 (.125, .146)	.344 (.336, .351)	.529 (.519, .540)
Minimum *d*	.087	.223	.354	.078	.203	.312
Maximum *d*	.139	.363	.549	.194	.491	.744
Average *d*	.111	.288	.443	.118	.316	.481

Table 4
Frequency of NSSE EI Effect Sizes by Suggested Ranges[a]

	Effect Size Range							
	Trivial		Small		Medium		Large	
	ES < |.1|		|.1| ≤ ES < |.3|		|.3| ≤ ES < |.5|		ES ≥ |.5|	
Engagement Indicator	First-year	Senior	First-year	Senior	First-year	Senior	First-year	Senior
Higher-Order Learning	45%	46%	44%	45%	9%	8%	1%	1%
Reflective & Integrative Learning	40%	40%	47%	44%	11%	12%	2%	4%
Learning Strategies	44%	38%	46%	46%	8%	15%	2%	1%
Quantitative Reasoning	47%	49%	42%	41%	8%	7%	3%	3%
Collaborative Learning	34%	30%	46%	48%	14%	14%	5%	7%
Discussions with Diverse Others	33%	35%	47%	47%	15%	14%	4%	4%
Student-Faculty Interaction	33%	23%	43%	34%	17%	23%	6%	20%
Effective Teaching Practices	38%	41%	48%	46%	12%	11%	1%	2%
Quality of Interactions	34%	30%	46%	48%	16%	18%	3%	4%
Supportive Environment	36%	30%	45%	46%	15%	18%	4%	6%

internships and study abroad, and fully 27% of effect sizes for culminating senior experiences were at least medium in magnitude. In contrast, over four-fifths of the institutional comparisons for first-year research with faculty were trivial, and 1% were at least medium in magnitude.

Research Question 2: Is it possible to derive more useful effect size interpretations that fit the context of institutional engagement results?

Our study aims to provide assessment professionals, policymakers, researchers and other users of NSSE data a framework to aid in assessing the practical significance of NSSE student engagement results.

Given the fact that a large majority of effect sizes were small or trivial according to Cohen's cut points, we analyzed effect sizes according to our proposed scheme based on the distribution of institutional scores. Table 3 shows the Cohen's *d* effect sizes and confidence intervals for the small, medium, and large model comparisons for first-year and senior students on all 10 EIs. While the effect size estimates in Table 3 varied somewhat between EIs and between student class levels, the ranges within the small, medium, and large categories were fairly consistent and, with the exception of a few instances, did not overlap. That is, the maximum small effect size was almost always lower than the minimum medium effect size, and the maximum medium effect size was usually lower than the minimum large effect size. For both first-year students and seniors, the average small effect size was about .1 and the average medium effect size was about .3. The average large effect size for first-year students was about .44 and for seniors was about .48. Compared to Cohen's recommendations, these effect size estimates tended to be lower in nearly every instance.

These results suggest that new criteria for the interpretation of Cohen's *d* effect sizes for EIs within the context of NSSE results are necessary. The consistency of effect size values among the EIs points toward a new set of criteria for their interpretation: small effects start at about .1, medium effects start at about .3, and large effects start at about .5. These new reference values were selected after an examination of the effect size values in Table 3, which when rounded to the nearest tenth approximated evenly-spaced intervals between .1 and .5. Table 4 reports the distribution of effect sizes based on the proposed reference values for the Engagement Indicators. As expected from our previous analysis of effect size distribution, the majority of effect sizes were trivial or small. Yet, there is a finer distribution within categories from what we saw in Table 2 based on Cohen's definitions. For the EIs, Table 4 shows that approximately 35% to 40% of all effect sizes were in the trivial range, 40% to 45% were considered small, 10% to 15% were medium, and large effect sizes were relatively rare.

Table 5 shows the Cohen's *h* effect sizes and confidence intervals for the small, medium, and large model comparisons on the six HIPs. Cohen's *h* effect sizes varied more across HIPs and across class year than did the effect size estimates for the EIs. While the effect size estimates for learning communities were generally similar to those of the EIs (.1, .3, and .5), the effect sizes for service-learning, internships, study abroad, and culminating senior experiences were considerably larger and in fact approximated Cohen's standards of .2, .5, and .8. Of the three HIPs measured for first-year students, service-learning had the widest range, with small, medium, and large estimates of .18, .43, and .73. On the other hand, research with faculty estimates for first-year students were smaller and in a fairly narrow range, with estimates of .06, .17, and .26, respectively. Effect size estimates for research with faculty also varied greatly between class level while estimates for learning community and service-learning were fairly consistent across class level. Average effect sizes for the three first-year HIPs were .11, .31, and .50 for small, medium, and large effects, respectively. Senior estimates for HIP effect sizes were generally larger in magnitude and ranged more. For instance, effect sizes for culminating senior experiences had the largest range, with small, medium, and large effects of .25, .60, and .92, respectively, while learning community effect sizes for seniors had the smallest range, .10, .27, and .43. With the exception of learning community (which typically had lower estimates) and culminating senior experiences (which typically had larger estimates), the other four HIPs for seniors had relatively similar effect size estimates: about .2 for small, between .4 and .5 for medium, and between .6 and .8 for large. Given the variability in Cohen's h effect size estimates both between HIPs and between class levels, it is difficult to provide a set of benchmarks for effect sizes applicable to HIPs in general.

Table 5
Effect Sizes from NSSE High-Impact Practices Percentile Group Comparisons (95% confidence intervals given in parentheses)

	First-year			Senior		
	Small	Medium	Large	Small	Medium	Large
Learning Community	.105 (.093, .118)	.345 (.337, .354)	.513 (.501, .525)	.096 (.086, .107)	.286 (.279, .293)	.434 (.424, .445)
Service-Learning	.179 (.166, .192)	.427 (.419, .437)	.728 (.714, .741)	.171 (.161, .182)	.434 (.427, .441)	.690 (.677, .702)
Research with Faculty	.058 (.045, .070)	.166 (.158, .175)	.255 (.242, .267)	.156 (.146, .165)	.407 (.400, .415)	.606 (.595, .616)
Internship[a]	--	--	--	.199 (.190, .208)	.501 (.494, .508)	.757 (.746, .768)
Study Abroad[a]	--	--	--	.199 (.189, .208)	.499 (.492, .506)	.784 (.775, .793)
Culminating Senior Experience[a]	--	--	--	.246 (.236, .257)	.604 (.596, .612)	.920 (.909, .931)
Minimum *h*	.058	.166	.255	.096	.286	.434
Maximum *h*	.179	.427	.728	.246	.604	.920
Average *h*	.114	.313	.498	.178	.455	.698

[a]Effect sizes for Internship, Study Abroad, and Culminating Senior Experience are not calculated for first-year students since these opportunities are typically not available until later in the undergraduate years.

Limitations

As with any research, ours is not without its limitations. First, our findings primarily apply to the NSSE Engagement Indicators and High-Impact Practice items. With the exception of the six HIP items, our analysis did not include all the individual items on the NSSE questionnaire. Thus, we urge readers to use caution when applying these recommendations to the individual item estimates in NSSE. Second, Cohen (1988) and others (e.g., Ellis, 2010b; Lakens, 2013; Vacha-Haase & Thompson, 2004) advocate for grounding effects in an area of research; as such we urge caution in applying the study's findings and recommendations on effect sizes to other surveys of undergraduates. Although NSSE is a widely adopted instrument used to assess the student experience, it is only one means by which to measure student engagement, and researchers are encouraged to adopt the study's methods to examine effect sizes in other contexts. Finally, the generalizability of the findings is also limited by the fact that institutions self-selected to participate in NSSE. Although our sample consisted of a wide cross-section of baccalaureate degree-granting institutions (Table 1), it was not necessarily representative of all four-year colleges and universities in the United States. Despite these limitations, we believe this study provides valuable insight to the types of effects that are possible for student engagement results with NSSE data and may guide these professionals in their interpretation of student engagement results.

Despite these limitations, we believe this study provides valuable insight to the types of effects that are possible for student engagement results with NSSE data and may guide these professionals in their interpretation of student engagement results.

Discussion

Knowing whether an institution scored statistically higher than its comparison group on a particular Engagement Indicator (EI) is not particularly helpful to an assessment professional or administrator. At the same time, raw score differences for abstract indices, like NSSE's Engagement Indicators, are difficult to interpret because they lack a meaningful measurement unit. Therefore, in order to communicate the importance of engagement survey results to assessment professionals, policymakers, and other users of NSSE, statistical comparisons need to be translated into a form that facilitates more practical interpretations. While professional organizations (e.g., AERA, APA, ASA) and journal editors (e.g., Smart, 2005; Lopéz et al. 2015) call for researchers to report effect sizes in their studies, researchers infrequently interpret what they mean or compare them to previous effects (Lakens, 2013; McMillan & Foley, 2011;

Peng et al., 2013). Absent a meaningful context grounded in data that are common to the field or area of research, an effect size by itself provides very little other than transforming the difference into standardized units. Interpreting the magnitude or practical significance of an effect size requires it to be compared with other appropriate effects that are relevant to the research study (Kelley & Preacher, 2012; Lipsey et al., 2012; Vacha-Haase & Thompson, 2004). Our study aims to provide assessment professionals, policymakers, researchers and other users of NSSE data a framework to aid in assessing the practical significance of NSSE student engagement results.

These effect size recommendations are not intended to be definitive judgments on the relative efficacy of NSSE's Engagement Indicators.

Our findings reinforce Cohen's (1988) caution against the use of universal benchmarks for interpreting effect sizes. Results from our study indicated that Cohen's benchmarks did not adequately fit effect sizes seen in NSSE, especially for the EIs. When examining the distribution of effect sizes within Cohen's benchmarks (Table 2), nearly all effects achieved would be considered trivial or small. Rarely did effect size estimates meet Cohen's thresholds for medium and large, particularly for the EIs. Using our contrived comparisons to mimic effect sizes of increasing magnitude, we found that the EIs could be better summarized using a .1, .3, .5 convention for small, medium, and large effects, respectively. Like Cohen's benchmarks, these new values should not be interpreted as precise cut points but rather are to be viewed as a coarse set of thresholds or minimum values by which one might consider the magnitude of an effect.

The proposed values for EIs may have intuitive and functional appeal for assessment professionals and other users of NSSE data. They are grounded in actual NSSE data, which allows for richer interpretations of the results. Institutions with meaningful differences will more likely find effect sizes of .3 or .5 and can be more confident in interpreting those effects as medium or large effects. Furthermore, although relatively small, one should not simply disregard effect sizes of .1 as trivial. In their review of psychological, educational, and behavioral treatment interventions, Lipsey and Wilson (1993) reached similar conclusions regarding findings with small effect sizes stating, "we cannot arbitrarily dismiss modest values (even 0.10 or 0.20 SDs) as obviously trivial" (p. 1199). Similarly, in their study of school reform, Konstantopoulos and Hedges (2008) remark that an effect of half a standard deviation (i.e., d = .5) should be interpreted as a very large effect in the context of school reform.

A goal of this article is to provide assessment professionals, policy makers, and researchers guidelines for interpreting NSSE student engagement effects sizes. Assessment professionals, in particular, can utilize these results by using effect sizes for guidance on which items to report to stakeholders. They can use our contextualized results and recommendations to identify areas of engagement where an institution is doing comparatively well, and to identify areas in need of improvement. For example, finding a negative, medium in magnitude effect size (such as -.30) in comparison to a group of peer institutions on the Student-Faculty Interaction indicator, an institution might set a goal to improve the quality and frequency of contact between students and faculty. Our findings can aid users in answering what is a meaningful difference, and what effect sizes are typical in this area?

These effect size recommendations are not intended to be definitive judgments on the relative efficacy of NSSE's Engagement Indicators. As Hill et al. (2008, p.176) states, "empirical benchmarks from a research synthesis do not indicate what effects are desirable from a policy standpoint;" instead, they serve to indicate what effects are likely and attainable. Our recommended benchmarks are a general gauge but can provide some guidance as to what magnitude effects are typical with student engagement results and NSSE data in particular.

Our effect size comparisons are most appropriate to serve as a reference for making institution-to-peer comparisons for the EI and HIP items on NSSE. While our analyses focused on comparisons among institutions, intra-institutional comparisons (e.g., comparisons across years, major fields of study, co-curricular involvement) are also often important and interesting to assessment professionals. Although our analyses did not focus on intra-institutional comparisons, our findings may be useful as a starting point when investigating these relationships since our results are grounded in NSSE data. However, we caution readers when making these comparisons that knowledge of the subject matter, and not blind reference to our findings, is warranted. For instance, an assessment professional interested in how often

students use quantitative reasoning skills across academic majors should keep in mind that certain majors emphasize these skills more than others (Rocconi, Lambert, McCormick, & Sarraf, 2013), and as such, should expect larger effect size differences among certain academic majors (e.g., humanities compared with physical sciences). Future research in this area needs to consider these intra-institutional comparisons.

For researchers or users interested in a specific EI, referring to the results in Table 3 would offer more accurate or meaningful information on the estimate of effect size for a particular indicator. Our recommended benchmarks fit better for some EIs than others. For instance, the Discussions with Diverse Others, Quality of Interactions, and Supportive Environment indicators closely follow the new recommended pattern of .1 for small, .3 for medium, and .5 for large. However, some indicators had effects slightly smaller than the recommended cut-off points. For instance, the largest effects for Higher-Order Learning, Reflective and Integrative Learning, and Learning Strategies were between .31 and .41. On the other hand, Student-Faculty Interaction and Collaborative Learning had slightly higher effect size estimates than the recommended benchmark values. Student-Faculty Interaction for seniors particularly stands out as an exception to our general guidelines with estimated effects closer to Cohen's recommendations of small, medium, and large: .2, .5, .8, respectively.

Our recommended benchmarks are a general gauge but can provide some guidance as to what magnitude effects are typical with student engagement results and NSSE data in particular.

We were unable to recommend a new set of benchmarks for interpreting the results from HIP comparisons. The effect size estimates among the HIPs and between class years varied so greatly that it was difficult to reduce them into a general recommendation for all HIPs. We encourage researchers and users of NSSE data to examine the effect size estimates in Table 5 to gauge the size or practical importance for a particular high-impact practice.

The effect size estimates we found were consistent with the claims of prior researchers in education and the social and behavioral sciences who found effect sizes rarely as large as Cohen's suggestions and often variable from one context to another (e.g., Bloom et al., 2008; Ellis, 2010a; Hill et al., 2008; Lipsey et al., 2012; Rosnow & Rosenthal, 1989, 2003). One reason the effect size estimates for the EIs were generally smaller in magnitude, compared with most of the HIPs, is because they are more abstract concepts, as opposed to the HIPs which are more concrete educational outcomes. Cohen (1988) cautioned that with more abstract and difficult to measure phenomena, the statistical noise brought on by uncontrollable factors and measurement error can lead to smaller effect sizes. Compared with the EIs, institutions have more direct control over HIPs. Program faculty or other institutional leaders can implement policies that require seniors to complete a culminating thesis or that implement a college-wide initiative with a service-learning component. In addition, HIPs are measured using a single item on the survey while the EIs are a collection of individual items used to create a scale measuring the desired construct.

The ability to contextualize effect sizes found in NSSE will aid assessment professionals and policymakers in judging the relative importance of student engagement results within the context of the survey and better enable these professionals to make more informed decisions on the relative size and practical value of student engagement results.

As Ellis (2010b) argues, effect sizes are "meaningless unless they can be contextualized against some frame of reference" (p. 32). Unfortunately, contextualizing the meaning of an effect grounded within the specific research context is not that common in the educational research literature (see McMillan & Foley, 2011; Peng et al., 2013). Our study provides researchers and users of NSSE the ability to contextualize the effects found in their studies against a frame of reference grounded in actual NSSE data. Contextualizing the interpretations of effect sizes not only helps facilitate the interpretation of results but can also aid researchers in building on previous findings. Our study provided new guidelines for considering the size of effects with NSSE's EI and HIP data. We believe the empirical results we have presented provide better guidance to a user of NSSE data than the conventional guidelines provided by Cohen. The ability to contextualize effect sizes found in NSSE will aid assessment professionals and policymakers in judging the relative importance of student engagement results within the context of the survey and better enable these professionals to make more informed decisions on the relative size and practical value of student engagement results.

References

American Educational Research Association. (2006). Standards for reporting on empirical social science research in AERA publications. *Educational Researcher, 35*(6), 33–40.

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, D. C.

Astin, A. W. (1993). *What matters in college? Four critical years revisited*. San Francisco: Jossey-Bass.

Bloom, H. S., Hill, C. J., Black, A. B., & Lipsey, M. W. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *Journal of Research on Educational Effectiveness, 1*(4), 289¬328.

BrckaLorenz, A., & Gonyea, R. M. (2014). *The NSSE update: Analysis and design of ten new engagement indicators*. Bloomington, IN: Indiana University Center for Postsecondary Research.

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

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159.

Cohen, J. (1994). The earth is round (p<.05). *American Psychologist, 49*(12), 997–1003.

Ellis, P. D. (2010a). Effect sizes and the interpretation of research results in international business. *Journal of International Business Studies, 47*(9), 1581–1588.

Ellis, P. D. (2010b). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. New York: Cambridge University Press.

Fisher, R. A. (1925). *Statistical methods for research workers*. London: Oliver & Boyd.

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice, 40*(5), 532–538.

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology, 141*(1), 2–18.

Grissom, R. J., & Kim, J. J. (2005). *Effect sizes for research: A broad practical approach*. Mahwah, NJ: Lawrence Erlbaum Associates.

Grissom, R. J., & Kim, J. J. (2012). *Effect sizes for research: Univariate and Multivariate Applications* (2nd ed.). New York: Routledge.

Hedges, L. V. (1982). Estimation of effect size from a series of independent observations. *Psychological Bulletin, 92*(2), 490-499.

Hill, C. J., Bloom, H. S., Black, A. B., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives, 2*(3), 172–177.

Hill, C. R., & Thompson, B. (2004). Computing and interpreting effect sizes. In J. C. Smart (Ed.). *Higher Education: Handbook of Theory and Research* (Vol. 19, pp. 175-195). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Hojat, M., & Xu, G. (2004). A visitor's guide to effect size. Advances in Health Sciences Education, 9(3), 241-249.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). New York: Routledge.

Huberty, C. J. (2002). A history of effect size indices. *Educational and Psychological Measurement, 6*(2), 227–24.

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods, 17*(2), 137¬153.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56*(5), 746–759.

Kline, R. B. (2013). *Beyond significance testing: Statistics reform in the behavioral sciences* (2nd ed.). Washington, DC: American Psychological Association.

Konstantopoulos, S., & Hedges, L. V. (2008). How large an effect can we expect from school reform? *Teachers College Record, 110*(8), 1613–1640.

Kuh, G. D. (2008). *High-impact educational practices: What they are, who has access to them, and why they matter*. Association of American Colleges and Universities.

Kuh, G. D., & O'Donnell, K. (2013). *Ensuring quality & taking high-impact practices to scale*. Washington, DC: Association of American Colleges and Universities.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-test and ANOVAs. *Frontiers in Psychology, 4*, 1–12.

Lipsey, M., Puzio, K., Yun, C., Hebert, M., Steinka-Fry, K., Cole, M., Roberts, M., Anthony, K., & Busick, M.D. (2012). *Translating the Statistical Representation of the Effects of Education Interventions into More Readily Interpretable Forms*. (NCSER 2013-3000). Washington, DC: National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education.

Lipsey, M., & Wilson, D. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist, 48*(12), 1181–1209.

Lopéz, X., Valenzuela, J., Nussbaum, M., & Tsai, C. (2015). Some recommendations for the reporting of quantitative studies. *Computers & Education, 91*, 106-110.

McCormick, A. C., Kinzie, J., & Gonyea, R. M. (2013). Student engagement: Bridging research and practice to improve the quality of undergraduate education. In M. B. Paulsen (Ed.). Higher Education: Handbook of Theory and Research (Vol. 28, pp. 47–92). Dordrecht, The Netherlands: Springer.

McMillan, J. H., & Foley, J. (2011). Reporting and discussion effect size: Still the road less traveled? Practical Assessment, *Research & Evaluation, 16*(14), 1–12.

Miller, A. L., Sarraf, S. A., Dumford, A. D., & Rocconi, L. M. (2016). *Construct validity of NSSE Engagement Indicators*. Bloomington, IN: Center for Postsecondary Research.

National Survey of Student Engagement. (2015). *Lessons from the field–Volume 3: Using data to catalyze change on campus*. Bloomington, IN: Center for Postsecondary Research, Indiana University School of Education.

National Survey of Student Engagement. (2017). *Lessons from the field–Volume 4: Digging deeper to focus and extend data use*. Bloomington, IN: Center for Postsecondary Research, Indiana University School of Education.

Pace, C. R. (1979). *Measuring outcomes of college: Fifty years of findings and recommendations for the future*. San Francisco: Jossey-Bass.

Pascarella, E. T., & Terenzini, P. T. (1991). *How college affects students: Findings and insights from twenty years of research*. San Francisco: Jossey-Bass.

Pascarella, E. T., & Terenzini, P. T. (2005). *How college affects students: A third decade of research*. San Francisco: Jossey-Bass.

Pearson, K. (1900). On the correlation of characters not quantitatively measurable. *Philosophical Transactions of the Royal Society of London, 195*, 1–47.

Peng, C. J., Chen, L., Chiang, H, & Chiang, Y. (2013). The impact of APA and AERA guidelines on effect size reporting. *Educational Psychological Review, 25*(2), 157–209.

Rocconi, L. M., Lambert, A. D., McCormick, A. C., & Sarraf, S. A. (2013) Making college count: An examination of quantitative reasoning activities in higher education. *Numeracy, 6*(2), Article 10. DOI: http://dx.doi.org/10.5038/1936-4660.6.2.10

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44*(10), 1276–1284.

Rosnow, R. L., & Rosenthal, R. (2003). Effect sizes for experimenting psychologists. *Canadian Journal of Experimental Psychology, 57*(3), 221–237.

Smart, J. C. (2005). Attributes of exemplary research manuscripts employing quantitative analyses. *Research in Higher Education, 46*(4), 461–477.

Springer, R. (2006). Using effect size in NSSE survey reporting. *Research & Practice in Assessment, 1*, 18–22.

Thompson, B. (1998). In praise of brilliance: Where the praise really belongs. *American Psychologist, 53*(7), 799–80.

Thompson, B. (2001). Significance, effect sizes, stepwise methods, and other issues: Strong arguments move the field. *Journal of Experimental Education, 70*(1), 80–93.

Vacha-Haase, T., & Thompson, B. (2004). How to estimate and interpret various effect sizes. *Journal of Counseling Psychology, 51*(4), 473–481.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician, 70*(2), 129–133.

Wilkinson, L., & American Psychological Association Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*(8), 594–604.

Abstract

This qualitative study examines how service-learning pedagogy can facilitate graduate students' learning of assessment. Interviews with 14 students enrolled in a student affairs graduate program suggest that a.) direct application of content knowledge through a service-learning course enhanced students' learning of assessment, b.) exposing students to the utility of their assessment findings deepened students' understanding of the value of assessment in student affairs, and c.) students shifted their focus from grades to conducting a rigorous assessment study as they experienced the value others placed on their service.

AUTHORS

Blanca Rincón, Ph.D.
University of Nevada, Las Vegas

Milagros Castillo-Montoya, Ed.D.
University of Connecticut

Learning Assessment in Student Affairs Through Service-Learning

Rising college costs, coupled with declining resources, has prompted an accountability movement in higher education. Between 2003 and 2014, the Government Accountability Office reported that, on average, state appropriations for public colleges and universities decreased by 12% while tuition rates rose by 55% (Emrey-Aaras, 2014). As resources become scarce, higher education leaders are increasingly being asked to respond to constituents' needs for creating opportunities for social mobility, leading innovation, and preparing students for the workforce (Alexander, 2000). Consequently, policymakers are scrutinizing postsecondary education to determine how it fares in terms of access, affordability, student retention, graduation rates, job placement, and student learning (Callan, 2008).

Under this mounting pressure to show results, institutional leaders are asking student affairs professionals to increasingly engage in assessment to demonstrate their contributions to student learning and development, as a matter of survival, and for decision-making (Schuh & Associates, 2001). Indeed, the Joint Task Force on Professional Competencies and Standards representing College Student Educators International (ACPA) and the National Association of Student Personnel Administrators (NASPA) identifies assessment, evaluation, and research (AER) as one of 10 professional competency areas expected of student affairs educators (ACPA/NASPA, 2015). According to ACPA and NASPA (2015), student affairs professionals competent in AER will have the "ability to design, conduct, critique, and use various AER methodologies and their results to inform practice" (p. 12). Foundational AER outcomes also include being able to facilitate "appropriate data collection," and understand how to assess the "legitimacy, trustworthiness and/or validity of various methods" (p. 20). Importantly, AER foundational outcomes also include knowing how to communicate results in an "accurate, responsible, and effective" way as well as with sensitivity to "organizational hierarchies" (p. 20). These foundational outcomes necessitate that student affairs professionals are taught how to conduct assessment while responding to various stakeholders within higher education.

CORRESPONDENCE

Email
blanca.rincon@unlv.edu

Despite the increasing emphasis on assessment in higher education, and specifically in student affairs, efforts toward developing a culture of, knowledge in, and skills around assessment within student affairs have been slow. Faculty of graduate preparation programs and senior student affairs administrators rate assessment as one of the most desired competencies among new student affairs professionals, yet new student affairs professionals display large gaps in their knowledge for conducting assessment (Dickerson et al., 2011). This finding is troubling given that job postings increasingly ask that entry-level student affairs professionals demonstrate knowledge of assessment, evaluation, and research. In fact, almost half of all 2008 student affairs job postings through The Placement Exchange—an annual placement conference held at the national meeting of NASPA geared towards entry-level student affairs professionals—included assessment, evaluation, and research skills as part of the job description (Hoffman & Bresciani, 2012). Beyond inadequate preparation and the growing demand of assessment skills, Elkins (2015) finds that newer professionals are overwhelmingly represented in the lower stages of Erwin's (1991) five stages of reacting to assessment—discovery, questioning, resistance, participation, and commitment (as cited in Elkins, 2015).

Despite the increasing emphasis on assessment in higher education, and specifically in student affairs, efforts towards developing a culture of, knowledge in, and skills around assessment within student affairs have been slow.

The past two decades have seen a proliferation of assessment literature within higher education and student affairs. This literature spans from discussions about the philosophical underpinnings of assessment to "how-to" guides that discuss assessment plans and implementation efforts (Elkin, 2015). Absent, however, is a focus on pedagogical approaches that promote the learning of assessment for future student affairs professionals. As such, we need to better understand how student affairs professional preparation programs can help prepare graduate students entering the student affairs profession to engage in the practice of assessment. In response to this need, we aimed to address the following research question: How, if at all, does engaging in a service-learning assessment project facilitate the learning of assessment among graduate students enrolled in a student affairs professional preparation program?

We chose to study graduate students' learning of assessment within the context of a graduate class designated as a service-learning course because this type of teaching provides students with a form of experiential learning that aligns with the theory-to-practice model pursued in many student affairs graduate programs in the United States. We use Kuh's (2008) definition of service-learning as a "field-based experiential learning with community partners… [where] students have to both apply what they are learning in real-world settings and reflect in a classroom setting on their service experiences" (p. 11). For this project, we defined community partners broadly to include the campus community since the graduate students in this course engaged in service student affairs departments on campus. As such, we see graduate students engaging in service with campus partners as an opportunity to prepare student affairs professionals to engage in collaborative work and service to the field of student affairs. Further, because the assessment-project sites selected for this course during the year of this study served traditionally underserved student populations on campus (e.g., LGBT, Latino/a), students also had to engaged in discussions of power and 'otherness' within the context of learning assessment.

Conceptual Framework

To explore how engaging in an assessment project through a service-learning course facilitates the learning of assessment among graduate students enrolled in a student affairs professional preparation program, we constructed a two-part framework comprised of the concepts of situated cognition and service-learning.

Situated Cognition

Situated cognition entails supporting students' learning by having them engage in authentic activities—"ordinary practices of the culture"—that resemble what practitioners in that field would potentially face (Brown, Collins, & Duguid, 1989, p. 34). Brown et al. (1989) state, "people who use tools actively rather than just acquire them, by contrast, build an

increasingly rich implicit understanding of the world in which they use the tools and of the tools themselves" (p. 33). A valuable point here is that through situated cognition students learn tools—conceptual knowledge and skills—by directly using them as opposed to solely hearing about them from the instructor or reading about them in a book. Another feature of situated cognition is that students learn about the tools by using them within the context, in this case, community, where that tool would naturally be used. In doing so, students also learn *how* to use the tool within that community, and that experience becomes part of the learning too.

Situated cognition can be valuable because learning how to use a tool is one thing but knowing how to use it in a real context, where the context itself may shape how you use the tool, can be completely different. Brown et al. (1998) wrote, "The community and its viewpoint, quite as much as the tool itself, determine how a tool is used" (p. 33). They assert that often students learn tools in the abstract and without knowing how to use what they know within the context of their work. Yet, the context provides information, structures, and cues that would inform the use of tools, and therefore the learning of them. As such, situated cognition entails cognitive apprenticeship—an opportunity for students to "acquire, develop, and use cognitive tools in authentic domain activity" (Brown et al., p. 39).

Cognitive apprenticeship entails collaborative learning where novices and experts work together to learn.

> The apprenticeship system often involves a group of novices (peers) who serve as resources for one another in exploring the new domain and aiding and challenging one another… The 'master,' or expert, is relatively more skilled than the novices, with a broader vision of the important features of the culturally valued activity. However, the expert too is still developing breadth and depth of skill and understanding in the process of carrying out the activity and guiding others in it. (Rogoff, 1990, pp. 39)

Through cognitive apprenticeship, students learn from each other, and skilled experts, to collectively solve problems, engage in multiple roles, and work through ineffective strategies, misunderstandings, and misconceptions (Brown et al., 1998; Hennessy, 1993). The collective learning, however, is always grounded in the authentic activity to deepen students' knowledge and skills. For this project, the graduate students were the novices and the experts were the faculty teaching the assessment course—having expertise in conducting assessment, evaluation, and research through qualitative, quantitative, and mixed-method designs —and the campus partners who have expertise as practitioners in their functional areas.

A valuable point here is that through situated cognition students learn tools—conceptual knowledge and skills—by directly using them as opposed to solely hearing about them from the instructor or reading about them in a book.

Service-Learning

Service-learning, as a way of teaching, is a high-impact practice that integrates community service with instruction and reflection (Kuh, 2008; National Service-Learning Clearinghouse, 2007). Service-learning has been associated with student gains in content knowledge, critical thinking (Astin & Sax, 1998), and identity development of undergraduate students (Jones & Abes, 2004). Researchers have documented the long-term impact of service-learning for undergraduate students. For example, Fullerton, Reitenauer, and Kerrigan (2015) found that students identify service-learning as a significant learning experience 3–15 years after completing a service-learning course. Further, participants in the study identified specific "epiphanic" moments that led to their learning about others, altered their perspectives, and enhanced interpersonal communication. The ability to vividly remember these experiences is likely a product of the service-learning environment. That is, service-learning settings produce strong emotional experiences often not experienced in traditional course offerings (Noyes, Darby, & Leupold, 2015). Moreover, project-based service-learning strategies, such as those employed in the course that informs this study, develop undergraduate students' technical, critical thinking, and interpersonal skills (Gomez-Lanier, 2016).

Service-learning has also been applied to undergraduate students' learning of research methods (Curwood, Munger, Mitchell, Mackeigan, & Farrar, 2011; Nigro & Wortham, 1998; Stocking & Cutforth, 2006). For example, Nigro and Wortham (1998) find that students engaged

in community action research value the direct hands-on experience gained from thinking through complex problems on their own. Despite the educational benefits of service-learning at the postsecondary level, little is known about the benefits of service-learning courses at the graduate level or specifically within student affairs graduate preparation programs. As such, this study draws on graduate student data from students enrolled in a service-learning assessment course to examine students' engagement and learning of assessment.

These two concepts, situated cognition and service-learning, frame this project in useful ways. Situated cognition helps us see the value of having in- and outside-classroom experts supporting the learning of novices regarding assessment within authentic situations of practice, or what Brown et al. (1998) refer to as cognitive apprenticeship. Through service-learning, we are able to frame students' opportunity to learn content knowledge, apply it in a real-world setting, and importantly to service-learning, reflect on their learning of the content and the experience of applying knowledge and skills in service to the field. Together, situated cognition and service-learning ground this project in ways reflective of the course structure offered to participants.

Methods

To address the question of how engaging in an assessment project through a service-learning course facilitates the learning of assessment among graduate students enrolled in a student affairs professional preparation program, we conducted a qualitative study. We collected an in-person questionnaire, student reflections, and conducted two one-on-one in-person interviews with the graduate students who matriculated in the service-learning assessment course offered as part of a master's degree student affairs program at Northeast University (pseudonym). Northeast University is a large, public research (R1) university located in the Northeastern region of the United States. We selected Northeast University as the site for this study because students in this program are required to enroll in a two-part *assessment in student affairs* course sequence with a service-learning designation as part of the core curriculum. Students in the program represent diverse gender, racial, ethnic, and sexual identities. They also have a wide range of professional experiences, but most students enroll directly after completing their bachelor's degrees. Because the instructors of the course were also the researchers of the study, student consent was not requested until the conclusion of each semester to reduce the possibility that students experienced any pressure to participate in the study.

As the instructors, we worked in partnership with four offices within student affairs departments to identify the assessment projects. We selected the assessment projects based on need, scope, and office resources. After selecting the projects, we assigned students to groups based on students' prior experiences with assessment as well as interest in the project. Stocking and Cutforth (2006) suggest that students who feel a sense of connection to their community partners display flexibility, patience, and personal investment when engaging in their research projects, regardless of prior research experience.

Situated cognition can be valuable because learning how to use a tool is one thing but knowing how to use it in a real context, where the context itself may shape how you use the tool, can be completely different.

Students enrolled in the year-long course sequence in the first semester of their first year. We used Jacobson's (2015) method for conducting rigorous, scholarly assessment to guide our teaching of the course. This includes developing clear goals, leveraging and building expertise, using appropriate research methods, interpreting results, disseminating work, and engaging in the peer review process. The fall semester consisted of students developing the research design and plan that informed their assessment activities for the spring semester (e.g., data collection and data analysis). Further, the two-course sequence is essential to addressing pedagogical challenges that may arise when students lack the readiness to engage with basic principles of assessment or cultural competencies for engaging with community partners (Stocking & Cutforth, 2006). Each class session was intentionally structured so that students applied the content covered in class to their assessment projects through a variety of exercises. For example, students learned about the components of developing a good questioning route and then created an interview protocol in line with their assessment project that reflected that learning.

The course structure aligns with the conceptual framework in that the graduate students developed and used assessment tools to enact assessment in actual professional practice, thus reflecting situated cognition. In terms of service-learning, the graduate students in this course conducted an assessment as a service to a program or office situated within student affairs departments at the research site. In this sense, the course took on another element—not only enacting assessment in actual practice but doing so to the benefit of student affairs programs and services.

Data

Cognitive apprenticeship entails collaborative learning where novices and experts work together to learn.

In the fall semester when this study took place, students in the course completed an in-person questionnaire during the first day of class that helped instructors place them into their assessment groups. The questionnaire asked for graduate students' background information including their gender, race, education, and prior experience with assessment, evaluation, and research. This information helped us understand their demographic backgrounds as well as the transferrable skills and knowledge they may have brought to their learning of assessment.

In addition to completing the questionnaire, students who agreed to participate in the study were also asked to participate in two one-on-one interviews, one at the end of each academic semester. Interviews provide useful data for understanding how people make meaning of their experiences (Rubin & Rubin, 2012). In terms of assessment, Newhart (2015) argues that qualitative approaches more accurately depict the "complexity" and "depth" of student learning. Not only can qualitative approaches measure what students are learning they can also help us understand "why students are or are not learning" (Suskie, 2009, p. 24). Since we sought to better understand how students experienced learning about assessment through a service-learning course, interviews were a fitting method.

A member of the research team followed up with each student to schedule interviews at the end of each semester. Of the 20 potential participants, 14 participated in the spring interviews. The semi-structured interviews lasted 40–60 minutes, were audio-recorded, and were transcribed by a third party. The findings presented here are drawn exclusively from the 14 interviews conducted at the end of the spring semester where the semi-structured interview protocol intentionally asked questions about learning assessment through a service-learning course. For example, students were asked: "What aspects of the spring course do you think were most helpful to your learning?" and "How did it feel to learn about assessment through a service-learning course?" To systematically examine the role of service-learning in learning about the process of assessment for all students, we added these questions after the fall interview data yielded some student responses that spoke to the service-learning component of the class.

Participants. Table 1 provides an overview of the demographic characteristics of spring interviewees. A total of six participants identified their gender as male and eight identified as female. Seven of the participants self-identified their race as White, three as Latino/a, two as mixed race, one as Asian American, and one as African American. The average age of participants was 23 years. Overwhelmingly, participants indicated that they had prior experience with assessment, evaluation, and research. Most of these experiences were the result of undergraduate research experiences under the guidance of faculty.

Analytical Approach

Before analyzing the interview data that informed this study, researchers de-identified each transcript by replacing student and program names with pseudonyms. Then, researchers reviewed the audio files to ensure that the transcripts were accurate. Next, researchers read each transcript to identify emerging concepts and codes (Rubin & Rubin, 2012) to develop a qualitative codebook—which kept a log of emerging codes and definitions. To this end, each researcher reviewed and coded two interview transcripts to identify initial concepts and codes (Saldaña, 2013) such as "service-learning_relationships," "service-learning_prior experience," and "service-learning_emotion."

Table 1
Participant Demographics (n=14)

	n	%
Gender		
Female	8	57
Male	6	43
Race/ethnicity		
White	7	50
Latino/a	3	21
Mixed race	2	14
Asian American	1	7
African American	1	7
Prior Experience with Assessment, Evaluation, and/or Research		
Yes	12	86
No	2	14

Note. Numbers may not equal 100 due to rounding.

Once we defined an initial set of concepts and codes, we engaged in the independent coding of the remaining data using NVivo software. We then brought our coding together to identify the similarities and differences in our coding by conducting an interrater reliability report through NVivo. The level of agreement across all codes averaged at 99.2% and ranged between 99.4% and 99.9%. This step strengthened the definition of concepts and codes and increased the reliability of the analysis (Miles, Huberman, & Saldaña, 2014). Upon completion of coding, we engaged in "second cycle coding" to identify emerging themes (Miles, Huberman, & Saldaña, 2014).

Findings

Our analysis of the data yielded three main findings related to the learning of assessment through a service-learning course. First, direct application of content knowledge through a service-learning course enhanced students' learning of assessment. Second, exposing students to the utility of their assessment findings deepened students' understanding of the value of assessment in student affairs. Finally, students shifted their focus from grades to conducting a rigorous assessment study as they experienced the value others placed on their service.

Direct Application

Study participants described the direct application of course content to their assessment projects as key to their learning of assessment. This was especially helpful for students who described themselves as learning best by engaging in authentic problems of practice. David summarized how his learning, as well as that of his peers, may have been limited using a different approach, "if you didn't have a service-learning component and you were talking about research from a more theoretical perspective, I think it would be difficult to connect with student affairs students in general." Tam echoed this sentiment when asked how she enjoyed learning assessment through a service-learning course:

> So, I think yes, the service-learning component was very helpful and allowed me to learn the theory part and then put that theory into practice. It was very helpful for my type of learning style because I want to know how it can be applicable to real life and how it can be applicable to student affairs in general, not just learning about it and hearing about it and hearing examples. But actually, being able to do it myself.

Beyond application, which could have emerged from engaging in a generic assessment project, or case studies, students also described the benefits of learning assessment by engaging in assessment projects situated within the context of student affairs, and across various functional areas. This is illustrated by Carlos who reflected on his learning assessment through a service-learning course: "I think I was able to really make that connection. For me, it was like if I could do this for this learning community, I could do this all the time in my professional career, but I think some people may not look at it like that." Many of the student participants expressed this point that having an opportunity to directly apply what they were learning in the assessment course helped them make a connection between principles of assessment and the doing of it.

Utility of Assessment

Several students also indicated that their learning of assessment was enhanced because they were working with community partners who intended to use their findings to inform program improvement. This is especially true for one group of students who worked on an assessment project for a program that was at risk of losing program funds for a mentoring program that provided college outreach to underserved students of color due to state budget cuts. By working on an assessment project of a program that was in the position of having to defend its existence, students were exposed to the financial realities of programs that exist almost exclusively on soft funds, as explained by Brittney:

Through service-learning, we are able to frame students' opportunity to learn content knowledge, apply it in a real-world setting, and importantly to service-learning, reflect on their learning of the content and the experience of applying knowledge and skills in service to the field.

> I realized that this wasn't just about me, theory-to-practice, getting to learn while doing, which is very beneficial, but literally hearing about Engaging Children in Higher Education [pseudonym] being in a state where it might– the grant might not be renewed. That's– a lot has to do with the school districts and the governor, all these budget cuts, etcetera. But just seeing all the sponsors really say like we're literally going to use these findings.

Abigail, another student from the same group, expressed similar sentiments:

> So that just made it so much more meaningful, to be working with real people on a real project. These are real experiences, and our research could– I mean they've already submitted the grant now, but in the future, it could help them to keep this program because statistics like 96% of [mentors] were satisfied with their experience, like hello, that's a really good number. And they're gaining skills, and they're learning and all the other things in the presentation.

As Abigail described, the application of the findings to program improvements made engaging of assessment "real," with real stakeholders and consequences.

While students were invested in their assessment projects and seeing the results of their assessments being used, they also experienced how engaging in assessment prepared them as student affairs professionals. They saw how learning assessment, and applying the findings of their assessments, was useful for their development as future student affairs professionals who could be running similar programs in the near future. Brittney shared, "I think me being able to take that and say not only did we do theory-to-practice, the departments actually utilize our findings and our recommendations on a job interview or wherever, feeling like a little consultant." Brittney saw assessment as a way to set herself apart from her peers.

A Shift from Grades to Rigor

Lastly, students reported being invested in their learning of assessment because of the "real" implications of their work. Students were not asked to apply course content to a fictitious project. Instead, students were asked to conduct an actual assessment project with community partners that had real assessment needs. Students in this study indicated that they were less focused on their grade in the course and more focused on conducting a rigorous assessment that could provide the most useful information for their community partners and affiliated program.

Many students spoke of feeling that they had to work harder than they would have if they were just submitting a paper for a grade. For example, Benjamin reflected that while the theory-to-practice model was one of the most helpful components of his learning in the course, knowing that the project and the final paper was in service to a program raised the stakes. As such, he focused on "really" learning the course content so that he could apply it to his project and produce a better product for their community partners:

> I think having that in-class time, the reading time to say that and then apply it, but also apply it with the idea of giving. I think it was a little extra boost to be like I better do this really good because I'm providing a resource, I'm providing information to [community partners] so I want to really make sure that I get this… So that when I'm applying it, I can really serve to the best of my abilities as well. I think that that was– it was up'd level of attention and focus that I needed to give to the learning happening in this class because there was going to be a result produced that wasn't just a hypothetical result.

Students often expressed their shift from grades to learning and producing "good work" within the context of building relationships. They did this by describing their connection with the assessment projects and/or the relationship they developed with their community partners. This was expressed by Abigail, "if this was just like any mentoring program that wasn't tied to a cultural center, I don't think I would care as much as I do because it's [Engaging Children in Higher Education] and because I know Leah [Director]."

Discussion

In terms of service-learning, the graduate students in this course conducted an assessment as a service to a program or office situated within student affairs departments at the research site. In this sense, the course took on another element—not only enacting assessment in actual practice but doing so to the benefit of student affairs programs and services.

A service-learning assessment course, where students are in service to offices and programs within the larger university community that have assessment needs, provides an opportunity for students to learn assessment through cognitive apprenticeship, which entails having the processes of the task made visible to students, having abstract tasks situated in authentic contexts, and varying the situations to promote transfer of learning (Collins, Brown, & Holum, 1991). Students, in this service-learning course, had opportunities to learn about assessment in class and in the field, through promoting transfer of learning. This transfer of learning was evidenced directly with the finding of "direct application" whereby students in this course indicated deepening their learning of assessment because of the opportunity to apply immediately what they were learning in class. For instance, in class, they discussed course readings and began to consider how readings helped them develop tools for their assessment project. These readings contributed to students developing knowledge about the difference between assessment, research, and evaluation; how to assess the legitimacy of studies; consider strengths and limitations of different methodologies; and how to use scholarly literature to inform the content and design of assessment tools. These outcomes align with ACPA and NASPA (2015) AER outcomes.

In the field, they learned about the practices of the office and gained insight from community partners that informed how they carried out their assessment, when, and what they assessed. This approach made the learning of assessment processes visible and situated it in student affairs contexts, thus contributing to the situated cognition that is part of a cognitive apprenticeship. Students in this study experienced situated cognition by engaging in authentic activities of practice (Brown et al., 1989), which included meeting student affairs stakeholders to gain a sense of the purpose and value of the assessment, to deepen their understanding of the community and context, and to develop relationships with the community being served. This authentic practice of meeting with the stakeholders aligns with ACPA and NASPA's (2015) intermediate AER outcomes of knowing how to appropriately design assessment "based on critical questions, necessary data, and intended audience(s)" (p. 20). The students had the opportunity to hone this outcome in their meetings with stakeholders (their intended audience). Those meetings were real, and not hypothetical, situating their learning in authentic contexts where they had to manage the relationships while asking critical questions to guide the assessment project. In addition, they developed the foundational

AER competency of developing sensitivity regarding the raw data and "handling them with appropriate confidentiality and deference to organizational hierarchies" (p. 20).

Students in this course also developed assessment tools (i.e., surveys, interview protocols) in partnership with the community partner, as well as in a team-based approach, thus exemplifying the cognitive apprenticeship that can contribute to situated cognition. Developing assessment tools with their community partner also meant that these students learned how to "select… tools that fit with the research and evaluation questions and with assessment and review purposes" as well as facilitate "appropriate data collection;" both of which are foundational AER competencies (ACPA/NASPA, 2015, p. 20)

This experience of carrying out assessment in an authentic student affairs office helped these future student affairs professionals learn how to conduct assessment in student affairs in a meaningful way. Students did not learn only for themselves but also engaged in learning assessment to serve others. They learned about how to meet stakeholders' needs and how assessment findings can be used to inform programmatic improvements or to leverage future funding during difficult economic times. In doing so, students could see how assessment could be useful for the work they do in their assistantships, as well as the work they will do in the future as student affairs professionals. Equally as important, students saw for themselves the potential consequences of opting-out of doing assessment—that is, not having the "evidence" to defend a program during hard economic times. These learning gains align with what Brown et al. (1998) refer to as learning tools by using them specifically within the community where they will be practiced.

Future inquiry should also seek to investigate whether service-learning, and specifically situated cognition, is helpful in the development of other student affairs competencies.

Limitations

The authors identify several limitations to consider when interpreting the results of this study. First, student accounts of how service-learning facilitated their learning of assessment is an indirect account of student learning. Future research should seek to combine both direct and indirect measures of student learning when determining the significance of service-learning pedagogical strategies in the learning of assessment. Second, the sample of students is limited to students in one graduate preparation program at one institution. As such, the design of the study limits our ability to generalize findings to other programs at other institutions.

Implications for Research and Practice

This study begins a line of inquiry about how service-learning can prepare student affairs professionals enrolled in graduate preparation programs to develop their competency in assessment, evaluation, and research. Future research is needed to examine how learning assessment through a service-learning course may affect student views of engaging in assessment over time. That is, are students who complete such a course more apt than their colleagues who have not taken such a course to engage in assessment as part of their work as student affairs practitioners? Future inquiry should also seek to investigate whether service-learning, and specifically situated cognition, is helpful in the development of other student affairs competencies.

Study findings also have implication for practice. Findings indicate that students who engage in learning about assessment through a service-learning assessment course can deepen their understanding of how assessment can have direct application for practice. Assessment, with its research foundation, may not initially be a practice student affairs professionals view as essential to their work. However, by seeing the direct application, students can develop into practitioners who see the value of conducting assessment as a form of student affairs practice.

Study findings have implications for campus partners as well. Campus partners who sponsor students may find that having enthusiastic students doing assessment for them may lead to synergy in their offices that can contribute to a culture of assessment—an intermediate AER outcome (ACPA/NASPA, 2015). They may also continue to sharpen their own assessment skills and knowledge as they collaborate in the cognitive apprenticeship experience.

Lastly, study findings have implications for faculty teaching in professional preparation programs such as student affairs. Graduate faculty who teach assessment may want to consider using service-learning pedagogy, with a cognitive apprenticeship lens, to deepen students' subject-matter learning while serving local communities. Even more, graduate faculty may want to consider other courses that could benefit from a service-learning approach anchored in situated cognition to enhance learning across the graduate education curriculum.

Conclusion

In conclusion, this study found that direct application of content knowledge through a service-learning course enhanced students' learning of assessment. Also, exposing students to the utility of their assessment findings deepened students' understanding of the value of assessment in student affairs. Lastly, students shifted their focus from grades to conducting a rigorous assessment study as they experienced the value others placed on their service. Such findings can inform how graduate preparation programs in student affairs, and faculty who teach in these programs, can leverage service-learning as a pedagogical tool when teaching assessment courses to build the competency of assessment among future student affairs professionals.

References

ACPA: College Student Educators International & NASPA: Student Affairs Administrators in Higher Education (2015). *Professional competency areas for student affairs educators*. Washington, DC.

Alexander, F. K. (2000). The changing face of accountability: Monitoring and assessing institutional performance in higher education. *The Journal of Higher Education*, *71*(4), 411–431.

Astin, A. W., & Sax, L. J. (1998). How undergraduates are affected by service participation. *Journal of College Student Development*, *39*(3), 251–263.

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher*, *18*(1), 32–42.

Callan, P. (2008). The 2008 national report card: Modest improvements, persistent disparities, eroding global competitiveness. *Measuring Up 2008: The National Report Card on Higher Education*, 1–31.

Collins, A., Brown, J. S., & Holum, A. (1991). Cognitive apprenticeship: Making thinking visible. *American Educator*, *15*(3), 6–11.

Curwood, S. E., Munger, F., Mitchell, T., Mackeigan, M., & Farrar, A. (2011). Building effective community-university partnerships: Are universities truly ready? *Michigan Journal of Community Service Learning*, *17*(2), 15–26.

Dickerson, A. M., Hoffman, J. L., Anan, B. P., Brown, K.F., Vong, L.K., Bresciani, M. J., Monzon, R., & Oyler, J. (2011). A comparison of senior student affairs officer and student affairs preparatory program faculty expectations of entry-level professionals' competencies. *Journal of Student Affairs Research and Practice*, *48*(4), 463–479.

Elkins, B. (2015). Looking back and ahead: What we must learn from 30 years of student affairs assessment. *New Directions for Student Services*, *151*, 39–48.

Erwin, T. D. (1991). *Assessing student learning and development: A guide to the principles, goals, and methods of determining college outcomes*. San Francisco, CA: Jossey-Bass.

Emrey-Aaras, M. (2014). *Higher education state funding trends and policies on affordability*. Report to the Chairman, Committee on Health, Education, Labor, and Pensions, United States Senate. GAO-15-151. Washington D. C.: Government Accountability Office.

Fullerton, A., Reitenauer, V. L., & Kerrigan, S. M. (2015). A grateful recollecting: A qualitative study of the long-term impact of service-learning on graduates. *Journal of Higher Education Outreach and Engagement*, *19*(2), 65–92.

George-Jackson, C.E., & Rincón, B. (2012). Increasing sustainability of STEM intervention programs through evaluation. *ASQ Higher Education Brief* (Special Issue on STEM), 5(1).

Gomez-Lanier, L. (2016). The effects of an experiential service-learning project on residential design student attitudes toward design and community. *International Journal for the Scholarship of Teaching and Learning*, *10*(2), 1–7.

Hennessy, S. (1993). Situated cognition and cognitive apprenticeship: Implications for classroom learning. *Studies in Science Education*, *22*(1), 1–41.

Hoffman, J. L., & Bresciani, M. J. (2012). Identifying what student affairs professionals value: A mixed methods analysis of professional competencies listed in job description. *Research & Practice in Assessment*, *7*, 26–40.

Jacobson, W. (2015). Sharing power and privilege through the scholarly practice of Assessment. In S. K. Watt (Ed.), *Designing transformative multicultural initiatives: Theoretical foundations, practical applications, and facilitator considerations* (pp. 89–102). Sterling, VA: Stylus.

Jones, S. R., & Abes, E. S. (2004). Enduring influences of service-learning on college students' identity development. *Journal of College Student Development, 45*(2), 149–166.

Kuh, G. D. (2008). *High-impact educational practices: What they are, who has access to them, and why they matter*. Washington, DC: Association of American Colleges and Universities.

Miles, M. G., Huberman, A. M., & Saldaña, J. (2014). *Qualitative data analysis: A methods sourcebook* (3rd ed). Los Angeles, CA: Sage.

National Service-Learning Clearinghouse (2007). *What is service-learning?*

Newhart, D. W. (2015). To learn more about learning: The value-added role of qualitative approaches to assessment. *Research & Practice in Assessment*, *10*, 5–11.

Nigro, G., & Wortham, S. (1998). Service-learning through action research. In R. G. Bringle & D. K. Duffy (Eds.) *Collaborating with the community: Psychology and service-learning*. Washington DC: American Association for Higher Education.

Noyes, E., Darby, A., & Leupold, C. (2015). Students' emotions in academic service-learning. *Journal of Higher Education Outreach and Engagement, 19*(4), 63–84.

Rogoff, B. (1990). *Apprenticeship in thinking: Cognitive development in social context*. Oxford: Oxford University Press.

Rubin, H. J., & Rubin, I. S. (2012). *Qualitative interviewing: The art of hearing data* (3rd ed.). Thousand Oaks, CA: Sage.

Saldaña, J. (2013). *The coding manual for qualitative researchers* (2nd ed.). Thousand Oaks,CA: SAGE.

Schuh, J. H., & Associates. (2009). *Assessment methods for student affairs*. San Francisco, CA: Jossey-Bass.

Stocking, V. B., & Cutforth, N. (2006). Managing the challenges of teaching community-based research courses: Insights from two instructors. *Michigan Journal of Community Service Learning, 13*(1), 56–65.

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

Abstract

Teacher education programs are under considerable pressure to evaluate their effectiveness in training new teachers. Over the last several decades there have been repeated calls for more systematic research on preservice teacher preparation programs. One institution has heeded this call, using the Video Assessment of Interactions and Learning (VAIL: Jamil, Sabol, Hamre, & Pianta, 2015) to annually assess teacher noticing of teacher-student interactions in preservice teachers. While there was no significant difference in first-test and last-test VAIL scores, VAIL scores were shown to be related to final college GPA. Additionally, there was a difference in student effort on the VAIL assessment, as participants provided fewer responses at the end of their program than at the beginning. These data show that assessments can be overused with regard to preservice teacher education programs. Future VAIL implementation should consider assessment fatigue when designing preservice teacher evaluation.

AUTHORS

Peter D. Wiens, Ph.D.
University of Nevada, Las Vegas

Matthew D. Gromlich, Ed.D.
University of Nevada, Las Vegas

Five Years of Video-Based Assessment Data: Lessons from a Teacher Education Program

Teacher education programs are under considerable pressure to evaluate their effectiveness in training new teachers. These pressures come from a variety of sources including state and federal governments, accreditation agencies, media, and potential preservice teachers (Feuer, Floden, Chudowsky, & Ahn, 2013). The fractured nature of the United States education system has led to different approaches across various institutions, as each institution must answer to its own unique set of stakeholders (Feuer et al., 2013). The Council for the Accreditation of Education Programs (CAEP: Council for the Accreditation of Education Programs, 2013), the largest national teacher education accreditation agency in the United States, includes standards requiring programs to show evidence that graduates are ready to be effective teachers. Yet, there remains a need in the field of teacher education to collect data that allows policymakers, researchers, and educators to better understand preservice teachers and the nature of their learning.

There have been repeated calls over the past decades for more systematic research on teacher education (Grossman & McDonald, 2008; Worrell et al., 2014; Zeichner, 2005). Reports issued by the National Academy of Education (Feuer et al., 2013) and an American Psychological Association Task Force (Worrell et al., 2014) provide guidance for the evaluation of teacher education programs. Both reports speak to the difficulty of effectively evaluating such a complex endeavor as training new teachers. Feuer and colleagues (2013) speak to the need for programs to use their core principles when designing evaluation systems, and Worrell and colleagues (2014) state:

> The data and methods required to evaluate the effectiveness of teacher education programs ought to be informed by well-established scientific methods that have evolved in the science of psychology, which at its core addresses the measurement of behavior. (p. 2)

Pianta and Hamre (2009) call for studies that identify early markers of teacher quality through standardized measures. Using such measures, links can be drawn between certain programmatic relationships and quality teaching interactions (Pianta & Hamre, 2009).

CORRESPONDENCE

Email
peter.wiens@unlv.edu.

Teacher education programs use a variety of measures to provide programmatic data in order to address the requirements of various stakeholders (Feuer et al., 2013). One teacher education program made the decision to add to its assessment portfolio by adopting an empirically and theoretically supported video-based assessment of teachers' ability to identify effective teaching interactions. The Video Assessment of Interactions and Learning (VAIL: Jamil, Sabol, Hamre, & Pianta, 2015) was administered to all preservice teachers each year they were in the teacher education program. After five continuous years of data collection, this paper examines what lessons can be learned about the assessment and its usefulness for teacher education evaluation. This data-gathering effort is unique in the teacher education field and examining the data provided may inform other teacher education programs in designing their own assessment frameworks.

Video as an Assessment Tool

Teacher education programs around the world have adopted the use of video for training purposes (Christ, Arya, & Chiu, 2017; Gaudin & Chaliès, 2015). However, assessing preservice teachers' ability to examine videos of real-world classrooms is less frequently cited in the literature. The VAIL, built upon an empirical and theoretical framework that includes teacher noticing and teacher-student interactions, uses video analysis as a means of assessing preservice teachers' knowledge of teaching interactions. The noticing framework is first credited to Goodwin (1994) who wrote, "Professional vision is perspectival, lodged within specific social entities, and unevenly allocated" (p. 626). Van Es and Sherin (2002) further refined the noticing framework in the context of teacher education. They contend that there are three key components of noticing:

> (a) identifying what is important or noteworthy about a classroom situation; (b) making connections between the specifics of classroom interactions and the broader principles of teaching and learning they represent; and (c) using what one knows about the context to reason about classroom interactions. (van Es & Sherin, 2002, p. 573)

Noticing includes observing a situation, making interpretations, and then making a decision on what has been observed (Kaiser, Busse, Hoth, König, & Blömeke, 2015).

Yet, there remains a need in the field of teacher education to collect data that allows policy makers, researchers, and educators to better understand preservice teachers and the nature of their learning.

A key component of effective noticing is that experts in a field notice different things when observing a situation than do novices—a concept that has been developed and supported in cognition research (Feldon, 2007). Glasser and Chi (1988) developed a list of expert characteristics that includes perceiving problems at a deeper level and spending a larger time analyzing problems. Expert teachers examine information in different ways than do novice teachers (Bransford, Brown, & Cocking, 1999). When examining still photographs, Carter, Cushing, Sabers, Stein, and Berliner (1998) found that experts were more cautious in their observations and were more sensitive to the sequence of events in the classroom, whereas in the case of the novices, "the schema they brought to these visual information processing tasks did not seem as richly developed as experts" (p. 31). Using the expertise framework, the examination of videos can be used as an assessment to differentiate between novices and more expert teachers.

Video as an assessment tool has promise because the ability to effectively notice in videos of teaching has been shown to correlate with the ability to teach effectively (Kersting, Givvin, Sotelo, & Stigler, 2010; Santagata & Yeh, 2014). Specifically, the VAIL has been associated with observed teaching quality in both in-service (Hamre, et al., 2012) and preservice teachers (Wiens, 2014). The VAIL is an assessment conducted online where participants watch short videos (2–3 minutes) of real-world classrooms. The participants then are prompted to identify effective teaching strategies and specific behavioral examples of those strategies by typing in open text boxes. Participant responses are then coded by trained coders for accuracy. Previous research showed a moderate correlation between performance on the VAIL and the observed quality of teaching interactions in a student teaching placement (Wiens, 2014). Additional validity support for using the noticing

framework as an assessment in teacher education is based on evidence that preservice teachers can be trained to become better at noticing (Sherin & van Es, 2005; Star & Strickland, 2008; Stürmer, Seidel, & Schäfer, 2013).

Basing Video Analysis on Understanding Student-Teacher Interactions

Building on the concepts of noticing and expertise, effective assessment of video analysis must be based on a clear vision of effective teaching. The VAIL is also supported by theory and research on teacher-student interactions. These interactions are proximal processes that take place regularly, over an extended time, and serve as an important part of children's development (Bronfenbrenner, 1993). Developed through extensive classroom observations, the Teaching Through Interactions Framework (TTIF) organizes interactions into three domains: Emotional Supports, Classroom Organization, and Instructional Supports (Hamre et al., 2013).

After five continuous years of data collection, this paper examines what lessons can be learned about the assessment and its usefulness for teacher education evaluation. This data-gathering effort is unique in the teacher education field and examining the data provided may inform other teacher education programs in designing their own assessment frameworks.

The TTIF provides a framework for understanding teacher-student interactions. It is most often measured using the Classroom Assessment Scoring System (CLASS: Pianta, La Paro, & Hamre, 2008; Pianta & Hamre, 2009) in a standardized and reliable way (Cadima, Leal, & Burchinal, 2010; Graue, Rauscher, & Schefinski, 2009). All three domains of the TTIF have been linked to positive academic outcomes including vocabulary growth (Cadima et al., 2010), phonological awareness (Curby, Rimm-Kaufman, & Ponitz, 2009), reading (Pianta, Belsky, Vandergrift, Houts, & Morrison, 2008), and grades (Reyes, Brackett, Rivers, White, & Salovey, 2012). CLASS has been used as an assessment in teacher education (see Wiens, Hessberg, LoCasale-Crouch, & DeCoster, 2013), but as an assessment of teaching interactions—such as those observed during the student teaching experience—it is most valid late in teacher education programs because that is when preservice teachers have their most authentic teaching experiences.

Originally developed for a large-scale study of pre-kindergarten teachers (Hamre et al., 2012), the VAIL has been implemented as an assessment in teacher education evaluation due to the fact that it can be implemented at multiple points in a teacher education program including as a pretest before preservice teachers even begin their training (Wiens et al., 2013). The VAIL is based on the TTIF framework (Jamil et al., 2015) as it uses videos of in-service teachers selected because they demonstrate the different domains of the TTIF. The VAIL uses the CLASS framework to understand teaching interactions and the VAIL videos match to CLASS domains as shown in Table 1. This study examines longitudinal data collected over five years of administering the VAIL in a teacher education program.

Student Motivation in Teacher Education Assessments

The implementation of the VAIL in teacher education is unique, as it was administered to preservice teachers in their introduction to education course and every subsequent year they were enrolled in the teacher education program. Therefore, not only were we able to assess student scores on the VAIL itself but also student motivation over time. It is important to consider preservice teacher motivation on the VAIL assessment over time, as motivation has been shown to impact academic performance (Dev, 1997) and over-surveying preservice teachers may lead to reduced responsiveness (Porter, Whitcomb, & Weitzer, 2004).

Most teacher education programs use a variety of assessments to evaluate their programs. These assessments can be considered either high stakes for the preservice teachers or low stakes. High-stakes assessments have negative consequences for the individual if they do not pass. For example, many programs use passing rates on licensure exams as one assessment in their evaluations. These are high stakes because if the preservice teacher does not pass then he/she cannot become a licensed teacher. Low-stakes assessments have no potentially negative consequences to the preservice teacher. The VAIL is an example of a low-stakes assessment. The preservice teachers were required to complete the VAIL; however, the results of the VAIL were not reported to the preservice teachers and their performance had no impact on their movement through the teacher education program or ability to become a licensed teacher.

Table 1
Alignment of CLASS and VAIL domains and dimensions (adapted Piante and Hamre, 2009)

Domains	Pre-K Dimensions	Indicators
Emotional Supports	**Positive climate**	Relationships, Affect, Respect, Communication
	Negative climate	Negative Affect, Punitive Control, Disrespect
	Teacher Sensitivity	Awareness, Responsiveness, Action to Address Problems, Comfort
	Regard for Student Perspectives*	Flexibility, Autonomy, Peer Interactions, Student Expression
Classroom Organization	**Behavior Management**	Clear Expectations, Proactiveness, Redirection
	Productivity	Maximizing Learning Time, Efficient Routines and Transitions
	Instructional Learning Formats*	Learning Targets, Variety of Modalities, Active Facilitation, Student Engagement
Instructional Supports	**Concept development**	Analysis/Reasoning, Creativity, Integration
	Quality of feedback*	Feedback Loops, Scaffolding, Building on Responses, Encouragement
	Language modeling	Conversation, Open-endedness, Repetition/Extension, Advanced Language

*Diminsions included in the VAIL instrument

The VAIL, built upon an empirical and theoretical framework that includes teacher noticing and teacher-student interactions, uses video analysis as a means of assessing preservice teachers' knowledge of teaching interactions.

In the arena of low-stakes assessments, where there is no external motivation, it may fall on individuals' intrinsic motivation for them to successfully complete the task. "Intrinsic motivation is motivation that is animated by personal enjoyment, interest, or pleasure" (Lai, 2011, p.4). In academics, intrinsic motivation has been linked to task persistence and the amount of time a student will spend on a task (Brophy, 1983). However, a task where the student has little or no interest will generate less intrinsic motivation (Deci & Ryan, 1985; Woolfolk, 1990). Research indicates that students with high levels of intrinsic motivation function more effectively in school (Dev, 1997). Given the connection between intrinsic motivation and academic success it is important to examine the relationship between academic success (in this study grade point average) and both success and effort on the VAIL.

Study Purpose

Teacher education programs seek to find innovative ways to administer assessments that contribute to effective evaluation. In order to address this need, one teacher education program administered the VAIL (Jamil et al., 2015) which conforms to Worrell and colleagues' (2014) call for evaluation "informed by well-established scientific methods" (p.2). In this study we examine the following research questions:

1. Does the ability of preservice teachers to identify effective teaching interactions change over the course of a teacher education program?
2. When taking the VAIL multiple times, do preservice teachers continue to demonstrate equal effort?
3. Are there characteristics that predict either final VAIL scores or final effort on the VAIL?

The five-year experience of the teacher education program can inform the discussion of program evaluation.

Methods

Procedures

Preservice teachers were required to complete the VAIL every year they were in the teacher education program. The VAIL was administered through a website. The first opportunity participants had to take the VAIL was during the first two weeks of their introduction to education course, prior to being enrolled in the education program. Once enrolled in the teacher education program, preservice teachers were required to participate in a data pool (Wiens et al., 2013) where they needed to earn research credits every spring semester. The VAIL was a requirement of the data pool and the preservice teachers could take the VAIL online any time during the spring semester. The online interface for the VAIL is shown in Figure 1. The data pool and the administration of the VAIL were both done by a program-funded doctoral graduate assistant.

Given the connection between intrinsic motivation and academic success, it is important to examine the relationship between academic success (in this study grade point average) and both success and effort on the VAIL.

Every summer the teacher education program paid four doctoral students $1000 each ($4000 total each summer) to code the VAIL responses. The coding team attended a three-hour training session and were required to pass a reliability test with 80% agreement with a master code list prior to beginning coding. Once coding began, the coding team would have weekly drift-check meetings to ensure that coding was reliable. Any time a coder fell under 80% agreement with the master code list that coder would stop coding, retrain, pass a new reliability test, and then resume coding.

Figure 1
VAIL Online Interface

Context and Participants

Data for this study come from a highly selective public university in a mid-Atlantic state. The university has two teacher education programs that lead to teacher licensure: a five-year bachelor's plus master's degree (n=226) and a two-year postgraduate degree (n=48). There are four different programs: early childhood (n=3), elementary (n=114), secondary (n=113), and special education (n=44). Of the participants, 71% were female, 13% male, and 3% unspecified.

Data for this study included all preservice teachers with multiple VAIL scores. For preservice teachers with more than two VAIL scores we used only the first score and the last score. For some bachelor's students the scores may be spread over multiple years. However,

for the post-graduate students the two scores were always in consecutive years, as it is a two-year program. The total number of preservice teachers with multiple years of VAIL scores were 281.

Measures

Video Assessment of Interactions in Learning (VAIL). The VAIL (Jamil et al., 2015) consists of three videos of pre-school language arts classrooms. These videos are followed by prompts instructing participants to identify teaching strategies and specific examples of those strategies from the video. After watching the video, participants had the opportunity to provide five effective teaching strategies they identified from the video in an open-ended format. Examples of effective teaching strategies included in the VAIL would be scaffolding, eliciting student ideas, and variety of instructional modalities.

For each strategy the participant had the opportunity to provide a specific example of the strategy taken from the video. The assessment defines an example as, "a teaching method used to meet a specific goal" (VAIL, 2010). In other words, examples constituted specific actions observed in the video. For example, if a participant noted scaffolding as a strategy a matching example might consist of the teacher helping the student sound out the word the student was struggling to read.

Responses supplied by participants were open ended and were coded for accuracy against a master code list created by master coders. Any differences between coders and the master code list were reconciled based on standards identified in the CLASS (VAIL 2010). The VAIL was designed so that CLASS-specific terminology was not necessary to perform well on the assessment. Participants could use any synonymous terms that identified the teaching strategies indicated in the VAIL manual. The VAIL uses a standardized rating description as outlined in the VAIL Coding Manual (2010) to guide all coding decisions.

To analyze the VAIL data, sum scores were calculated. Previous analysis of VAIL data with in-service teachers presented evidence to support using a one-factor model for compositing VAIL scores using the strategy, example, match and breadth scores (Jamil et al., 2015). The completion variable is analyzed separately because it does not conceptually measure a participant's ability to detect effective teaching interactions; instead, it measures participants' persistence in completing the assessment.

As every teacher education program is should be guided by best practice as well as internal and external stakeholders, finding innovative ways to administer assessments that contribute to the effective evaluation of programs is critical.

When a CLASS-matched strategy was identified by the participant, a breadth score was also assigned. Each assigned breadth score corresponded to a specific CLASS indicator. The number of unique indicators supplied by participants was then summed to create a breadth score for the entire set of responses for that video. Two of the videos had four possible strategy categories while the third video contained five possible strategy categories. Additionally, if both the strategy and example supplied were correct, the response was coded based on whether the example was an accurate example of the strategy identified.

The completion score measured how many responses the participants wrote for each video. Participants were coded for each attempt at identifying a strategy and example even if the strategy and example were not correctly identified. Each participant was required to provide at least one strategy and example to continue in the assessment. While there was the opportunity to identify five strategies and examples, only one response was required to continue with the assessment. Any strategy-example pairs that were left blank were coded as a zero.

Jamil and colleagues (2015) suggest an analysis strategy that standardizes values within the different videos and then composites the videos into a single score. However, it may be easier to understand the results of the VAIL, particularly when examining longitudinal change, using a sum score. Additionally, using a sum score also facilitates comparison of participant scores across contexts and administrations of the VAIL by providing a fixed number for the final score. The drawback of this approach is that the videos do not all have the same total possible points and therefore one video might have a slightly smaller weight in the overall score than the other videos. The total possible points for the Regard for Student Perspectives video is 19, Instructional Learning Formats is 19 as well, and the Quality of Feedback video

total is 20. The differences in possible points comes from the breadth score which has a maximum of four strategies in Regard for Student Perspectives and Instructional Learning Formats, while there are five total strategies in Quality of Feedback. While a sum score makes the Quality of Feedback video slightly more important, the benefits of a sum score outweigh these disadvantages.

Grade Point Average (GPA). The GPA data used in this study was taken from the end of program, cumulative GPA. GPA at this institution is on a four-point scale. The GPA data was taken from administrative records provided by the Teacher Education Office. GPA scores ranged from 2.61 to 4.00. The mean GPA was 3.57 with a standard deviation of .28.

Completion scores are all significantly correlated with each other. Data analysis did not show a significant difference between first-test and last-test VAIL scores in this sample.

Analysis

For our analysis, we used the first time they took the VAIL (first-test) and the last time they took the VAIL (last-test). The completion score was used as a test of effort. We examined both the VAIL totals and examined the three individual videos. We began with descriptive analysis and correlation estimates to better understand the data. Next, we computed paired sample t-tests to examine differences in variables—particularly focused on examining the differences between first-tests and last-tests. Finally, we computed multiple regression analysis to determine the relationship between effort, the amount of times participants took the VAIL, GPA, and teaching area and VAIL scores and effort.

Results

We began with an examination of the data. Mean scores for GPA and VAIL times taken are in Table 2 while first-test and last-test VAIL and Completion scores are presented in Table 3. Correlations, illustrated in Table 4, indicate that the first- and last- VAIL and first- and last- Completion scores are all significantly correlated with each other. Data analysis did not show a significant difference between first-test and last-test VAIL scores in this sample. We did find that preservice teachers scored higher on the first video than the last video (difference=.303, p=.09); however, this was only significant at the less stringent .1 value. Additionally, t-test analysis found that participants provided fewer responses to the third video last-test (Mean difference=.347, p=.001). In total, participants had lower Completion scores in the last-test than in the first-test (Mean difference= .518, p=.019).

Table 2
Mean Values for Variables

	Mean	SD
VAIL Times Taken	2.04	.20
GPA	3.57	.28

Table 3
Regression Table with Standardized Betas

	Mean (SD)			
	First-test	Last-test	First-test Completion	Last-test Completion
VAIL total	15.99	15.49	12.12	11.60
	(6.69)	(6.96)	(2.95)	(3.12)
Video 1	4.51	4.81	4.10	4.05
	(2.96)	(2.97)	(1.11)	(1.11)
Video 2	5.13	5.03	4.09	5.03
	(2.78)	(3.17)	(1.14)	(3.17)
Video 3	6.21	5.56	3.95	3.60
	(3.39)	(3.64)	(1.25)	(1.57)

Our regression analysis is shown in Table 5. When entering the times the VAIL was taken (Times Taken), GPA, and Teaching Area as predictors, we found the overall model to be significant for both last-test VAIL effort (Final R = .218, p=.02) and last-test VAIL scores (Final R = .237, p=.05). Within the regression model predicting final VAIL effort, Times Taken (Standardized β= -.168, p=.01) and Teaching Area (Early Childhood compared to Elementary: Standardized β= -.156, p=.02) were both significantly associated with the VAIL. Within the regression model predicting last-test VAIL score, GPA was the only individual variable that was significant (Standardized β= .172, p=.01).

Table 4
Bi-variate Correlations

	1	2	3	4	5	6
1. First-test VAIL	1	.353***	.501***	.148**	-.083	.076
2. Last-test VAIL		1	.147*	.589***	-.012	.178**
3. First-Completion			1	.270***	-.122*	.010
4. Last-Completion				1	-.137*	.012
5. Times Taken					1	-.064
6. GPA						1

*$p < .05$
**$p < .01$
***$p < .001$

Table 5
Bi-variate Correlations

		VAIL Effort		VAIL Total	
Predictors		ß	Std. Error	ß	Std. Error
Times Taken		-.168*	.865	-.006	2.020
GPA		-.019	.679	.172*	1.586
Teaching Area[a]					
Early Childhood		-.156*	2.074	-.104	4.844
Secondary		-.002	.422	-.071	.985
Special Education		-.067	.565	-.078	.985
	Final R	.237*	2.901	.218*	6.775
	Final ΔR^2	.056*	2.901	.047*	6.775

[a]Elementary is the comparison group.
*$p<.05$

Discussion

Many teacher education experts have called for improved teacher preparation instruments that can contribute to efforts to strengthen teacher evaluation (Worrell et al., 2014; Zeichner, 2005). To contribute to one teacher education program's evaluation efforts we examined the use of the Video Assessment of Interactions in Learning (Jamil, et al., 2015) over a five-year period. We found that teacher education students did not demonstrate improved performance on the VAIL from the beginning to the end of their program; however, we did find that participant effort towards the VAIL measure decreased from the beginning to the end of the program. Data indicate that repeatedly expecting teacher education students to take the same assessment may lead to measurement fatigue and lack of effort. Given the cost of implementing the VAIL and the results of overuse, future use of the assessment should be adjusted accordingly.

Worrell and colleagues (2014) call for valid and scientifically based assessments when evaluating teacher preparation programs. An important step in determining the validity of an assessment is understanding the data it provides in practice. The VAIL, which includes watching videos and responding to open-ended prompts, was required of teacher education students every year they were in the preparation program. The VAIL did not show differences

between participants' ability to identify effective teaching interactions at the beginning and end of the program. These might be attributable to the fact that participants demonstrated less effort at the end of their program than at the beginning. The only portion of the VAIL that did show a difference was the first video which also had the most consistent effort of participants at the beginning and end of their program. However, the VAIL does have the benefit of being a standardized measure that can be implemented at various points in the teacher education program (Wiens et al., 2013). It might be advisable to reduce the number of times participants are required to take the VAIL and see if they are more motivated to expend more effort at the end of their program.

There appears to be an element of assessment fatigue in our data, as seen in the reduced completion scores in the last-test Completion score compared to the first-test. Assessment fatigue is also supported by the regression analysis that showed a negative relationship between the number of times a participant took the VAIL and the effort he/she was willing to put into the final attempt. Even within the assessment the third and final video had the lowest completion score, and in the last-test the third video also had the lowest completion of any video from any time point. In this teacher education program the VAIL is a low-stakes assessment and it relies on preservice teachers' intrinsic motivation to do well. Since intrinsic motivation is related to personal enjoyment, interest, or pleasure (Lai, 2011), it might be difficult to motivate students to do their best work. While there is little empirical literature related to assessment fatigue in these situations, there is evidence that university students who are expected to complete multiple surveys may be unlikely to participate fully (Porter et. al., 2004) and this may be especially true in longitudinal surveys (Apodaca, Lea, & Edwards, 1998). In this sample, fatigue may be an issue due to low motivation and repeated administrations of the same measure; the more times teacher education students were asked to complete the VAIL the less effort they were willing to put into completing the measure.

Five years of data collection indicate that programs need to carefully consider the burden that their assessments place on participants. Assessing participants too often with the same measure may undermine the validity of the assessment if participants' effort decreases over time.

The VAIL has been shown to be a valid and reliable measure (Jamil et al., 2015), related to teaching performance with in-service (Hamre et al., 2012) and preservice teachers, and useful in teacher education contexts (Wiens et al., 2013). However, this study provides some important information for determining best practices for use of the VAIL as a teacher education program evaluation tool. The VAIL gives participants the opportunity to provide up to 30 different responses (five strategies plus five examples for each of the three videos). Future implementation of the VAIL should revisit the length or layout of the measure to make it a more valid estimate of preservice teachers' ability to identify effective interactions. Another option is to require teacher education students to only take the VAIL at the beginning and end of the program to determine if participant effort improves on the last-test. This would also have the benefit of requiring less resources from the teacher education program in hiring and training reliable coders. A third option to increase participant effort on the VAIL would be to experiment with making it a higher-stakes assessment. If participants were more motivated to do well on the assessment then they may increase their effort and improve their overall performance.

Conclusion

Systematic research on teacher education is a necessity for the field (Grossman & McDonald, 2008; Worrell et al., 2014; Zeichner, 2005). The development of valid measures (Worrell et al., 2014) that can address the needs of multiple constituents (Feuer et al., 2013) can help to move the field forward and provide robust program evaluations. One teacher education program used the VAIL (Jamil et al., 2015) as a component of its evaluation program. Five years of data collection indicate that programs need to carefully consider the burden that their assessments place on participants. Assessing participants too often with the same measure may undermine the validity of the assessment if participants' effort decreases over time. Continual examination of program assessments is required to ensure that teacher education programs are preparing future generations of quality teachers.

References

Apodaca, R., Lea, S., & Edwards, B. (1998). The effect of longitudinal burden on survey participation. In *American Statistical Association Proceedings of the Survey Research Methods Section* (pp. 906–10).

Bransford, J.D., Brown, A., & Cocking, R. (1999). *How people learn: Brain, mind, experience, and school*. Washington, DC: National Academy Press.

Bronfenbrenner, U. (1993). Ecological models of human development. In M. Gauvain & M. Cole (Eds.), *Readings on the development of children, 2nd ed*. (pp. 37–43). New York, NY: Freeman.

Brophy, J. (1983). Conceptualizing student motivation. *Educational Psychologist, 18*(3), 200–215.

Cadima, J., Leal, T, & Burchinal, M. (2010). The quality of teacher-student interactions: Associations with first graders' academic and behavioral outcomes. *Journal of School Psychology, 48*, 457–482. doi: 10.1016/j.jsp.2010.09.001

Carter, K., Cushing, K., Sabers, D., Stein, P., & Berliner, D. (1988). Expert-novice differences in perceiving and processing visual classroom information. *Journal of Teacher Education, 39*(3), 25–31.

Christ, T., Arya, P., & Chiu, M.M. (2017). Video use in teacher education: An international survey of practices. *Teaching and Teacher Education, 63*, 22–35. doi:10.1016/j.tate.2016.12.005

Council for the Accreditation of Educator Preparation. (2013). *2013 CAEP standards*. Washington, DC: Author.

Curby, T.W., Rimm-Kaufman, S.E., & Ponitz, C.C. (2009). Teacher-child interactions and children's achievement trajectories across kindergarten and first grade. *Journal of Educational Psychology, 101*(4), 912–925. doi: 10.1037/a0016647

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. New York: Plenum Press.

Dev, P.C. (1997). Intrinsic motivation and academic achievement: What does their relationship imply for the classroom teacher? *Remedial and Special Education, 18*(1), 12–19.

Feldon, D.F. (2007). The implications of research on expertise for curriculum and pedagogy. *Educational Psychology Review, 19*, 91–110. doi: 10.1007/s10648-006-9009-0.

Feuer, M.J., Floden, R.E., Chudowsky, N., & Ahn, J. (2013). *Evaluation of teacher preparation programs: Purposes, methods, and policy options*. Washington, DC: National Academy of Education.

Gaudin, C., & Chaliès, S. (2015). Video viewing in teacher education and professional development: A literature review. *Educational Research Review, 16*, 41–67. doi:10.1016/j.edurev.2015.06.001

Glaser, R. & Chi, M.T.H. (1988). Overview. In M.T.H. Chi, R. Glaser, & M.J. Farr (Eds.), *The nature of expertise* (pp. xv-xxviii). Mahwah, NJ: Lawrence Erlbaum Associates.

Goodwin, C. (1994). Professional vision. *American Anthropologist, 96*(3), 606–633.

Graue, E., Raushcer, E., & Sherfinski, M. (2009). The synergy of class size reduction and classroom quality. *The Elementary School Journal, 110*(2), 178–201. doi: 10.1086/605772.

Grossman, P., & McDonald, M. (2008). Back to the future: Directions for research in teaching and teacher education. *American Educational Research Journal, 45*(1), 184–205. doi: 10.3102/0002831207312906

Jamil, F.M., Sabol, T.J., Hamre, B.K., & Pianta, R.C. (2015). Assessing teachers' skills in detecting and identifying effective interactions in the classroom: Theory and measurement. *The Elementary School Journal, 115*(3), 407–432.DOI: https://doi.org/10.1086/680353

Kaiser, G., Busse, A., Hoth, J., König, J., & Blömeke, S. (2015). About the complexities of video-based assessments: Theoretical and methodological approaches to overcoming shortcomings of research on teachers' competence. *International Journal of Science and Math Education, 13*, 369–387. doi: 10.1007/s10763-015-9616-7

Kersting, N. B., Givvin, K. B., Sotelo, F. L., & Stigler, J. W. (2010). Teachers' analyses of classroom video predict student learning of mathematics: Further explorations of a novel measure of teacher knowledge. *Journal of Teacher Education, 61*(1–2), 172–181. doi:10.1177/0022487109347875

Lai, E. R. (2011). *Motivation: A literature review*. Boston, MA: Pearson.

Pianta, R.C., La Paro, & Hamre, B.K. (2008). Classroom Assessment Scoring System (CLASS). Baltimore, MD: Paul H. Brookes.

Pianta, R.C., Belsky, J., Vandergrift, N., Houts, R., & Morrison, F.J. (2008). Classroom effects on children's achievement trajectories in elementary school. *American Educational Research Journal, 45*, 365–397. doi: 10.3102/0002831207308230

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher, 38*(2), pp. 109–119. doi: 10.3102/0013189X09332374

Porter, S. R., Whitcomb, M.E., & Weitzer, W.H. (2004). Multiple surveys of students and survey fatigue. *New Directions for Institutional Research, 121*, 63–73.

Reyes, M. R., Brackett, M. A., Rivers, S. E., White, M., & Salovey, P. (2012). Classroom emotional climate, student engagement, and academic achievement. *Journal of Educational Psychology, 104*(3), 700–712. doi: 10.1037/a0027268

Santagata, R. & Yeh, C. (2014). Learning to teach mathematics and to analyze teaching effectiveness: Evidence from a video- and practice-based approach. *Journal of Mathematics Teacher Education, 17*, 491–514. doi: 10.1007/s10857-013-9263-2

Sherin, M.G. & van Es, E.A. (2005). Using video to support teachers' ability to notice classroom interactions. *Journal of Technology and Teacher Education, 13*(3), 475–491.

Star, J.R. & Strickland, S.K. (2008). Learning to observe: using video to improve preservice mathematics teachers' ability to notice. *Journal of Mathematics Teacher Education 11*(2), 107–125. doi: 10.1007/s10857-007-9063-7

Stürmer, K., Seidel, T., & Schäfer, S. (2013). Changes in professional vision in the context of practice. *Gruppendynamik und Organisationsberatung, 44*(3), 339–355. doi: 10.1007/s11612-013-0216-0

Van Es, E. A., & Sherin, M. G. (2002). Learning to notice: Scaffolding new teachers' interpretations of classroom interactions. *Journal of Technology and Teacher Education, 10*(4), 571–596.

Video Assessment of Interactions and Learning: VAIL (2010). Coding manual (Unpublished Manual). Charlottesville, VA: Center for the Advanced Study of Teaching and Learning.

Wiens, P.D., Hessberg, K., LoCasale-Crouch, J., & DeCoster, J. (2013). Using a standardized video-based assessment in a university teacher education program to examine preservice teachers' knowledge related to effective teaching. *Teaching and Teacher Education, 33*, 24–33.

Wiens, P.D. (2014). Using a participant pool to gather data in a teacher education program: The course of one school's efforts. *Issues in Teacher Education, 23*(1), 177–206.

Woolfolk, A. E. (1990). *Educational psychology* (4th ed.). Boston: Allyn & Bacon.

Worrell, F., Brabeck, M., Dwyer, C., Geisinger, K., Marx, R., Noell, G., & Pianta R. (2014). *Assessing and evaluating teacher preparation programs*. Washington, DC: American Psychological Association

Zeichner, K. M. (2005). A research agenda for teacher education. In M. Cochran-Smith & K. M. Zeichner (Eds.), *Studying teacher education: The report of the AERA Panel on Research and Teacher Education* (pp. 737–759). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers; Washington, DC, US: American Educational Research Association.

Abstract

This study utilized generalizability theory to assess the context where the National Survey of Student Engagement's (NSSE) summary measures, the Engagement Indicators, produce dependable group-level means. The dependability of NSSE group means is an important topic for the higher education assessment community given its wide utilization and usage in institutional assessment and accreditation. We found that the Engagement Indicators produced dependable group means for an institution derived from samples as small as 25 to 50 students. Furthermore, we discuss how the assessment community should use NSSE data.

AUTHORS

Kevin Fosnacht, Ph.D.
Indiana University, Bloomington

Robert M. Gonyea, Ed.D.
Indiana University, Bloomington

The Dependability of the Updated NSSE: A Generalizability Study

Despite decades of dialogue, higher education still struggles with assessing the quality of undergraduate education and no longer enjoys respectful deference from governments, media, and the public who are collectively anxious about cost and quality. Such anxieties have stimulated considerable pressure for assessment and accountability. The dominant paradigm focusing on resources and reputation – most visible in the *U.S. News and World Report* rankings – has been roundly criticized for its neglect of students' educational experiences (Carey, 2006; McGuire, 1995). In response, higher education leaders, researchers, and assessment professionals have explored many ways for higher education to improve – through reforming the curriculum, faculty development, and improved assessment (Association of American Colleges and Universities, n.d.; Barr & Tagg, 1995; Gaston, 2010; Lumina Foundation, 2011). In recent years, the measurement of student engagement has emerged as a viable alternative for institutional assessment, accountability, and improvement efforts. Student engagement represents collegiate quality in two critical ways. The first is the amount of time and effort students put into their coursework and other learning activities, and the second is how the institution allocates resources, develops the curriculum, and promotes enriching educational activities that decades of research studies show promote student learning. (Kuh, 2003, 2009; Kuh, Hayek, Carini, Ouimet, Gonyea, & Kennedy, 2001; McCormick, Kinzie, & Gonyea, 2013).

CORRESPONDENCE

Email
kfosnach@indiana.edu

The National Survey of Student Engagement (NSSE) collects information at hundreds of bachelor's-granting universities to estimate how students spend their time and how their educational experiences are shaped. Institutions use NSSE primarily in two ways. The first is to compare, or benchmark, their students' responses with those of students at other institutions. Such an approach provides the institution with diagnostic information about how their students are learning, and which aspects of the undergraduate experience have been effective and which are in need of improvement. The second way institutions use NSSE is to assess subgroups of their students to determine how student

engagement varies *within* the institution and to uncover areas for institutional improvement for groups such as first-generation students, part-time students, adult and commuter students, students enrolled in different majors, transfer students, and so on. Both of these approaches utilize NSSE scores by comparing the aggregate score of one group with that of another group, whether they be different institutions or different types of students within the same institution.

Thus, the NSSE instrument depends foremost on its reliability at the group level, and upon its ability to generalize an outcome to the aggregated group. The reliability of a group mean score requires methodological techniques that can account for and identify multiple sources of error (Pike, 1994, 2013). Consequently, this paper explores the notion that generalizability theory (GT) may provide the proper methodological framework to assess the dependability of benchmarking instruments such as NSSE, and uses GT to investigate the number of students needed to produce a reliable group mean for the NSSE Engagement Indicators. Finally, the appropriate uses of NSSE data are discussed in light of the study's findings.

In recent years, the measurement of student engagement has emerged as a viable alternative for institutional assessment, accountability, and improvement efforts.

Updating NSSE

Since NSSE's initial launch in 2000, higher education scholars have learned more about collegiate activities and practices that positively influence student outcomes (Kuh, Kinzie, Cruce, Shoup & Gonyea, 2006; McClenney & Marti, 2006; Pascarella, Seifert, & Blaich, 2010; Pascarella & Terenzini, 2005). Many areas of higher education are seeing growth, innovation, and rapid adoption of new ideas such as distance learning and other technological advances. To meet these challenges and improve the utility and actionability of its instrument, NSSE introduced an updated version in 2013, which both refines its existing measures and incorporates new measures related to emerging practices in higher education (NSSE, 2018b). The new content includes items investigating quantitative reasoning, interactions among diverse populations, learning strategies, and teaching practices. Additionally, the update provides the opportunity to improve the clarity and consistency of the survey's language and to improve the properties of the measures derived from the survey. Despite these changes, the updated instrument is consistent with the purpose and design of the original version of NSSE (Kuh et al., 2001), as it continues to focus on whether institutions emphasize participation in effective educational practices, and is administered to samples of first-year students and seniors at various types of baccalaureate-granting institutions.

Validity of NSSE

With the updated survey, NSSE continues its core purpose of providing institutions with valid and reliable assessment information for the improvement of the educational experience such as helping faculty and senior academic leaders to shape faculty development programs, revise curricula, or develop student support programs. Studies that link student engagement to university outcomes such as critical thinking, moral development, and leadership capacity, or to other indicators of success such as grades, persistence, and graduation, give credence to NSSE's validity and support such valid uses of the data.

For example, research has found positive associations with persistence (Hughes & Pace, 2003; Kuh, 2008; Kuh et al., 2008; McClenney & Marti, 2006), critical thinking (Loes, Pascarella, & Umbach, 2012), GRE scores (Carini, Kuh, & Klein, 2006), moral reasoning (Mayhew et al., 2012), and need for cognition (Padgett et al., 2010). Using institution-level data, NSSE benchmarks had at least one significant positive association with institution-level outcome scores (effective reasoning and problem-solving, moral character, inclination to inquire and lifelong learning, intercultural effectiveness, and personal well-being) for first-year students after controlling for pre-test outcome scores (Pascarella, Seifert, & Blaich, 2010).

Prior research has supported the use of self-reported data on university students (see Pace, 1985 and Pike, 2011), although some (e.g., Porter, 2011) have raised questions about the validity of university student surveys. Cited concerns included a lack of a sufficient

theoretical basis for survey content, difficulties in the response process, the lack of a factor structure and adequate reliability for NSSE's benchmarks, and poor relationships between measures of student engagement and direct observations of the same behavior. In response, NSSE researchers explain that while the student engagement survey items are supported in the literature, the survey was created for institutional assessment, not for theory building or testing of a narrow theoretical construct. Also, students' ability to respond to the survey items has been established by extensive testing with hundreds of students at dozens of institutions using focus groups and cognitive interviews. For a more comprehensive discussion of NSSE's validity, see McCormick and colleagues (2013) and NSSE's (2018c) psychometric portfolio.

Generalizability Theory

Generalizability Theory, first detailed in a monograph by Cronbach, Gleser, Nanda, and Rajaratnam (1972), is a conceptual framework useful in determining the reliability and dependability of measurements. Unlike reliability coefficients such as Cronbach's α that provide a single statistic, GT provides a framework for determining the situations where drawing inferences from their samples would be appropriate. Researchers and assessment professionals can then use this information to design a study, or inferences can be responsibly inferred from their existing data. GT is perhaps best described in relation to classical test theory (CTT) where a person's true score (T) on an item or test is composed of their observed score (X) and measurement error (e): $T=X+e$. Thus, CTT focuses on determining the error of a measurement. In contrast, GT recognizes that multiple sources of error may exist and examines their magnitude rather than focusing on a single overall error score. These potential sources of error (e.g., individuals, raters, items, and occasions) are referred to as facets. More concretely, an error could be due to a student randomly guessing the correct answer on a test, differences in the calibration of a scale, or the implicit biases of a judge or rater. The theory assumes that any observation or data point is drawn from a universe of possible observations. For example, an item on a survey is assumed to be sampled from a universe of comparable items, just as individuals are sampled from a larger population. Consequently, the notion of reliability in CTT is replaced by the question of the "accuracy of generalization or generalizability" to a larger universe (Cronbach et al., 1972, p. 15).

As a methodological theory, GT is intimately associated with its methods. Generalizability Theory utilizes analysis of variance (ANOVA) which analyzes the amount of variation in a measure attributable to groups of people, test items, schools, or other things of interest to a researcher. In the GT context, ANOVA is used to estimate the magnitude of the variance components associated with the types of error identified by the researcher. However, it is important to note that while GT uses ANOVA, it departs from the traditional uses of ANOVA through its focuses on variance components, not testing statistical significance. The researcher subsequently uses the variance components to calculate the generalizability coefficient, which is analogous to the reliability coefficient in CTT. The generalizability coefficient is a type of intraclass correlation coefficient (which measures the proportion of total variance attributable to within-group differences). However, in the generalizability coefficient the true score variance of CTT is replaced with the universe score variance focused on in GT (Kane & Brennan, 1977). GT also distinguishes between a generalizability (G) study and a decision (D) study. The G-study uses ANOVA to estimate the variance components used to calculate the generalizability coefficient. The components can also be used in a D-study to estimate the generalizability coefficient in different contexts. The D-study allows a researcher to efficiently optimize a study or to determine the conditions under which a score is generalizable.

Thus, CTT focuses on determining the error of a measurement. In contrast, GT recognizes that multiple sources of error may exist and examines their magnitude rather than focusing on a single overall error score.

Due to the focus on groups in educational assessment, GT makes important contributions to determining the validity of surveys, such as NSSE. The flexibility of GT allows researchers to determine the conditions under which group means will be accurate and dependable. This is in contrast to the methods based on CTT that look at the internal consistency of a set of items (e.g., Cronbach's α), but fail to identify the conditions under which a measure is accurate. This weakness of CTT approaches may lead well-intentioned researchers to use a measure under conditions where its validity is questionable. Despite the benefits of GT, it has been underutilized in higher education research even after Pike's (1994) work that introduced GT and its methods to the field.

Research Questions

Guided by GT, the study answered the following questions:

1. How dependable are the NSSE Engagement Indicators?
2. How many students are needed to produce a dependable group mean for the NSSE Engagement Indicators?

Methods

Data

The study utilized data from the 2013 NSSE administration. The survey was administered to 334,808 first-year students and seniors at 568 baccalaureate-granting institutions in the United States. in the winter and spring of 2013. The characteristics of the institutions and respondents are available from NSSE (2013). The characteristics of the institutions roughly mirror the U.S. landscape, although public institutions and larger master's colleges and universities were overrepresented. Baccalaureate Colleges – Diverse Fields were slightly underrepresented in the dataset. Approximately, two out of three respondents were female, the same proportion was White, and the vast majority enrolled as full-time students. The average institutional response rate was 30%, which prior research using NSSE data has shown to produce estimates that can be generalized to the broader population of students within an institution accurately (Fosnacht, Sarraf, Howe, & Peck, 2017).

Due to the focus on groups in educational assessment, GT makes important contributions to determining the validity of surveys, such as NSSE. The flexibility of GT allows researchers to determine the conditions under which group means will be accurate and dependable.

The measures used in the study were the survey items that comprised the NSSE Engagement Indicators (EI), groups of related items designed by NSSE researchers to measure the extent to which an institution's environment promotes effective educational practices. The ten indicators are Higher-Order Learning (HO; 4 items), Reflective & Integrative Learning (RI; 7 items), Learning Strategies (LS; 3 items), Quantitative Reasoning (QR; 3 items), Collaborative Learning (CL; 4 items), Discussions with Diverse Others (DD; 4 items), Student-Faculty Interaction (SF; 4 items), Effective Teaching Practices (ET; 5 items), Quality of Interactions (QI; 5 items), and Supportive Environment (SE; 8 items). The full list of the items that comprise the EIs is available from NSSE (2018a), and the abbreviations used match those used by NSSE in its reporting to institutions. The QI items had a "not applicable" option which was recoded to missing for this analysis. All items within each indicator shared the same response set and were not recoded (except for the QI items).

Analyses

Guided by GT, the study examined the group mean generalizability of the NSSE Engagement Indicators at the institution level. We performed the following procedures to assess the generalizability of each EIs by class. We identified two facets, students and items, as potential sources of error for the indicators. Additionally, students in our sample were nested within institutions due to the design of NSSE. Thus, the G-study portion of our analyses which estimated the variance components utilized a split-plot, random effects ANOVA design, where students were nested within institutions and crossed with survey items (see Kirk, 2013 for more details on split-plot ANOVA designs). In this design, each institution has a different set of students, but all students answered the same items. The design was also balanced, with 50 students randomly selected from each institution. The value of 50 was selected to maximize the number of students and institutions included in the study after the exclusion of cases with missing data. The mathematical model of the ANOVA was:

$$X_{usi} = \mu + \alpha_u + \pi_{s(u)} + \beta_i + \alpha\beta_{ui} + \beta\pi_{is(u)} + e_{usi}$$

Where,

X_{usi} = Response by student *s* in institution u on item *i*

μ = grand mean

α_u = effect for institution *u*

$\pi_{s(u)}$ = effect for student s nested within institution *u*

β_i = effect for item *i*

$\alpha_u\beta_i$ = institution by item interaction

$\beta_i\pi_{s(u)}$ = item by student, nested within institution, interaction, and

e_{usi} = error term.

Apart from the grand mean, each of the parameter estimates varies by institution, student, and/or item. This variation allows for the estimation of the variance components which decompose the total model variation into portions attributable to each effect. We used the G1 program for SPSS to analyze the data and estimate the variance components (Mushquash & O'Connor, 2006).

After calculating the variance components in the G-study, we performed D-studies for each EI. A D-study allows a researcher to estimate how a generalizability coefficient would change if the study parameters changed for example by changing the number of students participating in a study or changing the number of items in a factor. We estimated the generalizability coefficients over sample sizes of 25, 50, 75, and 100 students within an institution. By varying the number of students in the D-studies, the results allow us to investigate the dependability of the EIs and describe situations where the use of a group mean is and is not appropriate. We did not calculate generalizability coefficients using different numbers of items as the design of the core NSSE instrument is static.

The study found that the means of the NSSE Engagement Indicators can be reliably generalized to a larger population from small samples of students at postsecondary institutions.

In the D-studies, we calculated two generalizability coefficients using formulas outlined by Kane, Gillmore, and Crooks (1976). We choose to follow Pike's (2006; 2013) approach by calculating both coefficients to obtain more knowledge on the circumstances where using the NSSE Engagement Indicators would be appropriate (see the discussion section for an interpretation of their appropriate uses). The first coefficient generalized over both facets – students and items – and can be interpreted as the expected correlation of the group means derived from two samples of students at the same institution, who answered separate, but comparable items. This generalizability coefficient should be used if a set of items is believed to represent a higher-order construct or a factor. This correlation could also be produced by developing a number of survey items, giving half of the items to half of the students at each institution and correlating the mean to the mean of the other half of items given to the remaining students. The formula used to calculate the coefficient that generalized over students (*S*) and items (*I*) was:

$$\varepsilon\rho^2(S,I) = \frac{\sigma^2(u)}{\sigma^2(u)+\frac{1}{k}\sigma^2(ui)+\frac{1}{n}\sigma^2(s,ui)+\frac{1}{n*k}\sigma^2(e)}$$

where,

$\sigma^2(u)$=variance component associated with the institution from the G-study

$\sigma^2(ui)$=variance component associated with the institution by item interaction from the G-study

$\sigma^2(s,ui)$=variance component associated with students nested within institutions crossed with items from the G-study

$\sigma^2(e)$=variance component associated with the error term from the G study

k=number of items in the factor

n=number of students per institution.

The second coefficient generalized only over students by treating the survey items as fixed, rather than random, effects. This coefficient can be interpreted as the expected correlation of the aggregated means of two samples of students who answered the same items. This formula should be used when a conclusion is to be drawn about a set of items, but not a higher-order construct. An analogous method to produce this correlation is to correlate the group means of two samples of students at each institution answering the same items. The formula used to calculate the coefficient that generalized over students was:

$$\varepsilon\rho^2(S) = \frac{\sigma^2(u)+\frac{1}{k}\sigma^2(ui)}{\sigma^2(u)+\frac{1}{k}\sigma^2(ui)+\frac{1}{n}\sigma^2(s,ui)+\frac{1}{n*k}\sigma^2(e)}$$

where the variables and variance components are the same as in equation 2.

Limitations

The primary limitation of the study is that it used the institution as the object of measurement. As there is more variability within than between institutions (NSSE, 2008), the results may exhibit a non-trivial difference if the object of measurement utilized was major field or a demographic characteristic. For example, the dependability of QR may be higher when the object of measurement is the group mean of a major field as this measure varies more between majors than it does between institutions (Rocconi, Lambert, McCormick, & Sarraf, 2013). Additionally, the QI item set included a "not applicable" response option that we recoded to missing for this analysis. As "not applicable" is not an ordered response, we were unable to include this response type in our analyses, but excluding students who answered this response could potentially bias the results. Thus, the QI results should be interpreted with caution. We must also note that the generalizability coefficient discussed in this study differs from Brennan and Kane's (1978) *Index of Dependability*. While the Index of Dependability utilizes GT, the index is designed for mastery tests (of which NSSE is not) and focuses on decisions regarding a cut score such as an admission requirement to score at least 1,000 on the SAT.

Thus, the NSSE EIs can efficiently discriminate institutional environments that promote engagement in effective educational practices. In other words, using a relatively small sample of respondents, the EIs can identify institutions with high and low levels of engagement.

Results

Figure 1 demonstrates the utility of a G- and D-study as it plots both generalizability coefficients by the number of students included in a group mean for the CL engagement indicator for first-year students. The grey dotted line contains the generalizability coefficient of the group mean for various numbers of students when generalizing over students and items. The black line does the same when generalizing over just students. In both lines, there is a steep change in the lines' slope until an N of roughly 20 students is reached. From about 20 to 60 students, there is a moderate increasing slope, which then flattens for larger numbers of students. Using $E\rho^2 \geq .70$ as a threshold for dependability, at least ten students are required to produce a dependable group mean when generalizing over just students for CL. In contrast, approximately 20 students are required to meet the same threshold when generalizing over students and items. Thus, depending upon the intended purposes of a study, 10 to 20 randomly-selected students would be required to create a dependable group mean that can be generalized to the larger first-year student body for CL. Group means containing fewer students should be viewed as less generalizable and used with caution by researchers or assessment professionals.

Table 1 contains the generalizability coefficients when generalizing over students and items by class and the four sample sizes investigated in the D-studies (the variance components from the G-study are available in Appendix). For first-year students, the CL, DD, SE, and SF EIs met accepted standards for dependability ($E\rho^2 \geq .70$) when a group mean was derived from a sample of 25 to 75 students. The other EIs required substantially larger samples to meet the

Figure 1
D-study generalizability coefficients for first-year students and collaborative learning by number of students

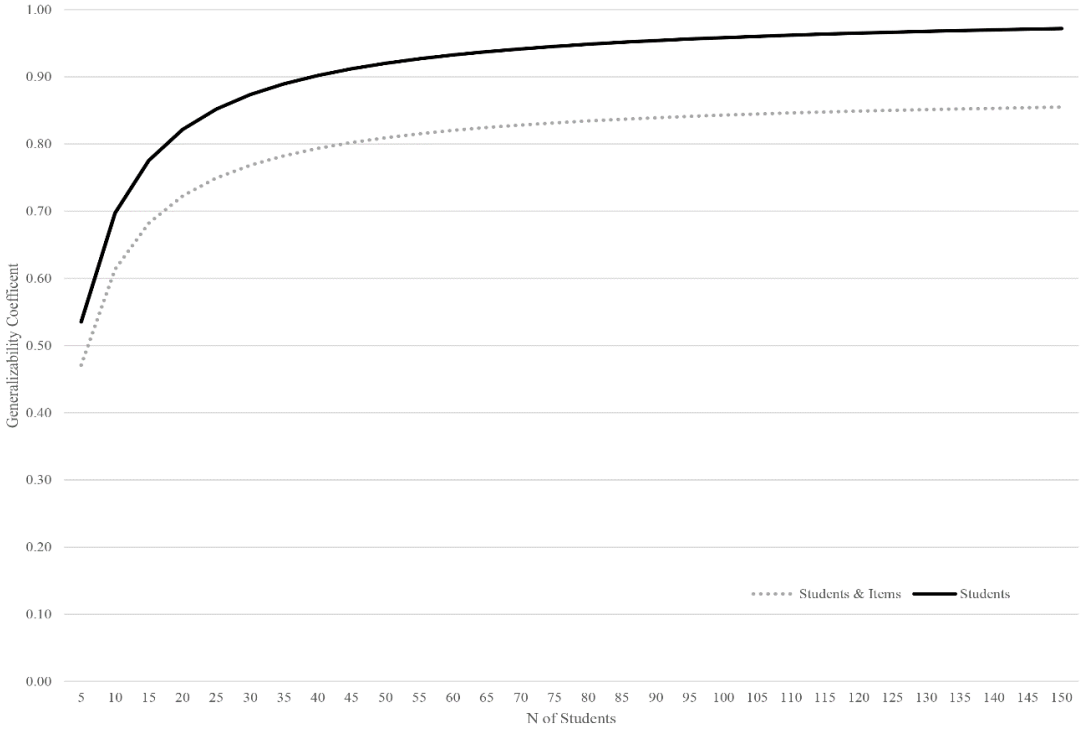


Table 1
D-study generalizability coefficients over students and items by class and sample size

	First-Year				Senior			
Engagement Indicator	25	50	75	100	25	50	75	100
Academic Challenge								
Higher-Order Learning	.48	.56	.59	.61	.46	.54	.57	.58
Reflective & Integrative Learning	.56	.62	.64	.65	.60	.65	.67	.68
Quantitative Reasoning	.39	.48	.52	.55	.46	.55	.59	.61
Learning Strategies	.47	.58	.63	.65	.60	.70	.74	.76
Learning with Peers								
Collaborative Learning	.75	.81	.83	.84	.75	.81	.83	.84
Discussions with Diverse Others	.62	.69	.72	.73	.62	.69	.71	.72
Experiences with Faculty								
Student-Faculty Interaction	.65	.72	.75	.76	.78	.83	.85	.86
Effective Teaching Practices	.47	.55	.58	.59	.50	.57	.60	.61
Campus Environment								
Quality of Interactions	.52	.60	.63	.65	.58	.67	.70	.72
Supportive Environment	.68	.72	.74	.75	.71	.76	.77	.78

same threshold. For seniors, the same EIs had generalizability coefficients greater than .70 using sample sizes of 25 to 50 students. The LS and QI EIs were dependable when a group mean contained at least 75 students. Coefficients for the remaining EIs were below the .70 standard using sample sizes of less than 100 when generalizing over students and items.

The generalizability coefficients when generalizing only over students are located in Table 2. In contrast to the coefficients over both students and items, nearly all of the EIs met standards for dependability using samples as low as 25 students. The exceptions were QR for both classes and LS for first-year students. All the generalizability coefficients were higher than .80 when group means contained 50 seniors, and all were greater than .80 when group means contained 75 first-year students.

Table 2
D-study generalizability coefficients over students but not by items by class and sample size

	First-Year				Senior			
	25	50	75	100	25	50	75	100
Academic Challenge								
Higher-Order Learning	.72	.84	.89	.91	.72	.84	.89	.91
Reflective & Integrative Learning	.81	.90	.93	.95	.85	.92	.94	.96
Quantitative Reasoning	.63	.77	.84	.87	.67	.80	.86	.89
Learning Strategies	.63	.77	.84	.87	.71	.83	.88	.91
Learning with Peers								
Collaborative Learning	.85	.92	.95	.96	.85	.92	.94	.96
Discussions with Diverse Others	.81	.89	.93	.94	.81	.90	.93	.95
Experiences with Faculty								
Student-Faculty Interaction	.80	.89	.92	.94	.88	.93	.96	.97
Effective Teaching Practices	.73	.84	.89	.92	.75	.86	.90	.92
Campus Environment								
Quality of Interactions	.74	.85	.90	.92	.74	.85	.90	.92
Supportive Environment	.87	.93	.95	.96	.89	.94	.96	.97

Discussion

The study found that the means of the NSSE Engagement Indicators can be reliably generalized to a larger population from small samples of students at postsecondary institutions. Therefore, the EIs appear to be dependable measurements of undergraduates' engagement in beneficial activities at an institution during university. Eight of the ten indicators had generalizability coefficients above .70 for both first-year students and seniors, when an institution's mean was derived from just 25 students. All EIs had generalizability coefficients in excess of .70 when the sample size increased to 50 students. Thus, the NSSE EIs can efficiently discriminate institutional environments that promote engagement in effective educational practices. In other words, using a relatively small sample of respondents, the EIs can identify institutions with high and low levels of engagement.

Future research should examine the generalizability of NSSE for subgroups (e.g., racial/ethnic groups, major fields, program participants), which will allow users of NSSE data to improve and target their educational offerings.

However, the results revealed that only some of the indicators could be dependably generalized to a higher-order construct. The CL, DD SF, and SE EIs appear to be dependable group-level measures when generalizing over students and items and using sample sizes of 25 to 75 students. LS and QI also appear to be dependable for seniors using a sample size of at least 75 students. However, the remaining EIs do not appear to produce dependable group means representing a higher-order construct, except when the sample contains hundreds of students. Therefore, when the object of measurement is an institution, the indicators with lower levels of dependability when generalizing over students and items would be most reliably treated as indexes (groups of items that, when combined, indicate a more general characteristic) rather than higher-order constructs. The lower level of dependability in these indicators, when generalizing over students and items, is generally caused by the small amount of variability accounted for by the institutional effects, which limits the ability to discriminate between institutional means.

It is not surprising that some of the indicators have poor dependability as higher-order constructs. NSSE was designed to estimate to undergraduates' engagement in effective educational practices "known to be related to important [university] outcomes" (Kuh et al., 2001, p. 3). Therefore, the Engagement Indicators do not contain items randomly selected from a domain of all possible questions related to a higher-order construct, but rather function as an index or snapshot of the level of engagement in specific beneficial activities known to improve university outcomes. While this study examined the generalizability of the Engagement Indicators over students and items, the purpose of and methods used to construct the survey suggest that this is not the appropriate criterion to assess the dependability of the NSSE Engagement Indicators. Instead, the more appropriate measure is the generalizability coefficient when generalizing over students, but not items.

NSSE briefly examines multiple forms of student engagement to increase its utility for institutions and to ensure a reasonable survey length for respondents. The downside of this approach is that NSSE is unable to ask a detailed set of questions about each type of student engagement. As the accuracy of a student's score is a function of a measurement's reliability (Wainer & Thissen, 1996) and the reliability is related to the number of items in a measurement (Brown, 1910; Spearman, 1910), the relatively small number of items in each Engagement Indicator suggests that an individual's score is associated with a nontrivial amount of error. However, NSSE overcomes this limitation by shifting the object of measurement from an individual student to the group level and aggregating the EIs into group means. Aggregation naturally increases the number of items in a measurement, which results in a higher degree of reliability for institution-level results.

The generalizability or the related concept of reliability in classical test theory does not alone indicate that a measure is valid. Validity is a multifaceted topic that includes construct validity (which this study focuses on), relevance, value implications, and social consequences (Messick, 1989, 1995). Thus, generalizability alone does not indicate that a measure is valid. We encourage readers to also review NSSE's (2018c, 2018d) psychometric portfolio and conceptual framework before concluding that the NSSE Engagement Indicators are accurate measures of student engagement.

Due to the relatively small number of students needed to produce a dependable group mean, the NSSE Engagement Indicators provide the opportunity for assessment professionals to investigate the level of student engagement in a variety of subpopulations.

The vast majority of variation in NSSE data occurs within institutions (NSSE, 2008). In other words, students vary considerably more than institutions. Research and assessment professionals can exploit this variation to examine how a program or academic unit with a high graduation rate impacts students. For example, by comparing the NSSE Engagement Indicators between participants and non-participants in a learning community with a high graduation rate, an institutional researcher may discover that the participants have more academic interactions with their peers and perceive a more supportive campus environment. Administrators may use this finding to justify expanding the program or to implement a portion of the program for all students. Similarly, enrollment in a major may be low because the faculty has poor pedagogical practices that can be improved upon through workshops or another type of intervention. These hypothetical examples illustrate how NSSE data can be used by institutions to identify areas of strength and weakness. After identifying these areas, institutions can intervene to improve areas of weakness and encourage other programs or academic units to adopt the practices of successful programs.

Future research should examine the generalizability of NSSE for subgroups (e.g., racial/ethnic groups, major fields, program participants), which will allow users of NSSE data to improve and target their educational offerings. Alternately, research could examine the generalizability of results of specific types of institutions like publicly controlled colleges and universities or Jesuit colleges. Researchers should also examine the relationship between the NSSE Engagement Indicators and important outcomes like institutional retention rates, completion rates, and student loan default.

In summary, the means of the NSSE Engagement Indicators can be dependably and accurately generalized to a broader population of students when derived from a relatively small sample of undergraduates. The number of students required to produce a dependable group

mean varies by Engagement Indicator; however, a sample of 25 or 50 students is typically sufficient. Due to the relatively small number of students needed to produce a dependable group mean, the NSSE Engagement Indicators provide the opportunity for assessment professionals to investigate the level of student engagement in a variety of subpopulations. Finally, researchers should keep in mind that NSSE is intended to be used as a group-level instrument and was not designed to predict the outcome of an individual student.

References

Association of American Colleges and Universities. (n.d.). *An introduction to LEAP*. Retrieved from https://www.aacu.org/sites/default/files/files/LEAP/Introduction_to_LEAP.pdf.

Barr, R. B., & Tagg, J. (1995). From teaching to learning: A new paradigm for undergraduate education *Change, 27* (November/December), 13-25.

Brennan, R. L., & Kane, M. T. (1977). An index of dependability for mastery tests. *Journal of Educational Measurement, 14*(3), 277-289.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3*(3), 296-322.

Carey, K. (2006). College rankings reformed: The case for a new order in higher education. Washington, DC: Education Sector.

Carini, R. M., Kuh, G. D., & Klein, S. P. (2006). Student engagement and student learning: Testing the linkages. *Research in Higher Education, 47*(1), 1–32.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

Gaston, P. L. (2010). *General education & liberal learning: Principles of effective practice*. Washington, D.C.: Association of American Colleges and Universities.

Fosnacht, K., Sarraf, S., Howe, E., & Peck, L. K. (2017). How important are high response rates for college surveys?. *The Review of Higher Education, 40*(2), 245-265.

Hughes, R., & Pace, C. R. (2003). Using NSSE to study student retention and withdrawal. *Assessment Update, 15*(4), 1–2, 15.

Kane, M. T., & Brennan, R. L. (1977). The generalizability of class means. *Review of Educational Research, 47*(2), 267-292.

Kane, M. T., Gillmore, G. M., & Crooks, T. J. (1976). Student evaluations of teaching: The generalizability of class means. *Journal of Educational Measurement, 13*(3), 171-183.

Kirk, R. E. (2013). *Experimental Design: Procedures for the Behavioral Sciences* (4th ed.). Thousand Oaks, CA: Sage.

Kuh, G. D. (2003). What we're learning about student engagement from NSSE: Benchmarks for effective educational practices. *Change: The Magazine of Higher Learning, 35*(2), 24-32.

Kuh, G. D. (2008). *High-Impact educational practices: What they are, who has access to them, and why they matter*. Washington, DC: Association of American Colleges and Universities.

Kuh, G. D. (2009). The national survey of student engagement: Conceptual and empirical foundations. *New Directions for Institutional Research, 141*, 5-20.

Kuh, G. D., Hayek, J. C., Carini, R. M., Ouimet, J. A., Gonyea, R. M., & Kennedy, J. (2001). *NSSE technical and norms report*. Bloomington, IN: Indiana University, Center for Postsecondary Research and Planning.

Kuh, G. D., Kinzie, J., Cruce, T., Shoup, R., & Gonyea, R. M. (2006). *Connecting the dots: Multi-faceted analyses of the relationships between student engagement results from the NSSE, and the institutional practices and conditions that foster student success: Final report prepared for Lumina Foundation for Education*. Bloomington, IN: Indiana University, Center for Postsecondary Research.

Loes, C., Pascarella, E., & Umbach, P. (2012). Effects of diversity experiences on critical thinking skills: Who benefits? *Journal of Higher Education, 83*(1), 1–25.

Lumina Foundation. (2011). *The Degree Qualifications Profile*. Indianapolis, IN: Author.

Mayhew, M. J., Seifert, T. A., Pascarella, E. T., Nelson Laird, T. F., & Blaich, C. (2012). Going deep into mechanisms for moral reasoning growth: How deep learning approaches affect moral reasoning development for first-year students. *Research in Higher Education, 53*, 26–46.

McClenney, K. M., & Marti, C. N. (2006). *Exploring relationships between student engagement and student outcomes in community colleges: Report on validation research*. Retrieved from http://www.ccsse.org/center/resources/docs/publications/CCSSE Working Paper on Validation Research December 02006.pdf

McCormick, A. C., Kinzie, J., & Gonyea, R. M. (2013). Student engagement: Bridging research and practice to improve the quality of undergraduate education. In M. B. Paulsen (Ed.) *Higher education: Handbook of theory and research* (pp. 47-92). Dordrecht, NL: Springer.

McGuire, M. D. (1995). Validity issues for reputational studies. *New Directions for Institutional Research, 88*, 45-59.

Messick, S. (1989). Validity. In R. L. Linn (Ed.) *Educational Measurement* (3rd ed., pp. 13–103). New York: American Council on Education and Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749.

Mushquash, C., & O'Connor, B. A. P. (2006). SPSS and SAS programs for generalizability theory analyses. *Behavior Research Methods, 38*(3), 542-547.

National Survey of Student Engagement. (2008). Promoting engagement for all students: The imperative to look within - 2008 results. Bloomington, IN: Indiana University Center for Postsecondary Research.

National Survey of Student Engagement. (2013). *NSSE 2013 overview*. Retrieved from http://nsse.indiana.edu/2013_Institutional_Report/pdf/NSSE_2013_Overview.pdf

National Survey of Student Engagement. (2018a). *Engagement indicators*. Retrieved from http://nsse.indiana.edu/html/engagement_indicators.cfm

National Survey of Student Engagement. (2018b). *Information about the 2013 update*. Retrieved from http://nsse.indiana.edu/nsse-update/

National Survey of Student Engagement. (2018c). *NSSE's commitment to data quality*. Retrieved from http://nsse.indiana.edu/html/psychometric_portfolio.cfm

National Survey of Student Engagement. (2018d). *NSSE's conceptual framework (2013)*. Retrieved from http://nsse.indiana.edu/html/conceptual_framework_2.cfm

Pace, C. R. (1985). *The credibility of student self-reports*. Los Angeles, CA: Center for the Study of Evaluation, University of California.

Padgett, R. D., Goodman, K. M., Johnson, M. P., Saichaie, K., Umbach, P. D., & Pascarella, E. T. (2010). The impact of college student socialization, social class, and race on need for cognition. *New Directions for Institutional Research, 145*, 99–111.

Pascarella, E. T., Seifert, T. A., & Blaich, C. (2010). How effective are the NSSE benchmarks in predicting important educational outcomes? *Change, 42*(1), 16-22.

Pascarella, E. T., & Terenzini, P. T. (2005). *How college affects students: A third decade of research*. San Francisco: Jossey-Bass.

Pike, G. R. (1994). Applications of generalizability theory in higher education assessment research. In J. C. Smart (Ed.), *Higher education: Handbook of theory and research* (Vol. 10, pp. 45-87). New York: Agathon Press.

Pike, G. R. (2006). The dependability of NSSE Scalelets for college- and department-level assessment. *Research in Higher Education, 47*, 177-195.

Pike, G. R. (2011). Using college students' self-reported learning outcomes in scholarly research. In S. Herzog & N. A. Bowman (Eds.), Validity and limitations of college student self-report data. *New Directions for Institutional Research, 150*, 41–58.

Pike, G. R. (2013). NSSE benchmarks and institutional outcomes: A note on the importance of considering the intended uses of a measure in Vvlidity studies. *Research in Higher Education, 54*(2), 149-170.

Porter, S. R. (2011). Do college student surveys have any validity? *The Review of Higher Education, 35*(1), 45–76.

Rocconi, L. M., Lambert, A. D., McCormick, A. C., & Sarraf, S. A. (2013). Making college count: An examination of quantitative reasoning activities in higher education. *Numeracy, 6*(2).

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3*(3), 271-295.

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice, 15*(1), 22-29.

Appendix

Variance components from the G-study by class

Engagement Indicator	$\sigma^2(u)$	$\sigma^2(i)$	$\sigma^2(s,ui)$	$\sigma^2(ui)$	$\sigma^2(e)$	k
First-year						
Higher-Order Learning	0.011	0.000	0.000	0.023	0.642	4
Reflective & Integrative Learning	0.011	0.000	0.000	0.036	0.671	7
Quantitative Reasoning	0.012	0.000	0.000	0.022	0.846	3
Learning Strategies	0.012	0.000	0.000	0.012	0.730	3
Collaborative Learning	0.037	0.000	0.004	0.020	0.715	4
Discussions with Diverse Others	0.026	0.000	0.001	0.030	0.794	4
Student-Faculty Interaction	0.025	0.000	0.000	0.024	0.798	4
Effective Teaching Practices	0.011	0.000	0.024	0.029	0.650	5
Quality of Interactions	0.047	0.000	0.047	0.097	2.683	5
Supportive Environment	0.022	0.000	0.000	0.052	0.839	8
Senior						
Higher-Order Learning	0.011	0.000	0.000	0.024	0.639	4
Reflective & Integrative Learning	0.015	0.000	0.000	0.043	0.659	7
Quantitative Reasoning	0.017	0.000	0.000	0.023	0.897	3
Learning Strategies	0.021	0.000	0.000	0.012	0.779	3
Collaborative Learning	0.043	0.000	0.029	0.023	0.745	4
Discussions with Diverse Others	0.026	0.000	0.000	0.032	0.767	4
Student-Faculty Interaction	0.061	0.000	0.012	0.030	0.910	4
Effective Teaching Practices	0.012	0.000	0.025	0.031	0.654	5
Quality of Interactions	0.059	0.000	0.124	0.083	2.700	5
Supportive Environment	0.030	0.000	0.000	0.059	0.894	8

Note: k= number of items in the Engagement Indicator; All analyses estimated with 50 students

Book Review

Demonstrating Results: Using Outcome Measurement in Your Library. Rhea Joyce Rubin Chicago: American Library Association, 2006, 160 pp. ISBN 0838935605 $60.30

REVIEWED BY:
Beyza Aksu Dunya, Ph.D.
Bartin University

In an era of budget restrictions and rapid environmental change, libraries increasingly need to demonstrate their value. Outcome measurement is commonly used by libraries to measure how their services and programs affect users. There are many guidelines available for libraries that plan outcome assessment to evaluate their impact on users. *Demonstrating Results: Using Outcome Measurement in Your Library* by Rhea Joyce Rubin (2006) provides guidelines and examples of developing and implementing outcome measurement using two case studies in a hypothetical public library (Anytown Public Library) in the United States. Rubin often uses questions to guide readers and provides applications of the concepts presented in each chapter. She provides several questions within the text that give readers a more active role by making the book less narrative driven and more thought provoking. A total of 14 work forms are presented as an appendix to help readers apply information presented in associated chapters. There are six chapters followed by six tool kits for practitioners. A brief glossary is provided at the end of the book to support a common terminology for readers with various levels of content knowledge.

Rubin explains the concept of *outcome* by giving examples of changes that may occur as a result of library programs: knowledge, skills, attitude, behavior, or condition. Outcome measurement is defined as "a user-centered approach to the planning and assessment of library programs or services" (p. 16). The process of designing an outcome measurement plan is presented with the aid of solid examples that distinguish between outputs and outcomes. For readers who are new to outcome measurement, this introductory chapter helps develop a basic understanding of outcome and output concepts with examples and case studies.

Chapter Two addresses how libraries plan programs to meet intended user needs. During implementation and evaluation phases libraries assess whether the planned outcomes are met. This chapter clarifies the difference between interim and long-term outcomes. Interim outcomes that are sometimes called *outputs* (i.e., participation rates, user statistics) facilitate determining long-term *outcomes* (e.g., behavioral change). After presenting different outcome types, Rubin walks the reader through the outcome statement development process. First, she explains how to gather data to detect and define potential outcomes, which are referred to as "candidate outcomes" in the chapter. When writing a candidate outcome statement, Rubin emphasizes not to use the word *library* but to focus on users, using general action verbs. She then provides sample outcome verbs and example if-then statements to explain concepts in more detail.

Rubin introduces logic models through a *so what* linkage in if-then statements. She emphasizes that people should keep asking "so what?" until reaching the last, long-term outcome. The chapter also exemplifies possible gaps in the logical chain while building if-then connections. For example, if there is something other than the proposed factor that can explain an observed change, there is a gap in the logic flow that should be fixed. Considering the importance of planning "if-then" flow as an initial step for building effective logic models, this chapter can serve as a guide for people who intend to create logic models for their programs. Yet, the discussion and examples used to illustrate candidate outcomes could have been explored in more detail in the chapter.

Assessments become more valuable and useful when they combine both purposes of accountability and seeking improvement.

Rubin describes steps in writing comprehensive and measurable outcome statements, specifying outcome indicators and setting targets in Chapter Three. She first explains the important distinction between an outcome and outcome indicator. An outcome indicator is a specific measure of change or action on the part of the user, and "a well-selected outcome indicator attempts to tell a story to emphasize the impact of the program on individuals" (p. 34). She provides specific examples to clarify the distinction between outcome and indicator. Then, she shows precise examples of four characteristics of an indicator: (a) verb, (b) object, (c) quantity of action, and (d) time frame. An outcome may require one to three indicators that cover all the dimensions of a concept. Rubin provides an example using a library program aiming to support the habit of reading. The potential dimensions of this outcome would be frequency of reading, positive attitude toward reading, and enjoyment. Each of these dimensions can be captured by measurable or observable indicators and compose an outcome statement. But, not all indicators are always direct. In some situations, "proxy" or "surrogate" outcome indicators can substitute for directly observable indicators and imply the outcomes. Some important considerations when specifying indicators and constructing indicator statements are a data analysis plan, timetable, and the context in which the library functions and the program is launched. In addition, external influences (e.g., economic, political, or social environment), program participant characteristics (e.g., literacy level, native language), and library setting (e.g., abilities of staff, funding sources) impact specific outcome indicators for a library program. Therefore, Rubin emphasizes that indicators should be decided by giving full consideration to the context of the library, program, and community. This is an important

point to emphasize, since public libraries are context-dependent, and one that applies to assessment situations in other context-dependent areas, such as assessment in higher education institutions (e.g. Suskie, 2009).

Prior to obtaining the data, when educators consider how they will use student performance data on these tests they are more likely to plan possible changes or action.

The other step to writing good outcome statements is to set targets for each indicator. Rubin describes targets as success indicators for the library which should be represented by both proportions and numbers of participants. She states that targets should not be used to make comparisons across different libraries, given the contextual differences, but to compare a program's functioning within a library over time. This statement overlaps with some other assessment professionals' (i.e. Banta & Palomba, 1999) arguments for the use of standardized methods for assessing accountability. Standardized measures can be developed to report retention, graduation, employment, and alumni satisfaction statistics; however, they should not be used to make comparative decisions for accountability purposes. After stating success indicators for individuals, and setting targets for the library, the last step is to compose outcome statements. In two separate figures, she lists components of an outcome statement and provides sample outcome statements.

Chapter Four starts with a discussion of the difference between outcome measurement and scientific, experimental research. Outcome measurement, as a specific type of assessment, is designed for assessing individual programs based on changes among participants and does not concern generalization of results. Outcome measurement is not grounded on a specific hypothesis and results should not/cannot be compared to larger populations. Rubin's stance on generalizability of the outcome measurement results corresponds with other authors in the assessment field. For example, Suskie (2000) encourages people to consider various factors such as cognitive style and cultural experience while assessing individual students. As each research design has specific approaches and data collection tools, outcome measurement often employs data collection tools which include: (a) existing records, (b) surveys, (c) tests, (d) interviews, and (e) observation. Under each data collection method, Rubin discusses their advantages and disadvantages. Despite the benefits of presenting cautionary issues associated with each tool for future users, though, I do not see those issues as *disadvantages*. For example, she lists the disadvantages of surveys as language burden, response rates, and social desirability concerns of respondents; however, each of these issues can be handled by careful survey design and should not be considered as *barriers* because surveys are an important data collection method for assessing attitude, behaviors, change, and even knowledge. Rubin provides a check list (work form) of relevant questions to guide users while selecting an appropriate data collection method. She adds that some outcomes may be assessed using multiple instruments (e.g., survey followed by interview).

The next step after choosing the data collection method is creating or adapting appropriate data collection instruments. She emphasizes the importance of this step: "Your data will only be as good as your data collection instruments" (p. 53). In a separate work form, she lists several criteria for evaluating the relevance of each question on an instrument to prevent redundancy. At the end of the chapter, she briefly mentions data analysis, with commonly used descriptive statistics including percentages, mean, mode, and cross tabulations. She warned readers not to use associational findings obtained from statistical tests (e.g., *t*-test) to draw causal conclusions. I found this part essential for readers who are new to quantitative methods since causality is often confused with association.

Chapter Five addresses the challenge of outcome measurement—getting people involved in outcome assessment knowing that they usually overestimate the work required to complete the assessment efforts. According to Rubin, the best way to overcome this issue is to create an outcome measurement plan (i.e., logic model). A sample outcome measurement plan, created by the California State Library, is provided along with a straightforward and applicable blank template for readers. She points out the importance of addressing participants' questions of "why" before starting actual measurement activities. Then, she explains the need for an external data collector to avoid using the direct program provider as the evaluator, and the need for pilot-testing the data collection. At the end of this chapter, Rubin explains how to design an action plan. The action plan is the operational form of a logic model, designed for answering who, what, and when. I agree with her point that a well-developed action plan facilitates implementation of outcome measurement, and helps predict time and resources needed for the actual implementation.

Cognitive bottlenecks relate to the difficulties students have with specific content. Cognitive bottlenecks create obstacles to student success and persistence in a discipline.

Chapter Six starts with interesting information about the use of outcome measurement results: in 2000, it was found that only 44% of public libraries used their survey data for improvement. Rubin then mentions the factors underlying this tendency to underutilize results; I think those factors are still relevant in library assessment practices today. She proposes key suggestions to make the most of outcome data, including how to interpret and communicate results. First and foremost she outlines potential data interpretation tools and methods in a straightforward manner, and mentions which data and analytical methods fit which data interpretation tools. One uncommon tip she provides for readers about interpretation

of open-ended responses is very useful—she states that in open-ended responses the interpreter/evaluator should focus on minority responses rather than common responses, as minority responses may reveal important patterns about the services. She then explains strategies for communicating results to service providers, funders, volunteers, users, and the public. The outlets she mentions are still frequently used to distribute findings and demonstrate library impact to people (e.g., newsletters, anecdotes and success stories, fact sheets, and annual reports). However, this chapter should be updated to include modern technologies in further editions. Lastly, Rubin explains how outcome measurement results can be used to make informed decisions and modifications to outcomes, indicators, data collection methods, timeline, staff, and other resources if needed.

Student learning outcome assessment data also led to changes in operational goals, such as increased retention and graduation rates or curriculum revision.

At the end of the book there are six tool kits that can be extremely useful for applications. These tool kits provide sample outcome statements for various user groups; sample reaction and benefit surveys, measuring not only satisfaction but also overall training input to participants; sample confidentiality statements; information about developing item types including ordering and formatting; and guidance on data cleaning, coding, and processing issues as well as sampling, deciding sample size, and sampling method.

Conclusion

This book has some weaknesses that might be addressed in further editions. First, definition of quantitative tools and approaches are too limited and simple. Although this is not a methodology book, I would expect a bit more detail and examples on the common quantitative approaches in outcome measurement. Second, Rubin's repeated statement that "sophisticated sampling and data analysis methods are not needed because outcome measurement does not attempt to make generalizations" (p. 42) may mislead some readers. Such a statement might be discouraging for people who are new to outcome measurement and intend to learn/use sophisticated methods. It should be a priority to employ the most valid and credible approaches, which can be sophisticated. Third, the book was first published in 2006 and the chapters should be updated to reflect new technology and tools in data collection and reporting for library science.

Despite the weaknesses, I recommend this book as an introductory resource for readers with various levels of understanding of outcome measurement due to the strengths it carries. First, for those who are new to the field of outcome measurement and library assessment, Rubin breaks down each stage of outcome measurement into smaller components, and walks the reader through using thought-provoking questions, blank templates, and case studies. Second, frequent use of figures throughout the book helps convey key points to the reader in a direct way. Third, the online work forms can be used in staff training activities and workshops on outcome measurement.

Demonstrating Results is a reference book for practitioners who aim to implement outcome measurement in public libraries. It can also guide other types of libraries such as academic and research libraries. It successfully extends discussion on the use of standardized measures, direct and indirect measures for evidence and contextual issues in assessment to the library field. Despite the weaknesses mentioned in this review, people who aim to learn about planning and conducting outcome measurement in libraries or conduct staff training should utilize this resource.

References

Banta, T.W., & Palomba, C.A. (1999). *Assessment essentials: Planning, implementing, and improving assessment in higher education*. San Francisco, CA: Jossey Bass.

Rubin, R. J. & Public Library Association. (2006). *Demonstrating results: using outcome measurement in your library.* Retrieved from http://www.loc.gov/catdir/toc/ecip0518/2005024218.html

Suskie, L. (2000, May). Fair assessment practices: Giving students equitable opportunities to demonstrate learning. *AAHE Bulletin, 52*(9), 7–9.

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

Notes in Brief

Planning the intended use of data and identification of bottlenecks are two best practices that faculty and administrators can use when they conduct assessments for the combined purposes of accountability and improvement. Prior to data collection, they need to have a clear plan of how the results should offer worthwhile insights. Upon identification of bottlenecks to learning and efficient operation of units, faculty can develop appropriate action steps to address these trouble spots. Planning for the use of data should improve the assessment process itself. The process of identifying bottlenecks will mostly help to improve outcomes. This article gives examples of how both practices were used effectively for both student learning and operational outcomes. Using these two best practices led to enhanced decision-making ability that completed the assessment loop. The examples show improvement in student learning, increased retention rates, and more effective educational programs.

AUTHORS

Phyllis Blumberg, Ph.D.
University of the Sciences

Two Underused Best Practices for Improvement Focused Assessments

Currently, the primary purpose of assessment of student learning in higher education is to document what is occurring (Hutchings, Huber, & Ciccone, 2011; Ikenberry & Kuh, 2015). These assessment efforts are often done to comply with regional or specialty accreditation standards. Such assessments are and will remain essential. Since educators want their programs and institutions to become or remain accredited, they often document a very high percentage of outcomes as met or even exceeded expectations. Yet, another essential purpose of assessment is to make improvements. These accountability assessments may not lead to data that can be used for improvement. When most expected outcomes are met, there is no reason to try to improve or to make changes. Although necessary for improvement, faculty and administrators may be reluctant to conduct assessments that reveal a program's weaknesses. Faculty fear they will look bad when students do not meet their learning outcomes or programs do not reach their operational goals. Faculty actually look bad if they never try to improve (Massa & Kasimatis, 2017). Assessments become more valuable and useful when they combine both purposes of accountability and seeking improvement.

This article showcases two best practices that faculty and administrators can use when they are conducting assessments for the combined purposes of accountability and improvement: plan intended use of data and identify bottlenecks in student learning. While both practices can lead to improvements in student learning and more effective operations, they come from different sources. Planning the intended use of assessment data comes from the mainstream current assessment literature (Kuh et al., 2015). Identification of bottlenecks is an evidence-based practice that educational developers use to help faculty revise their courses (Pace & Middendorf, 2017).

Both practices can be used for the two common types of assessment: student learning and operational outcomes. However, these practices offer different improvement benefits. Planning for the use of data should improve the assessment process itself. The process of

CORRESPONDENCE

Email
p.blumbe@usciences.edu

identifying bottlenecks will mostly help to improve outcomes. Changes to both the assessment process and the assessment outcomes are useful for both accountability and improvement assessment functions.

This article gives examples of how both practices were used effectively at the author's institution. This is a private, specialized, small university offering undergraduate and graduate degrees in the sciences and clinical professional degrees in the health sciences. The examples discussed come from recent assessment reports completed by directors of academic units.

Plan Intended Purpose of Assessment Data

Assessments become more valuable and useful when they combine both purposes of accountability and seeking improvement.

Prior to data collection, faculty and administrators need to have a clear plan of how the results should offer insights about student learning or effective operation of units. Without such a plan, the data may not be relevant or may not be acted upon (Kuh et al., 2015). Although this seems like a common-sense idea it is not always used. Skipping how data will be used is not explicitly mentioned in the often referred to assessment cycle heuristic (Kinzie, Hutchings, & Jankowski, 2015; Suskie, 2009). When faculty just need to report on assessment data, as they might do in the accountability function of assessment, they may not have planned how the data will be used. In such cases they may just collect data that looks relevant to the program. However, assessments take on additional meaning once the faculty explicitly plan how they will use the data. Explicit plans for data use lead to more precise questions about how well the program is meeting its goals about student learning or efficient operations. Planning provides an anticipated idea for how data will be interpreted. Thus, it leads to better decision making (Kuh et al., 2015).

The following two examples both use nationally normed, external exams as an appropriate measure of student learning. Prior to obtaining the data, when educators consider how they will use student performance data on these tests they are more likely to plan possible changes or action. However, the first example illustrates how an educational program initially collected data for accountability purposes without planning for its use. Once they identified the intended use of the data they were able to close the assessment loop both for accountability and improvement purposes. In contrast, the second example illustrates the planned use of data.

Prior to obtaining the data, when educators consider how they will use student performance data on these tests they are more likely to plan possible changes or action.

The biological sciences department requires that all graduating seniors take the ETS Major Field Test for biology as one of their major indictors of student learning. The requirement was to take the exam but the test had no impact on student grades or graduation. Students took this exam toward the end of their last semester and they did not take it seriously. While faculty were not pleased with the results they continued to require it because they felt they needed a valid, cumulative measure of student learning that was easy to administer. This is an example of conducting an assessment just to collect data for the sole purpose of accountability. Once the faculty asked what they would do with the student scores on this exam, they cared about student performance. This led them to make changes to try to improve performance. First, they moved the exam to an earlier semester. Students who performed significantly below the national average on the separate sections were asked to take another course relating to this content before graduating. Faculty assumed that this additional course should remediate these deficiencies. Upon further inspection of the results, the faculty found that many students, even some of their best students, were doing poorly on a few sections. This led the faculty to examine the alignment between their curriculum and the content on this national exam. They realized the exam was not a good indicator of mastery of the content emphasized in their major. Faculty are now considering using a different exam to measure cumulative student learning in their major. A possible operational outcome would be to identify or develop an appropriate and valid cumulative exam that aligns with their learning outcomes.

The faculty in the pharmacy program planned how they would use assessment data for both accountability and improvement purposes. For years the pharmacy program has been requiring students to take a test of mastery of pharmacy knowledge. The faculty use the results to gauge how well their students are doing in comparison to their national peers, as a stated student learning outcome. When repeated results indicated that pharmacy students toward

the end of their first professional year of training were below the national norms, the faculty decided to change the curriculum to help the students master the required content earlier and better. At first, they made small changes in the scope and sequence of material. These changes did not lead to significant improvements on this exam. When small changes did not lead to improvement in student performance, faculty were motivated to totally revise their curriculum and how it is taught. The new curriculum fully integrates the basic pharmaceutical sciences with the clinical applications. Instead of the traditional lecture-based courses they will be using many more active learning techniques, especially team-based learning. The new curriculum is being implemented this year. The scores on this nationally-normed exam will be used as a major indicator of the success of the changed curriculum.

In addition to using nationally normed exams, many faculty use course-embedded assessments with intended purposes, as the following example illustrates. The general education program assessment plan explicitly states the intended use of the data, "The evidence will be used to make informed decisions about curriculum, pedagogy, assessment, and instructional resources". This program requires that all undergraduate students gain competency in six skills. Specific courses have been approved to teach and assess students on one or more of these skills. Students are motivated to take them seriously since the assessment activity is part of the course grade. As the assessment plan states, faculty-directed, course-embedded assessments were chosen because they are more likely to be used for curricular improvement.

Each skill is considered every three years and two skills are reported on annually. Faculty who teach these skill-approved courses report on cohorts of student performance using a course-specific, summative assessment instrument that measures this skill. These direct measures of student learning may take different forms but must include a four-point scoring scheme (1. not met; 2. approaching; 3. met; and 4. exceed expectations) for the student learning outcome(s). In 2017, faculty who taught courses that included ethics or oral communication reported on their assessments. Greater than 90% of the students were reported as meeting or exceeding expectations for both skills, with some faculty reporting extremely high levels of students exceeding expectations (e.g., 100%). While these high scores were fine for accountability purposes, considering improvement caused the general education committee to delve deeper into the meaning of these results. These committee members found that different instructors use different criteria for meeting these levels. As a result, they decided to hold faculty focus groups to talk about how the skills are assessed. These focus groups led to the development of clearer criteria for scoring student achievement; these criteria will be used to develop skill-specific rubrics which should be used by all instructors whose courses satisfy the skill. The goal is to develop assessment standards that are similar across different courses and instructors. Next, the general education committee will conduct professional development with faculty to calibrate the instructors' use of the rubrics. Such development should lead to a consistent application of rubric criteria across instructors and courses.

The goal is to develop assessment standards that are similar across different courses and instructors.

In addition to planning the use of data, results of assessments can help identify ways to improve programs. When the data indicate students are not doing as well as expected, faculty can try to find why these results were obtained. Identifying bottlenecks can be a useful method for determining where the problems are.

Identify Bottlenecks in Student Learning and Operational Effectiveness

Since the beginning of this century, faculty at Indiana University have been engaged in a process designed to increase learning (Pace & Middendorf, 2004). The first step in this process is to identify bottlenecks in student learning. Bottlenecks can be either cognitive or emotional. Cognitive bottlenecks relate to the difficulties students have with specific content. Cognitive bottlenecks create obstacles to student success and persistence in a discipline. Emotional bottlenecks relate to student anxieties or fears about the content. Math anxiety and religious beliefs that might promote resistance to the concept of evolution are good examples of emotional bottlenecks. Faculty in at least ten countries now are using this evidence-based process to identify ways to increase student learning (Pace & Middendorf, 2017). Although not referred to as a formal assessment method, identification of trouble spots is frequently used for continuous program improvement in higher education.

Instead of asking faculty to identify weaknesses in their programs, ask them to identify bottlenecks that impede student learning and success in educational programs. Bottlenecks can be found by inspecting where student cohorts struggle. This turns assessment into looking for ways to improve and does not the carry the negative connotation of weaknesses. When applied to programs, the identification of bottlenecks can be a practical tool that faculty and administrators can use in assessing programs.

The bottleneck concept has a long history in manufacturing improvement initiatives whereby managers identify where and why product creation is reduced. A similar process can be applied to educational programs. Faculty can identify bottlenecks by reviewing semester-to-semester retention rates, student grades, and comments. Upon identification of trouble spots in educational programs, people can develop appropriate action steps to release these bottlenecks. To address the recent concern about timely graduation rates (program or institution-wide bottleneck), higher education administrators have adopted various approaches to increase completion rates. Programs geared toward increasing retention of beginning students are common (Hart Research Associates, 2012).

Once the bottleneck has been identified faculty can make appropriate changes to the program that attempt to address these trouble spots. For example, a program might identify that many students do not master required mathematics skills. An analysis of the items that many students got wrong on these skills assessments would provide diagnostic information about which types of questions or content are difficult for students. Thus, the assessment data identifies specific concepts or skills that the students find especially hard to master. The faculty could explore if they could find a different way to teach these concepts or skills to make it less difficult for the students. After changing how they teach this content, a resulting student learning outcome might be to attain a 15% increase in the number of students who achieve mastery scores on those questions that relate to these identified mathematics skills across several courses that assess them. This program also could identify an operational outcome that increases student retention by 10% in the program. Identification of bottlenecks and making changes because of this knowledge may be less threatening for faculty than stating assessment in terms of vulnerabilities.

Identification of bottlenecks and making changes because of this knowledge may be less threatening for faculty than stating assessment in terms of vulnerabilities.

Like their colleagues across the country, the faculty at this university are concerned with retention in STEM (Felder & Brent, 2016), as it traditionally has been a barrier to students remaining in their intended major—whether that is in STEM or in health professions that require a good STEM foundation. At this specialized science and health science university all students must do well in STEM courses not only to stay in their major but also to remain at the university. For example, doing well in organic chemistry is required for not only chemistry majors but also pre-health professional students who aspire to become pharmacists and physicians. Faculty members have been employing reform efforts to teach using best practices in most of the STEM introductory courses. The reform efforts in the general chemistry course, described next, illustrate a sustained effort to identify and overcome bottlenecks in a gateway STEM course required for most of the students at this university. This assessment has been used for both accountability and improvement for years.

In general chemistry, prior to 2002, more than 30% of the first year students earned a D or F or withdrew from the course (DFW). The course involved weekly three hours of lecture, two hours of laboratory, and an optional one hour for recitation where students had the opportunity to ask questions and the professor demonstrated the solution to chemistry problems. The faculty reviewed the mistakes students made on the exams and found that a majority of students had the most trouble with higher-order questions where they had to apply concepts. Thus, problem-solving skills were the bottleneck. In 2002 the faculty changed the format of the recitation from a large class to mandatory smaller recitation sections where the students solved problems in small groups. This restructuring led to a 10% reduction in DFW grades (Mahalingam, Schaefer, & Morlino, 2008). The following year the students were required to do homework where they solve problems prior to coming to the recitation. While students came to class with their homework done, many still did not understand how to solve these problems. Upon questioning the students, they indicated they copied their answers from others, as the assignments were mandatory. The faculty hypothesized that implementing

an online homework system should help overcome the bottleneck of understanding how to solve problems. Now the homework problem sets are more relevant to the exam questions, so students likely take the homework more seriously as meaningful preparation for exams, as opposed to busy work. This is a good example of aligning learning/practice activities with assessments. After experimenting with different online homework systems, faculty found that providing hints on how to solve problems throughout, and not just showing the steps of the problem, was the most helpful for student mastery of problem-solving skills (Mahalingam & Fasella, 2017).

While the dissertation is a major challenge for doctoral students, bottlenecks can occur at various stages of graduate education. Faculty should look at where attrition occurs to determine program-specific bottlenecks.

The passive nature of the lecture classes also served as a barrier to problem-solving skill acquisition (Weimer, 2004). Since 2009, the faculty who teach this course have incorporated an audience response system to allow all students to answer questions throughout the lectures. The audience response system gives students immediate feedback that allows them to evaluate their understanding of the content and its application to problems.

The grading system also changed since 2002 when only exam grades and laboratory performance counted. Now, performance on homework and recitation problems counts toward the final grade. Therefore, final course grades are not valid comparisons. Instead, performance on exams is the appropriate pre- and post- educational intervention comparison. In addition, over time, the percent of application questions on each exam has increased. The percentage of students earning D or F grades on exams dropped from over 30% to 15% even as the exams got harder. This example shows how faculty can identify bottlenecks to student success and implement changes that result in significant increases in student learning and understanding of the content.

Retention and graduation rates are of even more concern in graduate education because nationally there is about a 50% attrition rate from PhD programs (Lovitts & Nelson, 2000). While the dissertation is a major challenge for doctoral students, bottlenecks can occur at various stages of graduate education. Faculty should look at where attrition occurs to determine program-specific bottlenecks.

The director of the master's degree program in biomedical writing developed an operational goal of a 75% graduation rate, which he measured for both accountability and improvement purposes. This is a reasonable graduation rate because this program attracts nontraditional students, most of whom are employed. Some students discover that the field of biomedical writing is not for them or decide that they want to pursue other careers. Recently, this program had a retention-to-graduation problem, as far fewer than 75% of the students graduated. Once this decreased graduation rate occurred, the director determined that most of the attrition occurred either during or at the end of the recommended first course. Between 25–50% of students were either dropping out during the course or not continuing to the next semester after taking this course. Therefore, the first course was this program's bottleneck. The program director decided to gather data about the course from the students who dropped out and from those who continued in the program. Other faculty and nonfaculty practitioners in the field also examined the syllabus. In addition, the director looked at student weaknesses in more advanced courses.

The data indicated that over the years, the instructor increased the required content and tried to raise the rigor of the course through several writing assignments which required accurate use of the American Medical Association (AMA) writing style. When a new, adjunct instructor began teaching this course she continued to implement these changes and even increased the expectations. The students perceived that the course was intended to weed out the less-qualified students, especially those who were not yet employed in biomedical writing. This perception is contrary to the philosophy and goals of the program which aims to give students the skills to be able to be employed in biomedical writing or to advance their biomedical writing careers. No courses are expected to eliminate less-experienced students. The conclusion of the faculty and the external reviewers was that the course was too ambitious for beginners. Those students who were already employed as biomedical writers were able to

succeed with the assignments.

Because of this review the program director together with the instructor made a significant change in the content of the course. During this course the students are now taught how to write research reports using the industry's standard conventions, such as what goes into the introduction, methods, results, and discussion sections, and how to write using the AMA writing style—instead of assuming they knew how to do this. Some of the content was removed from this course and placed in a more advanced course. Since the implementation of these revisions, the dropout rate after taking this course fell to 5%.

These two examples show that identification of bottlenecks can be used for both student learning and operational outcomes. Once the bottlenecks are identified, the most critical step is to close the assessment loop by making changes to overcome the bottleneck. These changes can be made incrementally over a long period as the chemistry example illustrates or made quickly as was done with the biomedical writing example. In both cases, faculty were comfortable talking about assessments that showed previous students had struggled because they now fostered greater student success.

Discussion and Conclusion

The examples described here mirror the different types of recommendations that result from assessment (Massa & Kasimatis, 2017). As the examples show, assessments can lead to more than one type of recommendation. Course or curriculum revision occurred in pharmacy and biomedical writing. The faculty changed their pedagogy in chemistry and pharmacy. The general education assessment led to improved assessment of student learning, and a better alignment between the curriculum and the assessment tool. Repeatedly observing lower than expected student performance on exams can lead to different improvement action plans. Once the ETS Biology exam had a real purpose the faculty looked at the instrument itself. Since they were satisfied with their curriculum, they realized they needed an exam that aligned better with their learning outcomes. In the pharmacy example, the results suggested that the faculty needed to change their curriculum because the test was a valid measure of what the faculty expected the students to learn. By taking a deep dive into the data the faculty were able to close the assessment loop. The programs improved student learning and increased retention rates both in gateway undergraduate STEM courses and an introductory graduate course. Best of all, these improvements were made without needing many additional resources.

Student learning outcome assessment data also led to changes in operational goals, such as increased retention and graduation rates or curriculum revision.

These examples illustrate how faculty collect and study assessment data after planning the intended use of data or by identifying bottlenecks. Such data helped to determine whether student learning outcomes were met, which led to changes in what and how students were taught as well as how they were tested. Student learning outcome assessment data also led to changes in operational goals, such as increased retention and graduation rates or curriculum revision. The examples provide evidence for the framework used throughout this article: both common types of assessment (student learning outcome and operations) can support accountability and improvement purposes.

The two best practices discussed here—planning the intended use of data and the identification of bottlenecks—facilitate assessments for the dual purpose of accountability and improvement. Both practices encourage faculty to engage in meaningful student learning outcome and operational assessments. Perhaps the greatest benefit of these practices is that they are nonthreatening for those who use them. These practices do not make individual faculty members look bad or identify weaknesses of individual courses that could be held against individuals. The use of these practices reflects well on the people who use them because it shows they are trying to improve their programs and student learning. When provosts or deans actively promote the use of these best practices they are creating a supportive environment for meaningful assessments to occur.

References

Felder, R.M., & Brent, R. (2016). *Teaching and learning STEM: A practical guide*. San Francisco: Jossey-Bass.

Hart Research Associates. (2012). *The completion agenda: Post-secondary education leaders' perspectives on issues of strategies for increasing completion rate*.

Hutchings, P., Huber, M. T., & Ciccone, A. (2011). *The scholarship of teaching and learning reconsidered*. San Francisco: Jossey-Bass.

Ikenberry, S. O., & Kuh, G. D. (2015). From compliance to ownership: Why and how colleges and universities assess student learning. In G. Kuh et al. (Eds.), *Using evidence of student learning to improve student learning* (pp. 1–23). San Francisco: Jossey-Bass.

Kinzie, J., Hutchings, P., & Jankowski, N. A. (2015). Fostering greater use of assessment results. In G. D. Kuh et al. (Eds.), *Using evidence of student learning to improve higher education* (pp. 51–91). San Francisco, CA: Jossey-Bass.

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

Lovitts B.E., & Nelson, C. (2000). The hidden crisis in graduate education: Attrition from Ph.D. programs. *Academe, 86*(6), 44–50.

Mahalingam M., & Fasella E. (2017). Effective use of technology for asynchronous learning to elevate students' knowledge and problem solving ability. In S. P. Ferris & H. A. Wilder (Eds.), *Unplugging from the classroom* (pp. 149–158). Oxford, UK: Chandos Publishing (Elsevier).

Mahalingam, M., Schaefer, F., & Morlino, E. (2008). Promoting student learning through group problem solving in general chemistry recitations. *Journal of Chemical Education, 85*(11), 1577–1581.

Massa, L. J., & Kasimatis, M. (2017). *Meaningful and manageable program assessment*. Sterling, VA: Stylus.

Pace, D., & Middendorf, J. (2004). Decoding the discipline: A model for helping students learn disciplinary ways of thinking. *New Directions for Teaching and Learning, 98*(3), 1–12.

Pace, D., & Middendorf, J. (2017). Foreword to using the decoding the discplines framework for learning across the dsiciplines. *New Directions for Teaching and Learning, 150*(3), 9–11. doi:10.1002/tl.20243

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

Weimer, M. (2002). *Learner-centered teaching: Five key changes to practice*. San Francisco: Jossey-Bass.