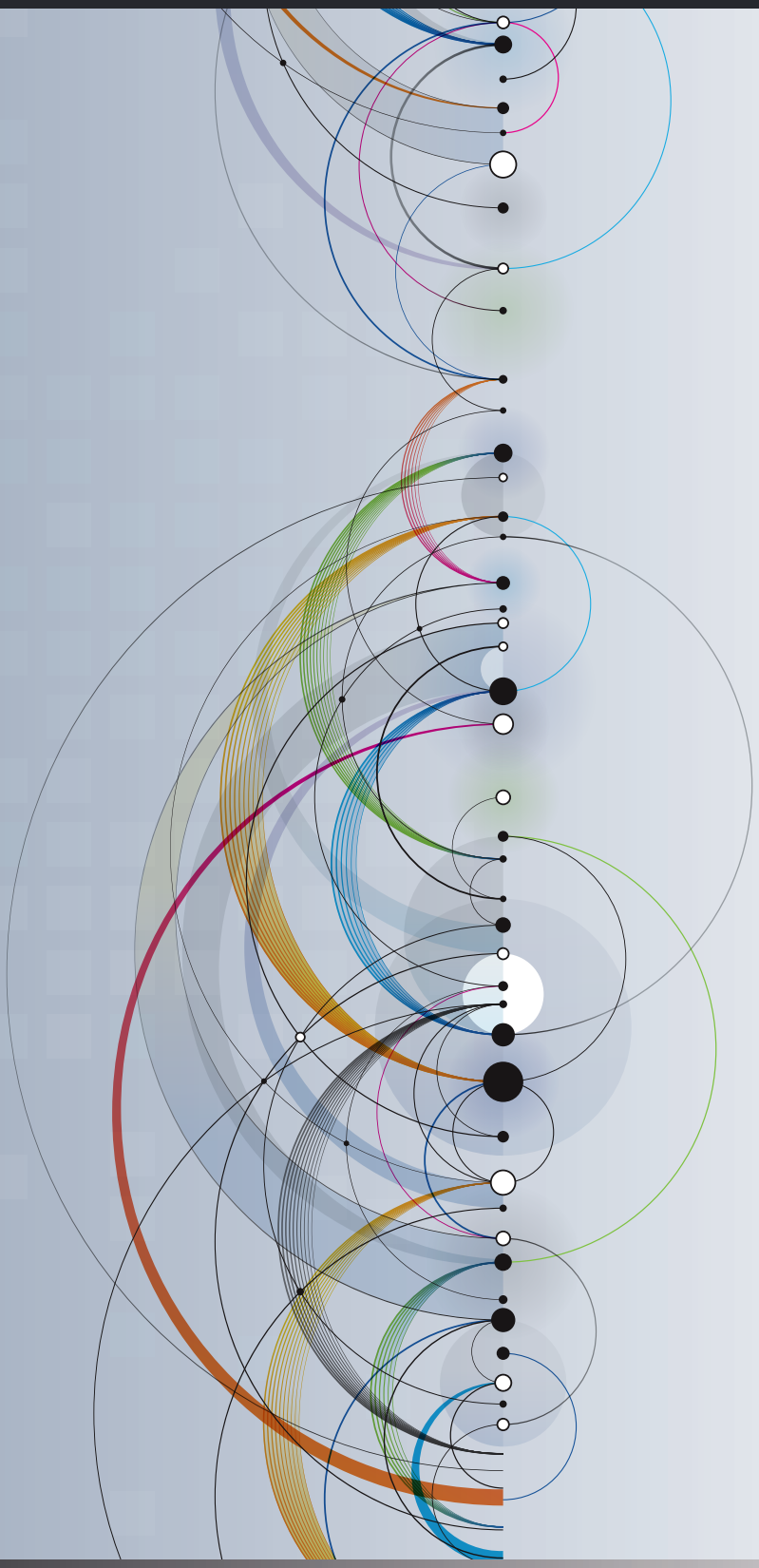
RESEARCH & PRACTICE IN ASSESSMENT

VOLUME EIGHT | WINTER 2013
www.RPAjournal.com
ISSN # 2161-4210



RPA

A PUBLICATION OF THE VIRGINIA ASSESSMENT GROUP

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, Gale, and ProQuest.

-Ruminate page 89

Published by:
VIRGINIA ASSESSMENT GROUP
www.virginiaassessment.org

TABLE OF CONTENTS

FROM THE EDITOR

The Embeddedness of Assessment

An understanding of embeddedness, the location of behavior and institutions within the social settings which condition and constrain them, can boost our understanding of the range of influences which affect organization, and so should have a critical impact upon the analysis of organizations.

(David Collins, Organizational change: Sociological perspectives, p. 133)

As the profession of higher education assessment advances, it continues to take on different forms at the field, organization, division, and department levels. One of these developments includes the introduction of a model which adopts a philosophy of embeddedness. This approach situates persons with a specialized knowledge of assessment in given organizational divisions such as student affairs, medical colleges, and academic units (e.g., arts and sciences). This issue of RPA focuses on a particular embedded sector of assessment within the university - student affairs assessment.

Assessment has been foundational to student affairs as a profession. Both the National Association of Student Personnel Administrators (NASPA) and American College Personnel Association (ACPA) professional associations have vibrant communities of practitioners and scholars focusing on this area. The Council for the Advancement of Standards in Higher Education (CAS) has been an influential component since 1979. More recently, a group of embedded student affairs assessment professionals has been able to successfully establish a professional association: the Student Affairs Assessment Leaders (SAAL). To this end, student affairs assessment offers advanced and established approaches to embedded assessment that warrant further discussion by the profession writ large.

The Winter 2013 issue of RPA opens with two provocative feature articles. In the first, Bowman presses student affairs professionals to critically examine the widespread practice of measuring outcomes by asking students how much they have learned. He addresses psychological processes that often result in flawed responses as well as factors that may lead to improvements in validity. In the second, Gerstner and Finney urge practitioners to ask, "Are students receiving the planned program?" Here, they offer a means of measuring the alignment between the planned program and implemented program.

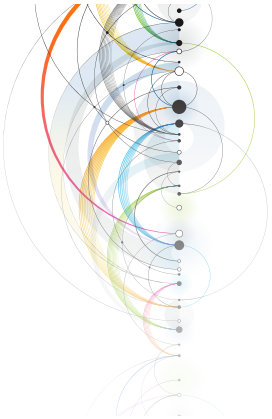
Four peer-reviewed articles are the mainstays for this issue. Ng, Skorupski, Frey, and Wolf-Wendel develop an instrument useful for exploring how commitments to diversity are reflected in teaching, research and service. Highly valuing a liberal studies emphasis, Puncochar and Klett offer a direct measure of student understandings of science inquiry processes. Athas, Oaks and Kennedy-Phillips suggest that we consider the value of university student employment with regard to the development of competencies and applied knowledge. Finally, DeMars, Bashkov and Socha examine gender differences in test-taking effort regarding three measures of motivation.

The reviews in this issue afford student affairs professionals with two substantive works to consider adding to their library. Successful Assessment for Student Affairs: A How-to Guide is reviewed by Lindsay, Stroud and Tubbs while Learning Is Not a Sprint: Assessing and Documenting Student Leader Learning in Cocurricular Involvement is reviewed by Endersby.

I would encourage readers to give pause and reflect on the first painting showcased in Ruminate. The issue concludes with original artwork by Keith Frew, coupled with excerpts from George Herbert Mead and Georg Simmel. Here, we are reminded the process of schooling has two primary foci, information and socialization, which the student affairs emphases on learning and growth suitably complement.

Regards,

Liberty University



Abstract

Asking college students how much they have learned or grown is a common assessment practice in student affairs and elsewhere. Unfortunately, recent research suggests that these self-reported gains do a very poor job of measuring actual student learning and growth. This paper provides an overview of the psychological process of how students likely respond to such questions and why their responses can be seriously flawed. It also discusses circumstances in which self-reported gains are somewhat more valid and offers concrete suggestions for student affairs professionals and other higher education constituents who seek to accurately measure student outcomes.

AUTHOR

Nicholas A. Bowman, Ph.D.
Bowling Green State University

Understanding and Addressing the Challenges of Assessing College Student Growth in Student Affairs

In an era of increased demands for accountability and limited financial resources in higher education, the assessment of college student outcomes has become crucial. Several recent books have provided excellent guidelines and examples for conducting college student outcomes assessments (e.g., Astin & antonio, 2012; Banta, Jones, & Black, 2009; Suskie, 2009; Walvoord, 2010). In general, these authors agree that multiple forms of assessment should be administered, that direct assessments should be employed when possible, and that assessment results should inform programmatic and institutional change. To measure academic outcomes, many institutions are using standardized examinations (e.g., Collegiate Learning Assessment, Collegiate Assessment of Academic Proficiency) as well as "authentic assessments," such as portfolios or rubrics of student work (Kuh & Ikenberry, 2009). These indicators can be used to assess the achievement of a particular level of skill or competence and/or the amount of growth that has occurred during the undergraduate years.

However, such formalized, direct learning assessments are rarely used to measure the effectiveness of student affairs in promoting student outcomes. These rigorous assessments not only require a great deal of resources, but they also indicate the types of academic and general cognitive skills that are generally not considered to be the primary focus of student affairs. As a result, student affairs professionals use a variety of other approaches for measuring learning and growth, including responses to broad national surveys (e.g., National Survey of Student Engagement), specific national surveys (e.g., ACUHO-I/EBI Resident Assessment), and a variety of locally developed surveys (I recently heard about a written questionnaire assessing student experiences and outcomes from a residence hall ice cream social!). In many cases, outcomes assessment in student affairs simply involves asking students what they have learned and how they have grown. The responses to these questions are then interpreted as indicating students' actual learning and growth.

CORRESPONDENCE

Email
nbowman@bgsu.edu

In many cases, outcomes assessment in student affairs simply involves asking students what they have learned and how they have grown. The responses to these questions are then interpreted as indicating students' actual learning and growth.

Recent research has cast serious doubt upon the (seemingly reasonable) assumption that college students can accurately report their own growth. If these self-reports were accurate, then one would expect a high correlation between students' self-reported gains on a particular outcome (e.g., critical thinking skills) and longitudinal changes on a well-validated measure of that same outcome. Across various samples and outcomes, the correlations between longitudinal and self-reported gains on the same construct are consistently low ($rs < .20$), and they are often not significantly different from zero (Bowman, 2010a, 2010b, 2011b; Bowman & Brandenberger, 2010; Gosen & Washbush, 1999; Hess & Smythe, 2001). In addition, the significant predictors of longitudinal growth (e.g., college experiences, student demographics, institutional attributes) often diverge considerably from the significant predictors of self-reported gains for the same construct (Anaya, 1999; Bowman, 2010b, 2011a, 2012; Bowman & Brandenberger, 2010; Porter, 2013). As a result, practitioners and researchers would arrive at remarkably different conclusions about the experiences that promote or hinder student growth depending on the type of outcomes assessment that they use. Through a synthesis of the existing literature and examination of several theory-driven hypotheses, Porter (2013) argues that college self-reported gains should not be used as indicators of actual student learning. Finally, relevant to many student affairs assessments, college students also have considerable difficulty reporting the educational impact of a particular experience or set of experiences; in general, students tend to overestimate the effects that their experiences actually have (Bowman & Brandenberger, 2010; Bowman & Seifert, 2011; Conway & Ross, 1984).

In this paper, I will first discuss why students may have such a difficult time reporting their own growth and why their self-reports may not even reflect their actual judgments. Next, I will propose several conditions under which students provide somewhat more accurate assessments of their growth. Finally, I will provide suggestions for student affairs practitioners and other higher education constituents who seek to measure and understand student outcomes.

The Psychology of Student Self-Reported Gains

Practitioners and researchers would arrive at remarkably different conclusions about the experiences that promote or hinder student growth depending on the type of outcomes assessment that they use.

In their seminal work, Tourangeau, Rips, and Rasinski (2000) proposed a four-stage model of the psychology of survey responses. The four steps involved, in order, are comprehension of the question, retrieval of memories associated with the question, judgment of the completeness and relevance of the memories, and mapping the judgment onto a response represented by one of the options provided. Below, a discussion of potential errors in college student self-reported gains is organized into these categories.

Comprehension

The language used in self-reported gain items, such as "thinking critically and analytically," is sometimes quite vague (Bowman, 2010a; Porter, 2011). Do students know what this phrase means? If so, do they all share the same definition? And are these definition(s) the same as the researchers' definition(s)? Even experts disagree considerably on the meaning of commonplace terms such as "intelligence" (e.g., Sternberg & Detterman, 1986), so it is reasonable to assume that students may also have different interpretations of terms used in self-reported gain items, such as "critical thinking skills," "general knowledge," and "leadership abilities" (Higher Education Research Institute [HERI], 2011, p. 1). This concern is further complicated by the fact that substantial cross-cultural differences exist on what constitutes complex thinking, interpersonal relationships, and even how a person defines oneself (for reviews, see Kitayama & Cohen, 2009; Markus & Kitayama, 1991; Nisbett, 2003). Thus, students from divergent cultural backgrounds may have systematically different interpretations of a given item. Moreover, some items are double-barreled in that they ask about two concepts at once. For example, if students are asked to report gains in "being an informed and active citizen" (National Survey of Student Engagement [NSSE], 2013, p. 6), then they might have a difficult time knowing how to respond, especially if they have become much more informed but not necessarily more active.

Retrieval and Judgment

The cognitive demands required to provide accurate self-reported gains are substantial. Ideally, students would estimate their own current skills or attributes, estimate their previous skills or attributes, and then have some means for directly comparing the two. However,

students generally do not follow this process; instead, they estimate their current skills and attributes and then attempt to determine whether or how these have changed over time (Ross, 1989). This distortion of the ideal process can lead to substantial errors, because students' estimates are biased toward their lay theories of change and stability over the lifespan. As Ross explains, most people think that their skills generally increase over time (with the exception of very late in life), whereas they think that their attitudes are quite stable. As a result, consistent with these lay theories, people tend to overestimate how much their skills and abilities have changed, yet underestimate how much their attitudes have changed (Conway & Ross, 1984; Goethals & Reckman, 1973; Markus, 1986; McFarland & Ross, 1987).

Interestingly, students may be reasonably accurate when estimating their current skills. Some early research found high correlations between self-reported knowledge and objectively tested knowledge (Berdie, 1971; Pohlmann & Beggs, 1974), and other studies found that self-reported and objectively tested skills on the same academic subject load onto the same factor within structural equation models (Pike, 1995, 1996). Moreover, a recent meta-analysis found a moderate relationship between objective measures of one's *current* knowledge level and self-assessments of knowledge (r = .34), whereas there was no relationship when examining *increases* in self-perceived and actual knowledge (r = .00; Sitzmann, Ely, Brown, & Bauer, 2010). Thus, the errors on self-reported gains may primarily occur not because of students' inadequate self-knowledge of their current attributes, but because they cannot or do not use adequate processes to estimate their growth over time.

Two additional biases may be considered to involve both difficulties with retrieval and failures to judge the adequacy of one's memories. Halo error occurs when students' perceptions of overall growth and development unduly influence their judgment of growth in specific domains. In a classic experimental example, Nisbett and Wilson (1977b) found that students were quite fond of a professor's European accent when he acted warm and friendly in a videotaped interview, whereas other students were annoyed by the same professor's accent when they saw him acting cold and distant in a different interview. Pike (1993) also observed direct evidence of halo error in self-reported gains when seniors reported on their overall collegiate experience. Other studies have provided indirect evidence by finding low correlations among longitudinal gains on various constructs, but moderate to high correlations among self-reported gains, which suggests that the interrelationships among self-reported gains may be inflated (Bowman, 2010b; Bowman & Brandenberger, 2010). Pike (1999) further demonstrated that halo error may account for up to 75% of the explained variance in self-reported gains among first-year students.

It is reasonable to assume that students may also have different interpretations of terms used in self-reported gain items, such as "critical thinking skills," "general knowledge," and "leadership abilities."

In addition, Pascarella (2001) argued that students may differ in the extent to which they perceive their educational experiences as beneficial; these chronic dispositions toward reporting (or not reporting) growth may also constitute an important source of error. He suggests that controlling for students' perceived gains during high school will largely or entirely correct for this error in college self-reported gains, but this practice has rarely been employed in higher education research. Recent studies have found that high school self-reported gains are at least moderately correlated with college self-reported gains (Bowman & Hill, 2011; Seifert & Asel, 2011) and that the results of regression analyses sometimes depend upon whether high school gains are included as a control variable (Seifert & Asel, 2011).

Response

Biases may also occur when students are asked to select a response option. On the NSSE, when reporting how much students' "experience at this institution contributed to [their] knowledge, skills, and personal development," the response options are "very much," "quite a bit," "some," and "very little" (2013, p. 6). All four of these categories are at least implicitly positive—and they are treated as positive in statistical analyses—so students are unable to state that they have not changed at all or that they declined. On the Cooperative Institutional Research Program (CIRP) College Senior Survey, students' response options for changes in their knowledge, skills, and understanding were "much stronger," "stronger," "no change," "weaker," and "much weaker" (HERI, 2011, p. 1). The CIRP scale eliminates some of the problems apparent on the NSSE scale, but only two options are available for reporting positive growth, which could lead to range restriction. Perhaps more importantly, the categories

on both surveys are quite vague. Do students draw similar distinctions between "quite a bit" and "very much" or between "stronger" and "much stronger"? The results from an older study on students' perceptions of college experience frequency descriptors (Pace & Friedlander, 1982) may be informative. When asked about the frequency of making appointments to see faculty members, 21% of students thought that the term "very often" meant more than once a week, whereas 33% of students thought that this meant 1-2 times a month, and a small percentage of students (2%) thought that this meant 1-2 times per year. Clearly, students can assign very different meanings to such descriptors.

Moreover, students may select a response category that portrays them in an overly positive light. For instance, a socially desirable response would be to say that they have gained a great deal while in college; the unappealing alternatives are to say that they have gained very little, not at all, or even regressed. Indeed, social desirability scales are significantly associated with college student self-reported gains (Bowman & Hill, 2011; Gonyea & Miller, 2011), and this relationship persists even when controlling for self-esteem, college satisfaction, and other potential confounding variables (Bowman & Hill, 2011).

Additional Problems and Processes

People tend to overestimate how much their skills and abilities have changed, yet underestimate how much their attitudes have changed.

As Krosnick (1991) explains, survey respondents are likely to become increasingly fatigued, disinterested, and distracted as they continue to take a survey. As a result, participants expend less energy (if any) on each of Tourangeau et al.'s (2000) four steps; Krosnick refers to this suboptimal responding as "satisficing." Self-reported gains may induce satisficing—particularly if they are included later in the survey—because these items require a great deal of cognitive effort, involve responses for which students likely do not have a preconceived answer, and often appear in succession with other such items that use the same response scale. Indeed, Barge and Gehlbach (2012) showed that satisficing is quite common when reporting college self-reported gains and that this tendency may substantially and adversely affect survey results (also see Chen, 2011).

Going a step further, Porter (2013) argues that a belief-sampling model of survey response more adequately captures students' thinking when considering their own growth. That is, instead of recalling actual memories and frequencies of events, students retrieve a host of beliefs, feelings, impressions, values and judgments (collectively referred to as "considerations") that are relevant to the question. The specific set of considerations that students retrieve is somewhat arbitrary and is based on what is accessible in that particular time and context. Porter offers an example of what this process might look like:

> Consider a student in a quantitatively-oriented major who is asked how her college experiences have contributed to her development in analyzing quantitative problems. Multiple considerations then enter her mind: memories of lectures from a statistics class; memories of having possibly worked on problem sets with other groups of students; a general impression that she [is] adept at math, based in part on her experiences in high school. These multiple, positive considerations then lead her to conclude that she has gained considerably in analyzing quantitative problems while in college. It is important to note that these considerations could easily be generated by a student, *but that none of them have anything to do with how much a student has learned while in college*. Because considerations that come into mind are a "haphazard assortment," it is clear that many, if not all, of the considerations that enter a student's mind will be related to their educational experiences, but not necessarily to how much they have actually learned in a specific content area. (p. 210, emphasis in original)

Of course, this hypothetical student may be "correct" in the self-assessment of her changes in quantitative skills, but the widespread use of this approach will be largely problematic for drawing conclusions about student growth in the aggregate. Porter tested several hypotheses regarding students' mental processes when reporting their own gains, and the results were quite consistent with predictions from the belief-sampling model. In addition, Bowman and Schuldt (in press) found that students' self-reported gains were higher when these appeared toward the beginning of a questionnaire than when presented toward the end (after reporting their college experiences), which also suggests that the mental availability of certain events likely influences student responses.

Conditions Associated with the Validity of Self-Reported Gains

The preceding discussion paints a rather gloomy picture of the use of self-reported gains as indicators of student learning and growth. However, there is reason to believe that this picture may be somewhat more optimistic under certain conditions. The validity of self-reported gains is substantially determined by the extent to which the outcome is salient and accessible to students. In their classic review, Nisbett and Wilson (1977a) argued that people generally have minimal access to their higher-order cognitive processes, and people's "introspection" on these processes is generally based on their lay theories of cognition. Psychologists have made similar arguments more recently about self-knowledge regarding one's own motivations (Wilson, 2002) and even which activities will lead to one's own happiness (Gilbert, 2007). While many people may have difficulty accessing introspective knowledge accurately, some students may be more attuned to their growth (or lack thereof) on a given outcome. For instance, many first-generation university students face considerable difficulties in their academics and social engagement (e.g., Pascarella, Pierson, Wolniak, & Terenzini, 2004; Zwerling & London, 1992), so they may be more aware of their cognitive and interpersonal growth. Consistent with this view, the correspondence between self-reported and longitudinal gains is greater among first-generation students than among other students (Bowman, 2010a, 2011b).

Instead of recalling actual memories and frequencies of events, students retrieve a host of beliefs, feelings, impressions, values and judgments (collectively referred to as "considerations") that are relevant to the question.

Moreover, students may be much better at estimating their growth on some outcomes than on others. For example, foreign language skills are largely developed through salient formal and informal experiences, and students receive regular feedback on these skills through course grades, instructor comments, and their (in)ability to communicate effectively. In contrast, leadership skills are harder to define, less subject to concrete feedback, and are not often quantified in terms of objective performance. A recent meta-analysis suggests that these outcome attributes are important; specifically, the correspondence between cognitive learning and self-assessments of knowledge is greater when participants are provided external feedback and when they have to opportunity to practice making their own self-assessments (Sitzmann et al., 2010). Perhaps for these reasons, the correlations between longitudinal and self-reported gains are virtually zero for abstract cognitive skills (which generally are not subject to direct feedback or frequent self-assessment), whereas these correlations are somewhat higher for non-cognitive attributes, such as attitudes, interpersonal skills, and intrapersonal knowledge (Bowman, 2010b, 2011b; Sitzmann et al., 2010).

Similarly, the phrasing of self-reported gain items may also affect their validity. For instance, even if students actually knew how much their cognitive skills had changed over time, it is unlikely that all students would have the same interpretation of "thinking critically and analytically," because this construct is quite broad and it contains academic jargon (Porter, 2011). Moreover, students' interpretations of the meaning of some outcomes might differ systematically. For example, "leadership skills" may connote something very different for White, middle-class North Americans (whose cultural contexts generally value individualism and uniqueness) than for Asians and Asian Americans (whose cultural contexts generally value collectivism and consensus; see Nisbett, 2003; Triandis, 1989). These problems can be remedied, in part, by using concrete language that has a similar meaning across diverse groups of students.

For example, foreign language skills are largely developed through salient formal and informal experiences, and students receive regular feedback on these skills through course grades, instructor comments, and their (in)ability to communicate effectively. In contrast, leadership skills are harder to define, less subject to concrete feedback, and are not often quantified in terms of objective performance.

The validity of self-reported gains also depends, in part, upon students' year in college. Several studies have indicated that biases in self-reported gains (e.g., socially desirable responding) appear to be greater among first-year undergraduates than among advanced undergraduates (Bowman & Hill, 2011; Pike, 1999; Seifert & Asel, 2011). This pattern may occur for multiple reasons. First, developmental research suggests that self-perceptions generally become more accurate among older children (Harter, 1999), and similar developmental processes may be driving these differences among traditional-age college students. Second, when students are in their last term of their undergraduate education, they may reflect upon their university experiences and how they have changed while attending college. As a result, these students may provide more accurate responses because they have previously considered their growth over time as opposed to providing answers that simply seem plausible (see Krosnick, 1991).

Although self-reported gains are more trustworthy under certain circumstances than in others, longitudinal studies are certainly preferable to cross-sectional studies for drawing inferences about change over time.

The final three attributes relate to issues that were discussed previously. First, self-reported gains will be more valid when an appropriate response scale is used; allowing students to say that a desired attribute did not change or has diminished is generally preferable. As one illustration, when graduating students were asked to provide self-reported gains in their religious beliefs and convictions (and were provided this full set of response options), almost half reported no change during university, and about 14% reported decreases (Lee, 2002). Second, social desirability also plays a role in the accuracy of self-reported gains. The prevalence of socially desirable responding may depend upon the phrasing of the instructions and items as well as the nature of the outcome itself. For example, it is probably less "threatening" for college students to report that they have not become more religious (which is not central to the mission and intended outcomes of most colleges and universities) than to report that their problem-solving skills have not changed. Third, halo error can be more problematic in certain circumstances. Some outcomes appear to be more susceptible to halo error than others; Pike (1993) found that self-reported gains in "personal development" (e.g., intrapersonal skills, self-directed learning) were much more strongly influenced by halo error than self-reported gains in quantitative skills and in understanding arts and cultures. The latter outcomes are fairly specific and not directly related to many students' undergraduate experiences, which likely explains why they are less conflated with general perceptions of growth.

Implications for Assessment in Student Affairs

The following suggestions are provided specifically with student affairs practitioners in mind, but these recommendations may also be useful for institutional researchers, higher education researchers, and others who want to design effective college student assessments.

1. Use longitudinal methods whenever possible. Although self-reported gains are more trustworthy under certain circumstances than in others, longitudinal studies are certainly preferable to cross-sectional studies for drawing inferences about change over time. After asking about self-reported gains for the past 20 years, the Cooperative Institutional Research Program (CIRP) removed these items from its 2013 Your First College Year and College Senior Surveys (see HERI, 2013), which suggests that this organization may have doubts about the usefulness of these items. Because the responses to these CIRP surveys are paired with The Freshman Survey—and all three surveys ask participants to report their current levels of various skills and attributes—CIRP datasets can still assess longitudinal changes during college.

2. Use specific language and multiple items to measure each student outcome. This recommendation actually combines two suggestions, but these are sufficiently related that they should be discussed together. For instance, asking students directly about "leadership skills" provides problems regarding both the ambiguity of language and the multidimensionality of this complex construct; in short, what exactly is meant by "leadership"? This problem can be remedied by providing items that measure behaviors, attitudes, values, and tendencies that exemplify various aspects of leadership. The original Socially Responsible Leadership Scale (SRLS) contained 104 items that indicate eight leadership constructs (Tyree, 1998). While this instrument constitutes an extreme example of the number of items (and subsequent versions of the SRLS contain fewer items), it illustrates the extent to which a complex concept can be measured in detail when it is the primary focus of a research or assessment project.

3. Never ask students to self-report their cognitive growth. There still may be some hope that a well-designed questionnaire can yield accurate estimates of student gains on some affective outcomes. However, self-reported and longitudinal assessments of cognitive outcomes provide such strongly divergent findings that these self-reports appear completely untrustworthy. As described earlier, standardized examinations and authentic assessments (e.g., portfolios or rubrics of student work) are likely the most effective means for assessing cognitive and academic growth.

4. Give pretests and posttests for content-based workshops and programs. As a way of exploring learning outcomes within a program or workshop, students could take a closed- or open-ended quiz on key concepts. This approach could be successful for a longer program (e.g., professional development over a semester), and a short quiz could also be useful for one- or two-hour workshops (e.g., regarding career planning). For the short version, some people may be skeptical of using a single quiz for both the pretest and posttest, because students'

responses may exhibit practice effects or students may be overly attentive to these specific pieces of information. If this seems problematic, two versions of the test could be created; half of the students complete Version A in the pretest and Version B in the posttest, and the other half of students would complete Version B and then Version A.

5. Collaborate across campus to conduct large-scale assessments. Coordinating efforts across departments, units, and divisions (including student affairs and academic affairs) can result in comprehensive assessments that would not otherwise be possible. For instance, students who take a critical thinking examination and/or other in-depth instruments might also report their involvement in various curricular and cocurricular activities so that one can determine whether these experiences predict performance and growth. This approach may also have the benefit of reducing survey fatigue, which has helped contribute to dramatic recent reductions in survey response rates (Adams & Umbach, 2012; Pew Research Center, 2012).

This emerging research on self-reported gains has strengthened my belief that having poor quality data is highly problematic and potentially misleading. The predictors of college student self-reported gains and longitudinal growth on the same construct differ considerably, and this divergence is sometimes systematic and even predictable.

Conclusion

A few years ago, a colleague and I had several discussions about whether it is preferable to have poor quality data or no data at all. This emerging research on self-reported gains has strengthened my belief that having poor quality data is highly problematic and potentially misleading. The predictors of college student self-reported gains and longitudinal growth on the same construct differ considerably (Bowman, 2010b, 2012; Bowman & Brandenberger, 2010), and this divergence is sometimes systematic and even predictable (Bowman, 2011a; Conway & Ross, 1984; Porter, 2013). Therefore, higher education practitioners and administrators can make faulty decisions about programs and practices if they rely too strongly upon students' subjective perceptions of learning and growth. Student affairs professionals face a host of circumstances that make them more likely to use this type of outcome assessment, so they must be particularly diligent about avoiding the problems associated with perceived growth. Although it is certainly more challenging and expensive to collect high-quality, longitudinal data on student outcomes, the long-term benefits will generally outweigh the costs.

AUTHOR'S NOTE:

I thank Vivienne Felix for her feedback on an earlier version of this manuscript.

References

Adams, M. J. D., & Umbach, P. D. (2012). Nonresponse and online student evaluations of teaching: Understanding the influence of salience, fatigue, and academic environments. *Research in Higher Education, 53*, 576-591.

Anaya, G. (1999). College impact on student learning: Comparing the use of self-reported gains, standardized test scores, and college grades. *Research in Higher Education, 40*, 499-526.

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

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

Barge, S., & Gehlbach, H. (2012). Using the theory of satisficing to evaluate the quality of survey data. *Research in Higher Education, 53*, 182-200.

Berdie, R. F. (1971). Self-claimed and tested knowledge. *Educational and Psychological Measurement, 31,* 629-636.

Bowman, N. A. (2010a). Assessing learning and development among diverse college students. In S. Herzog (Ed.), *Diversity and educational benefits* (New Directions for Institutional Research, no. 145, pp. 53-71). San Francisco, CA: Jossey-Bass.

Bowman, N. A. (2010b). Can 1st-year college students accurately report their learning and development? *American Educational Research Journal, 47*, 466-496.

Bowman, N. A. (2011a). Examining systematic errors in predictors of college student self-reported gains. In S. Herzog & N. A. Bowman (Eds.), *Validity and limitations of college student self-report data* (New Directions for Institutional Research, no. 150, pp. 7-19). San Francisco, CA: Jossey-Bass.

Bowman, N. A. (2011b). Validity of self-reported gains at diverse institutions. *Educational Researcher, 40*(1), 22-24.

Bowman, N. A. (2012, March). *Measuring student growth? The meaning and interpretation of self-reported gains*. Paper presented at the annual meeting of the American College Personnel Association, Louisville, KY.

Bowman, N. A., & Brandenberger, J. W. (2010). Quantitative assessment of service-learning outcomes: Is self-reported change an adequate proxy for longitudinal change? In J. Keshen, B. Holland, & B. Moely (Eds.), *Research for what? Making engaged scholarship matter* (Advances in Service-Learning Research, Vol. 10, pp. 25-43). Charlotte, NC: Information Age Publishing.

Bowman, N. A., & Hill, P. L. (2011). Measuring how college affects students: Social desirability and other potential biases in self-reported gains. In S. Herzog & N. A. Bowman (Eds.), *Validity and limitations of college student self-report data* (New Directions for Institutional Research, no. 150, pp. 73-85). San Francisco, CA: Jossey-Bass.

Bowman, N. A., & Schuldt, J. P. (in press). Effects of item order and response options on college student surveys. In N. A. Bowman & S. Herzog (Eds.), *Methodological advances and issues in studying college impact* (New Directions for Institutional Research). San Francisco, CA: Jossey-Bass.

Bowman, N. A., & Seifert, T. A. (2011). Can students accurately assess what affects their learning and development? *Journal of College Student Development, 52*, 270-290.

Chen, P.-S. D. (2011). Finding quality responses: The problem of low-quality survey responses and its impact on accountability measures. *Research in Higher Education, 52*, 659-674.

Conway, M., & Ross, M. (1984). Getting what you want by revising what you had. *Journal of Personality and Social Psychology, 47*, 738-748.

Gilbert, D. (2007). *Stumbling on happiness*. New York, NY: Vintage.

Goethals, G. R., & Reckman, R. F. (1973). The perception of consistency in attitudes. *Journal of Experimental Social Psychology, 9*, 491-501.

Gonyea, R. M., & Miller, A. (2011). Clearing the AIR about the use of self-reported gains in institutional research. In S. Herzog & N. A. Bowman (Eds.), *Validity and limitations of college student self-report data* (New Directions for Institutional Research, no. 150, pp. 99-111). San Francisco, CA: Jossey-Bass.

Gosen, J., & Washbush, J. (1999). Perceptions of learning in TE simulations. *Developments in Business Simulation & Experiential Learning, 26*, 170-175.

Harter, S. (1999). *The construction of the self: A developmental perspective.* New York, NY: Guilford.

Hess, J. A., & Smythe, M. J. (2001). Is teacher immediacy actually related to student cognitive learning? *Communication Studies, 52*, 197-219.

Higher Education Research Institute. (2011). *2011-2012 College Senior Survey*. Retrieved from http://www.heri.ucla.edu/researchers/instruments/FUS_CSS/2012CSS.PDF

Higher Education Research Institute. (2013). *Survey instruments, codebooks, & participation history*. Retrieved from http://www.heri.ucla.edu/researchersToolsCodebooks.php

Kitayama, S., & Cohen, D. (2009). *Handbook of cultural psychology*. New York, NY: Guilford.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology, 5*, 213-236.

Kuh, G., & Ikenberry, S. (2009). *More than you think, less than we need: Learning outcomes assessment in American higher education*. Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment.

Lee, J. J. (2002). Religion and college attendance: Change among students. *Review of Higher Education, 25*, 369-384.

Markus, G. B. (1986). Stability and change in political attitudes: Observed, recalled and explained. *Political Behavior, 8*, 21-44.

Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review, 98*, 224-253.

McFarland, C., & Ross, M. (1987). The relation between current impressions and memories of self and dating partners. *Personality and Social Psychology Bulletin, 13*, 228-238.

National Survey of Student Engagement. (2013). *National Survey of Student Engagement 2013* (U.S. English Version). Retrieved from http://nsse.iub.edu/pdf/survey_instruments/2013/NSSE%202013%20US%20English.pdf

Nisbett, R. E. (2003). *The geography of thought: How Asians and Westerners think differently…and why.* New York, NY: Free Press.

Nisbett, R. E., & Wilson, T. D. (1977a). Telling more than we know: Verbal reports on mental processes. *Psychological Review, 84*, 231-259.

Nisbett, R. E., & Wilson, T. D. (1977b). The halo effect: Evidence of unconscious alteration of judgments. *Journal of Personality and Social Psychology, 35*, 250-256.

Pace, C., & Friedlander, J. (1982). The meaning of response categories: How often is occasionally, often, and very often? *Research in Higher Education, 17*, 267-281.

Pascarella, E. T. (2001). Using student self-reported gains to estimate college impact: A cautionary tale. *Journal of College Student Development, 42*, 488-492.

Pascarella, E. T., Pierson, C. T., Wolniak, G. C., & Terenzini, P. T. (2004). First-generation college students: Additional evidence on college experiences and outcomes. *Journal of Higher Education, 75*, 249-284.

Pew Research Center. (2012). *Assessing the representativeness of public opinion surveys.* Washington, DC: Author.

Pike, G. R. (1993). The relationship between perceived learning and satisfaction with college: An alternative view. *Research in Higher Education, 34*, 23-40.

Pike, G. R. (1995). The relationship between self reports of college experiences and achievement test scores. *Research in Higher Education, 36*, 1-21.

Pike, G. R. (1996). Limitations of using students' self-reports of academic development as proxies for traditional achievement measures. *Research in Higher Education, 37*, 89-114.

Pike, G. R. (1999). The constant error of the halo in educational outcomes research. *Research in Higher Education, 40*, 61-86.

Pohlmann, J., & Beggs, D. (1974). A study of the validity of self-reported measures of academic growth. *Journal of Educational Measurement, 11*, 115-119.

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

Porter, S. R. (2013). Self-reported learning gains: A theory and test of college student survey response. *Research in Higher Education, 54*, 201-226.

Ross, M. (1989). Relation of implicit theories to the construction of personal histories. *Psychological Review, 96*, 341-357.

Seifert, T. A., & Asel, A. M. (2011). The tie that binds: The role of self-reported high school gains in self-reported college gains. In S. Herzog & N. A. Bowman (Eds.), *Validity and limitations of college student self-report data* (New Directions for Institutional Research, no. 150, pp. 59-72). San Francisco, CA: Jossey-Bass.

Sitzmann, T., Ely, K., Brown, K. G., & Bauer, K. N. (2010). Self-assessment of knowledge: A cognitive learning or affective measure? *Academy of Management Learning & Education, 9*, 169-191.

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

Sternberg, R. J., & Detterman, D. K. (1986). *What is intelligence? Contemporary viewpoints on its nature and definition*. Norwood, NJ: Ablex.

Tourangeau, R., Rips, L., & Rasinski, K. (2000). *The psychology of survey response*. New York, NY: Cambridge University Press.

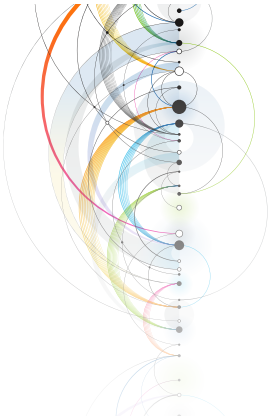
Triandis, H. C. (1989). The self and social behavior in differing social contexts. *Psychological Review, 96*, 506-520.

Tyree, T. (1998). *Designing an instrument to measure socially responsible leadership using the social change model of leadership development*. Unpublished doctoral dissertation, University of Maryland-College Park.

Walvoord, B. E. (2010). *Assessment clear and simple* (2nd ed.). San Francisco, CA: Jossey-Bass.

Wilson, T. D. (2002). *Strangers to ourselves*. Cambridge, MA: The Belknap Press of Harvard University Press.

Zwerling, L. S., & London, H. B. (Eds.) (1992). *First-generation students: Confronting the cultural issues* (New Directions for Community Colleges, no. 80). San Francisco, CA: Jossey-Bass.



Abstract

Implementation fidelity assessment provides a means of measuring the alignment between the planned program and the implemented program. Unfortunately, the implemented program can differ from the planned program, resulting in ambiguous inferences about the planned program's effectiveness (i.e., it is uncertain if poor results are due to an ineffective program or poor implementation). We demonstrate how inclusion of implementation fidelity in the outcomes assessment process increases the validity of inferences about program effectiveness and, ultimately, student learning. Although our didactic discussion of implementation fidelity focuses on its importance to assessing student affairs programming, the concepts and process are applicable to academic programs as well.

AUTHORS

Jerusha J. Gerstner, M.A.
James Madison University

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

Measuring the Implementation Fidelity of Student Affairs Programs: A Critical Component of the Outcomes Assessment Cycle

What is Implementation Fidelity and Why is it Important?

Implementation fidelity has been discussed in many domains (e.g., K-12 education, health, psychology). As a result, numerous definitions of implementation fidelity exist. The general definition provided by O'Donnell (2008) is "the determination of how well an intervention is implemented in comparison with the original program design during an efficacy and/or effectiveness study" (p. 33). Specific to the student affairs context, implementation fidelity examines the extent to which the planned student affairs program matches the implemented program. That is, student affairs programs (or any educational program) should be designed thoughtfully to meet particular learning and development outcomes. However, as Berman and McLaughlin (1976) noted, "The bridge between a promising idea and the impact on students is implementation, but innovations are seldom implemented as intended" (p. 349). Importantly, research has shown that programs implemented with high fidelity have more of an impact with respect to program outcomes than those with low fidelity (e.g., Durlak & DuPre, 2008). Thus, higher education practitioners and instructors need to ask themselves, "Are students receiving the *planned* program?"

Deviations from the planned program may involve excluding critical program components or curriculum, shortening program sessions or classes, changing the mode of program delivery, or adding extraneous information or activities (Ball & Christ, 2012). Program deviation or drift may occur for many reasons, including poor training of program implementers (e.g., instructors, facilitators, interventionists), lack of motivation of implementers, or insufficient time provided for program components (Century, Cassata, Rudnick, & Freeman, 2012; Durlak & DuPre, 2008; Lane, Bocian, MacMillan, & Gresham, 2004). Drift "refers to the unplanned, gradual altering of the implementation of an intervention by the interventionist" (Hagermoser Sanetti & Kratochwill, 2009a, p. 452). We agree with

CORRESPONDENCE

Email
gerstnjj@dukes.jmu.edu

Hagermoser Sanetti and Kratochwill that some flexibility in program implementation should be tolerated, but "such flexibility does not justify, however, an acceptance of interventionist drift, which may result from a host of factors" such as "forgetting intervention components, having limited resources, [or] believing the intervention requires too much response effort" (p. 452). Implementation fidelity assessment allows for a direct evaluation of the degree of program drift. If the program is not implemented as planned, it should not be surprising when program outcomes are not achieved.

Informal conversations with student affairs professionals, numerous consultations regarding assessment of student affairs programs, and observations of professional presentations on practice at conferences suggest that student affairs professionals are not asking themselves this implementation fidelity question, which aligns with similar observations in other domains (e.g., Cochrane & Laux, 2008; Hagermoser Sanetti & Kratochwill, 2009a). If the question is being asked, implementation fidelity results appear to be neither gathered nor reported; thus, the alignment between the planned and implemented program is not known. This lack of fidelity information greatly limits interpretation of outcomes assessment results and, ultimately, evaluation of the planned program (Ball & Christ, 2012). For instance, if an outcome measure is mapped directly to an objective and students are performing poorly on this measure, it could be inferred that students are not meeting this objective as a function of the planned program. However, if the programming aligned with this objective is not implemented as planned, the outcome measure reveals nothing about the efficacy of the planned program, because the planned program was not administered. In fact, the planned program may impact student learning and development in a powerful way if implemented correctly. The combination of low implementation fidelity and the lack of its assessment can result in changing or terminating a program that would be effective if implemented as planned (Hagermoser Sanetti & Kratochwill, 2008).

Implementation fidelity examines the extent to which the planned student affairs program matches the implemented program.

Although implementation fidelity has been a topic of interest and research in health-related fields (e.g., Breitenstein et al., 2010; Garvey, Julion, Fogg, Kartovil, & Gross, 2006) and K-12 education (e.g., Ball & Christ, 2012; Cochrane & Laux, 2008; Hagermoser Sanetti & Kratochwill, 2009a), a review of several highly-esteemed books focused on assessment in higher education and student affairs uncovers no mention of implementation fidelity (American College Personnel Association [ACPA], 2006; Bresciani, Gardner, & Hickmott, 2009; Erwin, 1991; Schuh, 2009; Schuh & Upcraft, 2001; Upcraft & Schuh, 1996). Moreover, authors, such as Shutt, Garrett, Lynch, and Dean (2012), have provided recommendations regarding how best practice, with respect to student affairs programs, centers on the assessment process: "In essence, then, the intentional use of the assessment process itself is what constitutes best practice" (p. 71). We could not agree more and echo their call for empirically-based programs and curriculum. However, the importance of implementation fidelity data for making valid inferences about program effectiveness was not stressed, much less was the process of collecting and using fidelity data to evaluate program efficacy. Nonetheless, there is clearly a place for implementation fidelity assessment within all outcomes assessment processes. Despite the lack of coverage in the higher education assessment literature, the concept of implementation fidelity is analogous to "process" or "implementation evaluation" discussed in the program evaluation literature (e.g., Patton, 1997; Posavac & Carey, 1997; Weiss, 1998). In addition, we applaud Aiken-Wisniewski et al. (2010) for not only discussing the concept of implementation fidelity (termed "process/delivery outcomes") but also noting the importance of gathering implementation data in their *Guide to Assessment in Academic Advising*. Unfortunately, practitioners in units other than advising may be unaware of this document and its recommendations regarding implementation fidelity assessment.

Three possible reasons why implementation fidelity is not assessed center on untested assumptions, lack of understanding of implementation fidelity, and lack of guidance on the practice of collecting and using implementation fidelity data. First, practitioners may assume the program "on paper" is implemented as planned (e.g., Hagermoser Sanetti & Kratochwill, 2009a; O'Donnell, 2008). Namely, practitioners may assume implementation fidelity is high because program implementers *should* present the program exactly as directed. However, this is an assumption that needs to be tested, as research indicates this assumption is often wrong (e.g., Ball & Christ, 2012; Durlak & DuPre, 2008; Hagermoser Sanetti & Kratochwill, 2008, 2009b; Lane et al., 2004). Second, practitioners may not understand that low fidelity can

attenuate program effectiveness. Third, even if practitioners are concerned about program implementation and understand the impact of low implementation fidelity, they may not engage in fidelity assessment because they do not understand how to assess the alignment of the planned and implemented programs. These barriers to implementation fidelity assessment align with noted barriers in other, related domains. More specifically, in the domain of school psychology, the following barriers were uncovered with respect to collecting implementation fidelity data: lack of general knowledge of implementation fidelity, lack of guidelines on procedures to collect these data, lack of resources, and lack of requirements to collect these data (Cochrane & Laux, 2008; Hagermoser Sanetti & DiGennaro Reed, 2012).

Given the push for accountability in higher education (U.S. Department of Education, 2006) and the assessment, evaluation, and research standards established for student affairs (ACPA, 2006), we propose that the measurement of implementation fidelity is past due in higher education. Moreover, high quality program assessment must incorporate implementation fidelity into the outcomes assessment process. Our goals in this article are to explicate implementation fidelity's place within the outcomes assessment cycle, to provide insight into quantifying fidelity, and to provide an example of how implementation fidelity was used to strengthen a student affairs program on our campus.

Higher education practitioners and instructors need to ask themselves, "Are students receiving the planned program?"

Implementation Fidelity in the Outcomes Assessment Cycle

The Typical Outcomes Assessment Cycle

The outcomes assessment cycle is used to evaluate how well programming functions with respect to meeting student learning and development objectives. Many of these cycles include the following six steps: establishing objectives/outcomes, mapping programming to these objectives/outcomes, selecting or designing measures of the outcomes, collecting outcomes data, analyzing and maintaining outcomes data, and using outcomes information (e.g., ACPA, 2006; Aiken-Wisniewski et al., 2010; Bresciani et al., 2009; Erwin, 1991; Suskie, 2009). First, practitioners must establish program objectives. These objectives outline intended outcomes of the program: what students should be able to know, think, or do as a result of participating in the program. Objectives provide a clear, detailed presentation of the program's purpose. Second, various programming components are developed to align with the stated objectives. These program components can be conceptualized as treatments that should result in the particular outcomes stated in the objectives. The intentional creation and mapping of programming to objectives is a critical part of the outcomes assessment cycle. Third, outcome measures are selected or designed to quantify whether students are meeting the objectives after being exposed to programming. During the fourth and fifth steps, outcomes data are collected and then analyzed. Finally, the outcomes assessment results are used to evaluate program effectiveness, with a specific focus on making informed changes to the programming components revealed to be suboptimal.

Incorporating Implementation Fidelity into the Assessment Cycle

In the standard assessment cycle, one never evaluates whether the planned program was implemented. In fact, the term "black box" has been used to characterize the disguised nature of any information regarding implementation of the program in standard outcomes-based assessment (Mowbray, Holter, Teague, & Bybee, 2003; Nelson, Cordrary, Hulleman, Darrow, & Sommer, 2012). Without implementation fidelity assessment, nothing is known about what occurred during the program, only what was planned—which could be radically different from the actual implemented program. That is, in absence of fidelity data, one is assessing the effectiveness of an unknown program (i.e., a black box). To open this black box, we advocate adding implementation fidelity into the assessment cycle (see Figure 1).

Researchers have proposed key components of implementation fidelity assessment (e.g., Dane & Schneider, 1998; Hagermoser Sanetti & Kratochwill, 2009a; Hulleman & Cordray, 2009; Mihalic, 2002; O'Donnell, 2008). However, none of these researchers focused specifically on assessing programming in student affairs. After reviewing and integrating the literature, we outlined five implementation fidelity components, with a specific focus on aligning these components with student affairs programming: program differentiation, adherence, quality, exposure, and responsiveness. Each component is defined in Figure 2, along with a means of assessing it. An understanding of implementation fidelity and its place in the assessment cycle

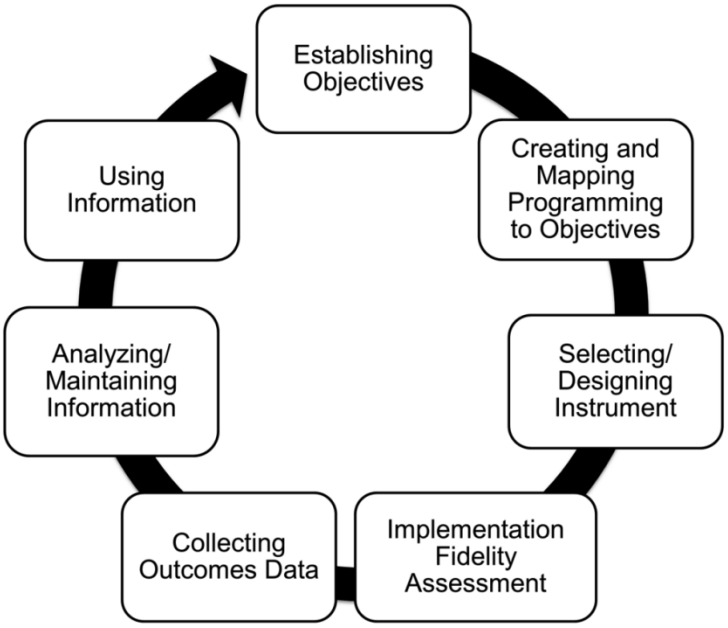


Figure 1. Outcomes assessment with implementation fidelity assessment included.

is best facilitated by a case study of implementation fidelity assessment. We therefore offer an example of implementation fidelity assessment from a large, multi-faceted student affairs program on our campus.

Gathering and Using Implementation Fidelity Data: Transfer Orientation Example

Three possible reasons why implementation fidelity is not assessed center on untested assumptions, lack of understanding of implementation fidelity, and lack of guidance on the practice of collecting and using implementation fidelity data.

Transfer Student Orientation (TSO) is a one-day program that occurs the summer prior to the start of fall classes designed to help transfer students adjust to the campus community. Approximately 650 incoming transfer students attend one of four identical days of TSO programming. TSO programming was intentionally created to meet three objectives: increase academic requirements knowledge (ARK), increase resource knowledge (RK), and increase social acclimation (SA). Throughout the day, students attend programming aligned with these objectives. It is important to note that given the wide scope of TSO, many programming aspects are necessarily implemented by staff outside of the Orientation Office.

Outcomes Assessment Process

The three TSO objectives have outcome measures mapped to them. Data from the three measures are collected before and after TSO. A matched pre- to posttest design is used to assess growth for each objective. In summer 2011, 441 transfer students provided responses to all items on the pretest and posttest.

Although valuable information was obtained through the outcomes assessment process (i.e., which objectives were or were not met), informed program changes could not be made using only the outcomes assessment data. For instance, it was unclear why students were meeting the ARK objective better than the RK objective. Were the planned program features associated with the RK objective administered with high quality for the intended duration, implying this programming simply did not "work"? Or were we observing these findings because the program was implemented with low fidelity? Given that the administered program was a black box, we could not draw many conclusions about the effectiveness of the planned program. However, incorporating implementation fidelity into the assessment cycle enabled stakeholders to make programmatic decisions that could not be made with outcomes data alone.

Implementation Fidelity Assessment Process

Implementation fidelity can be easily assessed by creating and completing a fidelity checklist (Swain, Finney, & Gerstner, 2013). Thus, an implementation fidelity checklist was developed for TSO to assess the five components of fidelity outlined in Figure 2. The checklist mapped a column of program objectives to a column of program features (i.e., program differentiation). Next to the column of program features, the planned duration of the feature was listed along with a space to record the actual duration. The next column was used to record adherence for each program feature ("yes" or "no"). The final column included a quality scale (1=*Low* to 5=*High*) so each implemented program feature could be rated for quality.

Without implementation fidelity assessment, nothing is known about what occurred during the program, only what was planned—which could be radically different from the actual implemented program.

Data were collected using this checklist in two ways. First, three university staff affiliated with the program posed as students and audited TSO. During the day-long program, they recorded their ratings on this checklist. Second, two implementers of the various program features rated their own adherence, duration, and quality. In addition, we collected data from students regarding their responsiveness. Specifically, we added a question regarding responsiveness (*How attentive were you throughout the day?*) on the posttest, which also included outcome measures.

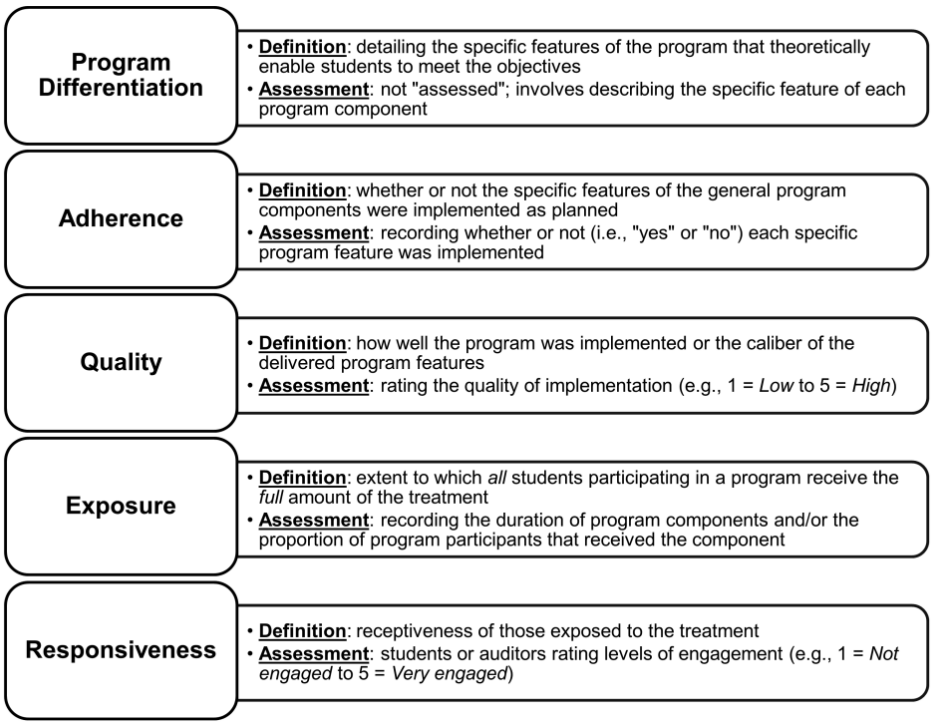


Figure 2. Implementation fidelity components: Definitions and assessment.

Interpreting Implementation Fidelity Data

In order to facilitate practitioners gathering and using implementation fidelity data, we expound on the definition and measurement of the five components of implementation fidelity within the context of TSO implementation fidelity data.

Program differentiation. The first component of implementation fidelity, *program differentiation*, involves detailing specific features of the program that theoretically enable students to meet program objectives (Dane & Schneider, 1998; Mihalic, 2002; Sheridan, Swanger-Gagné, Welch, Kwon, & Garbacz, 2009; Swain et al., 2013). As noted above, the TSO programming developed intentionally to enable students to meet each objective was specified in the fidelity checklist. For example, the program component of University Welcome, mapped to the SA objective, was broken down by stakeholders into the specific features intended to enable students to meet this objective (e.g., speech by the university president,

icebreakers in small groups). The act of program differentiation offered the stakeholders an opportunity to articulate their understanding of the link between the program outcomes and the program itself. That is, clarification of and commitment to the program objectives and programming was greatly facilitated by this differentiation process.

As noted in Figure 2, this component of implementation fidelity (unlike others) is not "assessed"; however, it is the *most* fundamental aspect of fidelity assessment. That is, program differentiation defines the program in the most specific way possible, which enables one to assess whether those features actually occurred (i.e., "adherence") and evaluate their quality (i.e., "quality"). If specific program features cannot be discerned, fidelity assessment is impossible.

Adherence. The second component of implementation fidelity is *adherence*, which addresses whether or not specific program features were implemented (Dane & Schneider, 1998; Swain et al., 2013). In the education literature, adherence is often labeled "opportunity to learn" (e.g., Boscardin et al., 2005; Gee, 2003). Practitioners need to determine whether students had the opportunity to acquire skills and knowledge needed to meet the stated outcomes. Although Suskie (2009), in her book on higher education assessment, stressed the importance of presenting opportunities to learn, she never suggested one should evaluate if those opportunities were provided as planned (i.e., assessment of implementation fidelity was not a component of her outcomes assessment process).

There are four common methods of assessing adherence: auditors of the "live" program, videotapes of the program that are later examined, program implementers, and/or an evaluation of presentation materials.

As noted above, adherence can be easily assessed using a checklist (Cochrane & Laux, 2008; Swain et al., 2013). There are four common methods of assessing adherence: auditors of the "live" program, videotapes of the program that are later examined, program implementers, and/or an evaluation of presentation materials. The first, and the most objective and valid, method to assess adherence is through the use of auditors of the live program (Cochrane & Laux, 2008). This method was employed during TSO. Auditors attended programming (undercover) as participants and indicated whether specific program features were implemented as planned (i.e., recorded opportunity to learn as "yes" or "no" on the checklist). This method allowed auditors to experience the program as "students." Readers should realize this approach could be resource-heavy, especially for long programs. The second method involves videotaping the program and having someone rate adherence by watching the videotape. This method may facilitate using a greater number of raters; however, the videotape may not allow an authentic representation of the actual program. Also, the presence of a camera could change the program's dynamic. Another useful method of assessing opportunity to learn is by asking program implementers to indicate their adherence to specific program features (Breitenstein et al., 2010). This approach was also employed during TSO. Gathering adherence data from implementers and auditors provides inter-rater reliability data (i.e., consistency in ratings from auditors and implementers). The assessment of inter-rater reliability is important, as some research has found that self-ratings indicate higher fidelity when compared to ratings from independent observers (Hagermoser Sanetti & Kratochwill, 2009a; O'Donnell, 2008), whereas other research has found accurate ratings from implementers (Hagermoser Sanetti & Kratochwill, 2009b). If it can be shown that implementers and auditors provide the same implementation fidelity data, then auditors would not be needed. Finally, if a program involves the presentation and discussion of informational materials (e.g., handouts), an examination of these materials can serve as a crude measure of adherence (Lane et al., 2004). Although not an ideal approach, this may be the only possible method for assessing adherence when a program audit or videotaping is not possible (e.g., private setting, lack of time) or if implementers will not participate in assessing fidelity (Cochrane & Laux, 2008).

With respect to TSO, the auditor and implementer adherence ratings were identical. Importantly, both auditors and implementers noted specific program features that were not executed. This finding was extremely valuable, as it indicated that implementers were willing to report their lack of coverage of program features and did so accurately. Moreover, the implementers indicated that simply engaging in rating their adherence to specific features served as an additional reminder of the content intended to be covered in the program. As noted, programs can drift unintentionally from the intended features. Requiring program implementers to review a list of program features and then indicate whether they implemented those features communicates the importance of executing the program as planned and can

protect against program drift. In addition, the process of gathering implementer adherence ratings may reduce time needed to retrain implementers (Durlak & DuPre, 2008).

Quality. The third component of implementation fidelity, *quality*, assesses the caliber of delivered program features (Dane & Schneider, 1998; Mihalic, 2002; Swain et al., 2013). With respect to TSO and higher education programming more generally, quality is an essential component of implementation fidelity. Implementers could deliver all specific program features (i.e., high adherence), yet low quality prevents the planned program from being administered fully. Anyone who has attended a presentation where information was presented quickly or unclearly can attest to the importance of assessing the quality of implementation. Although Schuh and Upcraft (2001) and Suskie (2009) mention the importance of developing high-quality student-centered programs and note that quality of presentation skills (i.e., "presenter effectiveness") could impact the functioning of the program, they focus only on this one narrow component of quality. Moreover, they fail to discuss how to measure presenter effectiveness or how to couple these data with the outcomes assessment to inform program changes. We believe the assessment of quality should include the concept of presenter effectiveness addressed by Schuh and Upcraft and Suskie when appropriate, but it should be defined widely enough to accommodate programs without a presenter. Whereas the concept of presenter effectiveness would be irrelevant for a student affairs program targeted at weight loss, the quality of the implemented program features (e.g., exercise regime) could be rated (e.g., exercises completed too quickly, without much effort, with poor form), thus providing useful information regarding program implementation.

Requiring program implementers to review a list of program features and then indicate whether they implemented those features communicates the importance of executing the program as planned and can protect against program drift.

Similar to adherence, quality can be rated by auditors and/or implementers (Swain et al., 2013). In the case of TSO, every specific program feature that received a "yes" for adherence was rated for quality (e.g., 1 = *Low* to 5 = *High*) by the auditors and implementers. A specific feature received a low quality rating if the feature was addressed, but not well. For example, TSO has an icebreaker activity intended to increase students' sense of belonging to campus. If the icebreaker activity occurred (i.e., adherence) but group facilitators did not present the activity in an engaging manner, then the students received the program feature but with poor quality.

With respect to TSO, the auditors and implementers were in agreement for most of the quality ratings. Although many features were adhered to, there was a range in quality. Fortunately, many features garnered high quality ratings; however, there were also some low ratings. As discussed below, these fidelity data helped to explain some unfavorable outcome results.

Exposure. The fourth component of implementation fidelity is *exposure*, which assesses the extent to which all students participating in a program receive the full treatment (Carroll et al., 2007; Dane & Schneider, 1998; Swain et al., 2013). In addition to detailing each program feature, program differentiation specified the planned duration of the program components. With respect to student affairs programming, practitioners intend for students to receive a "full dose" of each program component, but that does not always occur. If the planned 50-minute program component receives only 20 minutes of time, students are not being exposed to the "full treatment." Thus, students may not have the opportunity to learn to the extent intended by the program. With respect to TSO, exposure was assessed by auditors recording the actual duration of each program component. All components endured for the planned amount of time, providing confirmation that students had the opportunity to be exposed to the intended, "full dose" of programming.

In addition to assessing the duration of programming, one can also assess whether everyone was exposed to each aspect of the program. We would not expect positive outcomes assessment results if half of the participants "skipped" the programming aligned with the objective. Thus, even if the program was presented for the intended duration with high quality, the programming may appear ineffective if data from program attendees and those who skipped the program were analyzed together.

With respect to TSO, plans have already been established to further assess exposure. In the future we will ask students whether they attended various optional aspects of TSO. These attendance data will allow the outcomes data to be analyzed separately for those who did and did not attend optional programs. This type of analysis is important because if

attending optional aspects of programming has a strong impact on program outcomes, it may be beneficial to make that programming mandatory in future years. That is, exposure data can help highlight which combination of programming components are most effective in meeting outcomes, which can assist in the allocation of resources when implementing the program in the future.

Responsiveness. The final component of implementation fidelity is *responsiveness*, which addresses the receptiveness of those exposed to the treatment (Dane & Schneider, 1998; Swain et al., 2013). If students are not engaged with the TSO program, it does not matter whether the implementers deliver all the planned program features in a high quality manner for the intended duration. Students will not be impacted by a high quality program if they are disengaged. Thus, assessing responsiveness, rather than making the assumption that the program is being offered to a fully captive audience, can help illuminate why well-implemented programs may be associated with poor outcomes assessment results.

Thus, even if the program was presented for the intended duration with high quality, the programming may appear ineffective if data from program attendees and those who skipped the program were analyzed together.

Responsiveness can be assessed by asking students to self-report their level of attentiveness throughout the program. Another, more distal, measure of responsiveness would entail an auditor or implementer rating the responsiveness of the audience. Both assessments have their flaws. Students' self-reports of their responsiveness may be influenced by social desirability. Alternatively, an auditor may mistakenly perceive attendees as inattentive because they are looking down when in reality, they are taking notes. However, both measures of responsiveness can supply information otherwise lacking from the assessment process. These results can also be used to analyze the outcomes data separately for those who were or were not responsive.

With respect to TSO, as noted above, students indicated whether they were attentive throughout the day (1 = *Not at all*, 2 = *Somewhat*, 3 = *Very*). Fortunately, only 1.1% of students responded "not at all." A fairly comparable number of students responded either "somewhat" or "very" to the item. We tested for a possible differential effect of the programming across "responsiveness" groups and found no differential change in the three outcomes over time. That is, responsiveness did not moderate the change in outcomes assessment scores (i.e., there was no significant interaction).

Drawing Conclusions about Program Effectiveness by Combining Fidelity and Outcomes Assessment

The implementation fidelity results were used in numerous ways to strengthen the validity of inferences about the effectiveness of TSO. It is important to note that the TSO program director and program implementers were equal partners when interpreting implementation fidelity and outcomes assessment results and when using these results to strengthen TSO for subsequent years. This equal partnership, which had not been present in the past, was facilitated by the implementers' participation in fidelity assessment.

The implementation fidelity results coupled with the assessment results uncovered findings that neither set of results could have yielded independently. When fidelity and outcomes assessment results are combined, there are four possible scenarios that could occur (McIntyre, Gresham, DiGennaro, & Reed, 2007; Swain et al., 2013). All four scenarios presented in Figure 3 were evidenced in the assessment of TSO and we provide examples of each to model interpretation of such findings. Importantly, the combination of implementation fidelity and outcomes assessment results informed modifications to the programming components and allocation of resources.

High fidelity and favorable outcomes. Some results reflected high levels of implementation fidelity coupled with favorable outcomes assessment results (scenario 2 in Figure 3). For example, the SA objective has numerous specific program features, such as the University Welcome and Peer Discussion Groups, and the auditors observed and reported that all specific features were presented and in a fairly high-quality manner. Moreover, outcomes assessment results revealed a significant increase for the SA outcome measure from pre- to posttest. Thus, the fidelity results suggested the increase on the outcome measure might be a function of TSO programming.

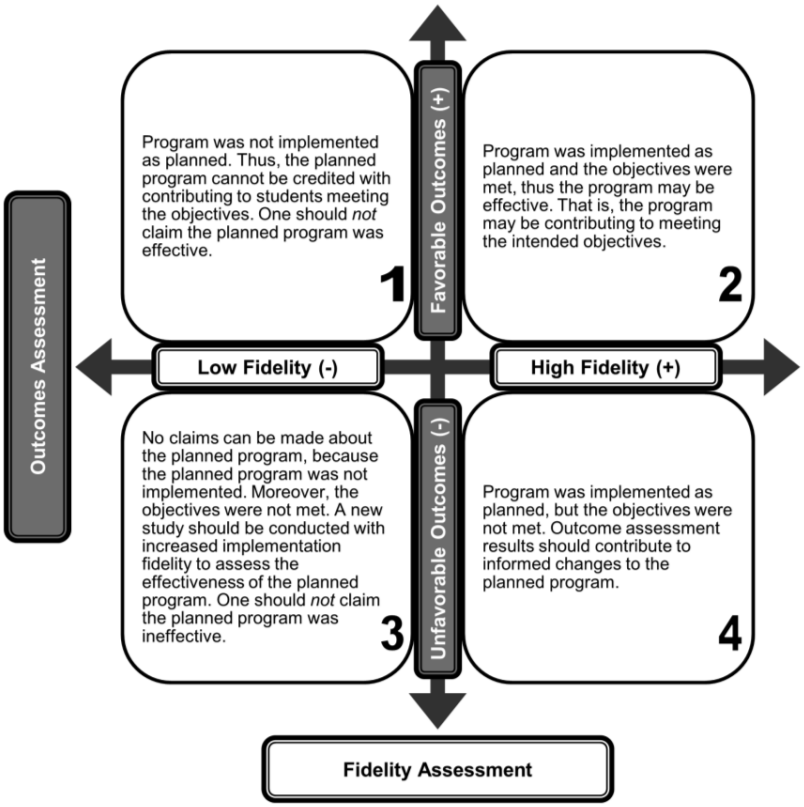


Figure 3. Four scenarios resulting from pairing implementation fidelity assessment results with outcomes assessment results. This figure is general with respect to research design; it does not assume a true experiment. Thus, positive outcomes assessment results do not imply program effectiveness; it simply reflects the objective was met. In quasi-experimental designs there could be several reasons other than program effectiveness that explain objectives being met.

Low fidelity and unfavorable outcomes. Some of the outcomes and fidelity assessment results aligned with scenario 3: the outcomes assessment results were poor and the fidelity assessment revealed the program had not been implemented as planned. Therefore, the obtained outcomes results were not reflective of the *planned* program. For instance, one specific program feature associated with the RK objective is explaining how and where to transfer credits. The fidelity assessment results indicated this specific program feature, although adhered to, had been delivered with extremely low quality. Not surprisingly, the outcomes assessment results indicated students did not understand the process of evaluating whether a course could transfer from another institution. By pairing implementation fidelity and outcomes assessment results, stakeholders discovered that the poor performance might be due to poor program implementation, which can be easily remedied before the next transfer orientation.

Exposure data can help highlight which combination of programming components are most effective in meeting outcomes, which can assist in the allocation of resources when implementing the program in the future.

High fidelity and unfavorable outcomes. Some results reflected fairly high levels of implementation fidelity coupled with poor outcomes assessment results (scenario 4). One specific program feature associated with the RK objective involved explaining how and where one pays tuition. Fidelity assessment results revealed this information was presented in a high-quality manner; however, students performed poorly on the outcome measure. Because the fidelity assessment results indicated this poor performance was not due to poor program implementation, additional or different types of programming may need to be developed to help students meet this objective. In short, it appears the programming in place is not working, thus resources should be allocated to replace or modify the existing programming.

Low fidelity and favorable outcomes. Finally, there were instances of low fidelity paired with good outcomes assessment results (scenario 1). Outcomes assessment results revealed students increased significantly from pre- to posttest on measures associated with ARK. From the perspective of a standard outcomes assessment cycle, one would conclude that

programming mapped to this objective may be effective in teaching students this information. However, fidelity data indicated programming mapped to this objective was *not* implemented. Thus, students were evidently learning this information elsewhere. Fortunately, during the fidelity assessment the auditors noted that the information was mentioned in another (albeit unplanned) programmatic component. Absent fidelity information, one would have wrongly attributed the favorable outcomes assessment results to the planned programming. Given the success in presenting this information in the unplanned, alternative programming component, the program director and implementers decided to adjust the program to reflect this change (i.e., no longer expend resources on the original programming but instead on the alternative programming).

Implications for Practice and Suggestions

In order to equip decision-makers with the necessary data to make informed decisions, implementation fidelity data must be presented and used to interpret outcomes assessment results.

The lack of implementation fidelity data challenges valid inferences and decision making regarding program impact (i.e., internal validity), as a lack of student learning and development could be due to poor implementation for which no data are available to aid administrators making program-related decisions (Ball & Christ, 2012; Durlak & DuPre, 2008). In turn, it is extremely difficult (if not impossible) to make informed, data-based decisions about resource allocation. Moreover, the lack of implementation fidelity data compromises conclusions concerning the replication and generalization of program effects (i.e., external validity; Swanson, Wanzek, Haring, Ciullo, & McCulley, 2013). Finally, lack of fidelity data makes evaluation of the properties of outcome measures ambiguous. Understanding the properties of outcomes measures is critical, as a measure may appear more difficult if students did not have the OTL (i.e., low implementation fidelity). That is, students will have trouble answering items correctly if they were not taught the material (i.e., did not have the OTL), thus making the measure appear more difficult than it would be if the students had the OTL (as intended). This failure to assess fidelity could result in practitioners discarding a high-quality measure that would have functioned properly (i.e., appeared adequate and not overly challenging) if students had the OTL (Coleman, Kaliski, & Huff, 2012; Huff & Ferrara, 2010; Polikoff, 2010). Thus, in order to equip decision-makers with the necessary data to make informed decisions, implementation fidelity data must be presented and used to interpret outcomes assessment results.

Allowing practitioners to allocate the necessary time and resources (both financial and human) to fidelity assessment is critical. Of course, this may result in assessing fewer programs each year, but that is weighed against having more accurate assessment of these programs.

Given the importance of implementation fidelity data, how do we engage higher education practitioners (e.g., faculty, administrators, staff) in the practice of gathering these data? We offer three suggestions to increase the practice of evaluating implementation fidelity and overcoming barriers. First, practitioners must be educated about the concept of implementation fidelity and its importance for evaluating program effectiveness. Research has shown that those trained in implementation fidelity are more likely to perceive it as important and engage in its measurement (Cochrane & Laux, 2008). Articles appearing in higher education, student affairs, or assessment journals that explicate the concept and model use of these data could increase awareness and support practitioners engaging in this practice for the first time. Participation in listservs, conferences, or other activities that focus on making empirically-based decisions regarding program effectiveness is also advised (Hagermoser Sanetti & Kratochwill, 2009a).

Second, resource barriers must be minimized. That is, human and financial resources have been found to be a barrier to implementation fidelity assessment (e.g., Cochrane & Laux, 2008; Hagermoser Sanetti & DiGennaro Reed, 2012). Thus, allowing practitioners to allocate the necessary time and resources (both financial and human) to fidelity assessment is critical. Of course, this may result in assessing fewer programs each year, but that is weighed against having more accurate assessment of these programs.

The final suggestion addresses the barrier associated with a lack of requirement to gather implementation fidelity data (Hagermoser Sanetti & DiGennaro Reed, 2012). Although we agree with Shutt et al. (2012) that program assessment is part of best practice and thus should be engaged in *without* mandates, requiring the gathering and use of implementation fidelity data would spur engagement in this practice. Research has shown that lack of perceived value of implementation fidelity data by administrators or the system serves as a barrier to fidelity assessment (Cochrane & Laux, 2008). Thus, we urge administrators to request these

data and the explication of how these data were used to provide a more complete and accurate picture of program effectiveness.

Moreover, implementation data are just as critical for research on higher education programs as they are for internal program effectiveness studies, which aligns with Hagermoser Sanetti and DiGennaro Reed's (2012) call for the integration of implementation fidelity and outcomes data in intervention research: "Treatment integrity and student outcome data are not only important in school and clinical settings but they are also essential to drawing valid conclusions in treatment outcome research" (p. 196). Thus, these authors call for journal editors and reviewers to require implementation fidelity data, which further addresses the barrier associated with lack of reporting requirements.

Obtaining implementation fidelity data ensures the correct program is being evaluated rather than one distorted, possibly substantially, due to implementers drifting from the planned program.

Finally, one may ask, "Is implementation fidelity a part of the outcomes assessment process or part of the program development process?" We present the following two thoughts in response to this question. First, program development and outcomes assessment should never be two separate processes. Program development has always been considered a key *part* of the outcomes assessment process (e.g., ACPA, 2006; Bresciani et al., 2009; Suskie, 2009). More specifically, when engaging in the program development process, the goal is to create programming that aligns with the stated student learning and development outcomes. To remove program development from the assessment cycle (Step 2 in Figure 1) would be nonsensical—the two are necessarily integrated. In fact, it is the clear link between the objectives and the programming that is critical to any assessment process or quality programming.

Second, and given our first point, implementation fidelity is part of the outcomes assessment process. In fact, implementation fidelity strengthens the outcomes assessment cycle. For example, program differentiation essentially makes the mapping of programming to objectives more overt, thus strengthening and better integrating the first (i.e., establishing objectives) and second (i.e., creating and mapping programming to objectives) stages of the assessment cycle. During the final stage of the assessment cycle, "Use of Information," fidelity data make diagnostics, program changes, and resource allocation much easier for stakeholders. In sum, we view the outcomes assessment cycle as including the key components of program development (e.g., ACPA, 2006; Bresciani et al., 2009; Suskie, 2009) and implementation fidelity.

Conclusions

When employing the standard outcomes assessment cycle, we have observed two common (although not necessarily appropriate) conclusions are often made following unfavorable performance on an outcome measure: the measure is not functioning properly and thus cannot reflect program effectiveness; or the program needs revision or termination. If the outcome measure was meticulously selected/designed for the program and has adequate psychometric properties, poor measurement would not seem to be a likely cause of poor performance. Moreover, concluding that program revision/termination is necessary would be premature without any information as to whether or not the planned program was truly the implemented program.

Obtaining implementation fidelity data ensures the correct program is being evaluated rather than one distorted, possibly substantially, due to implementers drifting from the planned program. Moreover, given the more complete understanding of the program's functioning afforded by implementation fidelity data, more accurate changes can be made to the program (McIntyre et al., 2007). Although this article focused on a one-day student affairs program, we applaud practitioners who conduct implementation fidelity assessment for programs of longer durations and more complex outcomes (e.g., K-12 education [Boscardin et al., 2005], school psychology [Hagermoser Sanetti & Kratochwill, 2009b]).

As Terenzini and Upcraft (1996) noted, "While assessing the purported outcomes of our efforts with students is probably the most important assessment we do, it is seldom done, rarely done well, and when it is done, the results are seldom used effectively" (p. 217). Implementation fidelity assessment can help address this problem. In sum, integrating implementation fidelity and outcomes assessment can assist us all in making more informed programmatic decisions, increasing communication between program directors and implementers of programs, and ultimately meeting the needs of students by offering empirically-supported, effective programming.

References

Aiken-Wisniewski, S., Campbell, S., Nutt, C., Robbins, R., Kirk-Kuwaye, M., & Higa, L. (2010). *Guide to assessment in academic advising* (2nd ed.). (NACADA Monograph No. 23). Manhattan, KS: The National Academic Advising Association.

American College Personnel Association. (2006). *ASK standards: Assessment skills and knowledge content standards for student affairs practitioners and scholars*. Washington, DC: Author.

Ball, C. R., & Christ, T. J. (2012). Supporting valid decision making: Uses and misuses of assessment data within the context of RTI. *Psychology in the Schools, 49*, 231–244.

Berman, P., & McLaughlin, M. W. (1976). Implementation of educational innovation. *Educational Forum, 40*, 344–70.

Boscardin, C. K., Aguirre-Munoz, Z., Stoker, G., Kim, J., Kim, M., & Lee, J. (2005). Relationship between opportunity to learn and student performance on English and Algebra assessments. *Educational Assessment, 10*, 307–332.

Breitenstein, S. M., Fogg, L., Garvey, C., Hill, C., Resnick, B., & Gross, D. (2010). Measuring implementation fidelity in a community-based parenting intervention. *Nursing Research, 59*, 158–165.

Bresciani, M. J., Gardner, M. M., & Hickmott, J. (2009). *Demonstrating student success: A practical guide to outcomes-based assessment of learning and development in student affairs*. Sterling, VA: Stylus.

Carroll, C., Patterson, M., Wood, S., Booth, A., Rick, J., & Balain, S. (2007). A conceptual framework for implementation fidelity. *Implementation Science, 2*, 1–9.

Century, J., Cassata, A., Rudnick, M., & Freeman, C. (2012). Measuring enactment of innovations and the factors that affect implementation and sustainability: Moving toward common language and shared conceptual understanding. *Journal of Behavioral Health Services & Research, 39*, 1–19.

Cochrane, W. S., & Laux, J. M. (2008). A survey investigating school psychologists' measurement of treatment integrity in school-based interventions and their beliefs about its importance. *Psychology in the Schools, 45*, 499–507.

Coleman, C., Kaliski, K., & Huff, K. (2012). *Examining the relationship between opportunity to learn and difficulty statistics on exams implementing evidence-centered design*. Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, BC.

Dane, A. V., & Schneider, B. H. (1998). Program integrity in primary and early secondary prevention: Are implementation effects out of control? *Clinical Psychology Review, 18*, 23–45.

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

Erwin, T. D. (1991). *Assessing student learning and development*. San Francisco, CA: Jossey-Bass.

Garvey, C., Julion, W., Fogg, L., Kartovil, A., & Gross, D. (2006). Measuring participation in a prevention trial with parents of young children. *Research in Nursing & Health, 29*, 212–222.

Gee, J. P. (2003). Opportunity to learn: A language-based perspective on assessment. *Assessment in Education, 10*, 27–46.

Hagermoser Sanetti, L. M., & DiGennaro Reed, F. D. (2012). Barriers to implementing treatment integrity procedures in school psychology research: Survey of treatment outcome researchers. *Assessment for Effective Intervention, 37*, 195–202.

Hagermoser Sanetti, L. M., & Kratochwill, T. R. (2008). Treatment integrity in behavioral consultation: Measurement, promotion, and outcomes. *International Journal of Behavioral Consultation and Therapy, 4*, 95–114.

Hagermoser Sanetti, L. M., & Kratochwill, T. R. (2009a). Toward developing a science of treatment integrity: Introduction to the Special Series. *School Psychology Review, 38*, 445–459.

Hagermoser Sanetti, L. M., & Kratochwill, T. R. (2009b). Treatment integrity assessment in the schools: An evaluation of the treatment integrity planning protocol. *School Psychology Quarterly, 24*, 24–35.

Huff, K., & Ferrara, S. (2010, June). *Frameworks for considering item response demands and item difficulty*. Paper presented at the annual meeting of the National Conference on Student Assessment, Detroit, MI.

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

Lane, K. L., Bocian, K. M., MacMillan, D. L., & Gresham, F. M. (2004). Treatment integrity: An essential–but often forgotten–component of school-based interventions. *Preventing School Failure, 48*(3), 36–43.

McIntyre, L. L., Gresham, F. M., DiGennaro, F. D., & Reed, D. D. (2007). Treatment integrity of school-based interventions with children in the Journal of Applied Behavior Analysis 1991-2005. *Journal of Applied Behavior Analysis, 40*, 659–672.

Mihalic, S. (2002, April). *The importance of implementation fidelity*. Boulder, CO: Center for the Study and Prevention of Violence. Retrieved from www.colorado.edu/cspv/blueprints/Fidelity.pdf

Mowbray, C. T., Holter, M. C., Teague, G. B., & Bybee, D. (2003). Fidelity criteria: Development, measurement, and validation. *American Journal of Evaluation, 24*, 315–340.

Nelson, M. C., Cordrary, D. S., Hulleman, C. S., Darrow, C. L., & Sommer, E. C. (2012). A procedure for assessing intervention fidelity in experiments testing educational and behavioral interventions. *Journal of Behavioral Health Services and Research, 39*, 374–396.

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

Patton, M. Q. (1997). *Utilization focused evaluation* (3rd ed.). Thousand Oaks, CA: Sage.

Polikoff, M. S. (2010). Instructional sensitivity as a psychometric property of assessments. *Educational Measurement: Issues and Practice, 29*(4), 3–14.

Posavac, E. J., & Carey, R. G. (1997). *Program evaluation: Methods and case studies* (5th ed.). Upper Saddle River, NJ: Prentice-Hall.

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

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

Sheridan, S. M., Swanger-Gagné, M., Welch, G. W., Kwon, K., & Garbacz, S. A. (2009). Fidelity measurement in consultation: Psychometric issues and preliminary examination. *School Psychology Review, 38*, 476–495.

Shutt, M. D., Garrett, J. M., Lynch, J. W., & Dean, L. A. (2012). An assessment model as best practice in student affairs. *Journal of Student Affairs Research and Practice, 49*, 65–82.

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

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

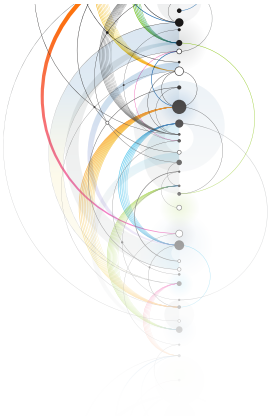
Swanson, E., Wanzek, J., Haring, C., Ciullo, S., & McCulley, L. (2013). Intervention fidelity in special and general education research journals. *The Journal of Special Education*, *47*(1), 3-13.

Terenzini, P. T., & Upcraft, M. L. (1996). Assessing program and service outcomes. In M. L. Upcraft & J. H. Schuh (Eds.), *Assessment in student affairs: A guide for practitioners* (pp. 217–239). San Francisco, CA: Jossey-Bass.

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

U.S. Department of Education. (2006). *A test of leadership: Charting the future of U.S. higher education*. *A report of the commission appointed by secretary of education Margaret Spellings*. Washington, DC.: Author.

Weiss, C. H. (1998). *Evaluation* (2nd ed.). Upper Saddle River, NJ: Prentice-Hall.



Abstract

Diversity is an increasingly important value for institutions of higher education. Yet, few measures exist to assess whether college and university faculty share in this objective and how their beliefs relate to specific aspects of their work. In this study, we gathered data from a sample of faculty at one American research university to develop a valid and reliable instrument useful for exploring how commitments to diversity are reflected in teaching, research and service. The resulting instrument, ACES, assesses four factors: (a) Attitude towards diversity, (b) Career activities and professional norms, (c) Environment conducive to diversity, and (d) Social interactions with diverse groups. Evidence for the validity and reliability of the scores produced by ACES is presented. How this psychometrically-sound instrument might benefit higher education research and practice in the assessment of diversity related goals is also considered.

AUTHORS

Jennifer Ng, Ph.D.
University of Kansas

William Skorupski, Ph.D.
University of Kansas

Bruce Frey, Ph.D.
University of Kansas

Lisa Wolf-Wendel, Ph.D.
University of Kansas

ACES: The Development of a Reliable and Valid Instrument to Assess Faculty Support of Diversity Goals in the United States

Diversity is an increasingly vital objective in American higher education. Although past rationales for this effort to emphasize diversity have focused upon the need to affirmatively remedy legacies of discrimination and prevent historically disadvantaged groups from remaining disadvantaged, colleges and universities now commonly relate diversity to broad statements of their institutional missions. They acknowledge the educational value of diversity in enriching perspectives within classrooms and across campus, and they recognize the social value of diversity in preparing students to live in a pluralistic and multicultural democracy (McGowan, 1996; Moses & Chang, 2006; Smith, 2009).

The ability of an institution to achieve its diversity goals arguably depends upon being able to accurately determine the willingness of its individual members to support and enact those same principles. Yet, no validated measures are currently available to assess whether university faculty share their institution's stated commitments to diversity and how these varied commitments are expressed in their teaching, research, and service. The purpose of this project is to gather psychometric data from a sample of faculty at one research university, and then develop a valid and reliable instrument intended to measure faculty beliefs and professional practices related to diversity goals in higher education.

CORRESPONDENCE

Email
jcng@ku.edu

Diversity is a challenging concept to capture narrowly enough for useful analysis. This is apparent both in scholarship that seeks to conceptualize its meaning and in studies that have sought to determine its significance. Whereas researchers commonly adopt the language of diversity to address race and ethnicity, others include gender and socioeconomic status, and still others include language, disability, sexual orientation, citizenship, and

religion. These multiple uses of the umbrella term "diversity" make it difficult to compare the results of one study with another and may also suggest implicitly or explicitly a hierarchy of importance where certain types of diversity are deemed more critical or worthy of consideration than others. Emphasis on diversity may simply refer to efforts that promote harmony across diverse ideas, lifestyles, dress, and other attributes or, it could mean paying particular attention to the legal principles of fair treatment and the historic struggles for equal opportunity particular to certain diverse groups (Edelman, Fuller, & Mara-Drita, 2001; Smith, 2009).

In designing the instrument for this project, we ultimately chose to define diversity as differences of race and ethnicity, national origin, and gender. Although this conceptualization is admittedly limited, our decision was guided by two primary rationales. First, generic and undifferentiated references to diversity obscure the fact that different groups have distinct experiences, perspectives, and needs. Clearly specifying groups of interest and related questions yielded more clarity in the items developed. Secondly, institutional policies and affirmative actions such as scholarships, admissions, and target hiring, for example, are typically made on the basis of certain types of diversity, but not other types of human difference. Whether the scope of diverse groups recognized by the university should ideally be more inclusive is beyond the scope of this study as our purpose was to explore whether faculty supported the existing diversity goals of the university where they worked.

Despite the assorted meanings of diversity, Terenzini, Cabrera, Colbeck, Bjorklund, and Parente (2001) posit that researchers tend to examine diversity in three relatively distinct ways: structural, in situ, and programmatic diversity. Researchers who focus on structural diversity look at the numerical makeup and proportional mix of diverse individuals within a given setting. This approach provides quantifiable evidence of access and representation in educational settings, especially as they relate to the involvement of historically marginalized groups in society. As Baird (1990) points out, examining the "differences between these 'is' and 'should be' ratings show how closely present campus goals match the goals that people prefer….and differences among groups of respondents on their preferred goals shows how much agreement exists about institutional purposes" (p. 38).

[Universities] acknowledge the educational value of diversity in enriching perspectives within classrooms and across campus, and they recognize the social value of diversity in preparing students to live in a pluralistic and multicultural democracy.

Researchers examining in situ diversity rely on participants' reports of the frequency or nature of their interactions with others who are different from themselves. This is important because a heterogeneous mix of individuals simply sharing a common physical space may be insufficient to yield the social and educational benefits of diversity that depend also on human interaction (Hurtado, 1992; McGowan, 1996). Understanding the psychosocial development, engagement, and identification of individuals contextualized by the institutional climates where they coexist provides valuable insight into how individuals interpret their experiences and perceive relevant relationships that ultimately influence behaviors and attitudes (Kossek & Zonia, 1993).

Lastly, studies of programmatic diversity explore the impact of curriculum and coursework, professional development, and other existing or planned reforms to promote diversity (Terenzini et al., 2001). Measurement strategies using this approach evaluate the access that underrepresented students have to an institution's programs and resources; comparative retention rates for students; institutional receptivity to being accommodating and responsive; and excellence in achievement. Research on programmatic diversity can provide comprehensive awareness of existing inequities, interpretation of related data, and actions to strategically remedy such disparities in the institutional structures of a university (Bensimon, 2004).

Many measures of diversity in higher education exist, and they solicit responses from students, faculty, administrators, staff, and alumni on varied topics like general campus climate, overall satisfaction, intergroup relations, student learning and involvement, and curriculum, for example (see Association of American Colleges and Universities, 2005; Shenkle, Snyder, & Bauer, 1998; Smith, 2009; Smith, Wolf-Wendel, & Levitan, 1994). These instruments are typically generated by institutional task forces or offices of institutional research, and a closer review of select items indicates their main purpose is assessing the effectiveness of past efforts or identifying areas in need of future attention. Specifically, most available measures focus on gauging student attitudes or perceptions of campus climate. Additionally, little attention is paid

to the reliability and validity of the surveys themselves because they are primarily intended for internal use. In one notable exception, Pohan and Aguilar (2001) discussed the development of a statistically valid and reliable instrument suitable for measuring elementary and secondary educators' personal and professional beliefs about diversity. The context of K-12 schooling and teachers' work is quite unlike that of faculty in higher education, though, with faculty members who are tenured or in tenure-track positions being expected to teach, conduct research, and provide professional service (Clark, 1987). Thus, the instrument presented in this study is unique because it reflects these particular dimensions of faculty life, recognizing that faculty members embody and negotiate multiple institutional, departmental, and disciplinary norms and values in their daily work (Austin, 1990, 1994).

A heterogeneous mix of individuals simply sharing a common physical space may be insufficient to yield the social and educational benefits of diversity that depend also on human interaction.

Method

Participants

A pilot instrument consisting of 100 items was developed and assembled into an online format. Email invitations to participate in the study were sent to all tenure track, full time faculty members (n=1,205) at a large, Midwestern, public research intensive university. The study sample included 332 individuals, which represented a 28% response rate. This sample size was somewhat less than ideal, but at least larger than many recommended minima for conducting factor analysis suggested by researchers (e.g., Thompson, 2004).

Thirty-eight percent of the sample consisted of full professors (compared to 42% in the population), 35% were associate professors (compared to 33% in the population), and 26% were assistant professors (compared to 25% in the population). Women represented 47% of the sample (compared to 39% in the population), international faculty represented 14% (compared to 12% in the population), and 17% were racial/ethnic minorities (as compared to 15% in the population). Aside from the slight overrepresentation of women in the sample, the demographics of the sample are reflective of the population at the institution.

Instrument

Diversity for this instrument was defined as differences of race and ethnicity, national origin, and gender. We chose this limited definition of diversity in order to be acute with respect to our operational definition, and to mirror what many individuals treat as the typical definition of diversity. In choosing domains and constructing items for the instrument, the three approaches to operationalizing diversity identified by Terenzini et al. (2001) provided a useful framework. The structural diversity approach was reflected in attitudinal questions about the ideal composition and amount of attention that faculty thought ought to be given at a university with regard to the structural diversity of students, faculty, and administrators. The in situ contextual approach to assessing diversity with an emphasis on climate was represented by asking respondents to evaluate their interactions with individuals from diverse backgrounds and a large number of questions where respondents were asked to determine the extent to which diversity is a priority in their respective departments, university, and professional communities—domains within which faculty might encounter particular professional norms (Austin, 1990, 1994). Terenzini et al.'s third approach to assessing diversity addresses programmatic initiatives and faculty activities. Faculty engage in teaching, research, and service activities which are presumably integrally tied to the objectives of the university overall. Consequently, questions were included to consider not only whether faculty members support the university's diversity goals in principle, but also how they enact those commitments in various aspects of their actual individual and programmatic work. This framework of three assessment strategies produced questions about general attitudes and beliefs about diversity (21 items); perceptions of institutional climate for diversity, (12 items); inter-personal relationships (15 items); professional norms (9 items); research (10 items); teaching (19 items); and service (14 items). Each item on the instrument consisted of a statement which represented a perspective on diversity. Using a Likert-type format, every item was scored from 1 – *Strongly Disagree* to 5 – *Strong Agree*. A balance of items with positive and negative valences was included.

Procedures

Responses from the instrument were analyzed by means of an exploratory factor analysis (EFA) with principal components analysis (PCA) extraction and varimax rotation. EFA was chosen because our framework for item generation was a fairly informal way of organizing our thoughts for item writing. Our goal was to explore the data structure to determine an optimum factor structure, not test an a priori hypothesis about dimensionality. PCA was used for the initial extraction for the EFA because of its utility in determining an optimal number of components by using eigenvalues and a scree test (Cattell, 1966). Varimax rotation was employed to maintain orthogonality among the components, thus increasing their interpretability. Our goal was to identify those aspects of attitude towards diversity which were meaningful and independent of each other. By these means, an appropriate number of components was identified with a balance of efficiency and explanatory power for the observed data.

The instrument presented in this study is unique because it reflects these particular dimensions of faculty life, recognizing that faculty members embody and negotiate multiple institutional, departmental, and disciplinary norms and values in their daily work.

The criteria for decisions regarding the number of components in the final solution and item retention/deletion were as follows: using Cattell's (1966) scree test, an optimal number of components was identified. All components that were interpretable, based on the pattern of factor loadings for items, were retained. Items were retained if they demonstrated strong (> |0.3|) loadings on one and only one component. Any items with loadings less than |0.3| were deleted. Cross-loading items, those with loadings greater than |0.3| on multiple components, were also deleted. Our goal in these analyses were to arrive at a reduce set of very discriminating items to be included on the final instrument.

Based on these results, a final solution was determined. The item pool was then revised to eliminate any items that did not have strong loadings on any component or had strong loadings on multiple components. As stated, for the purposes of this analysis, a factor loading was considered "strong" if it was greater than or equal to |0.3|, which is a common criterion (e.g., Thompson, 2004). After the final set of retained items was identified, all items with negative valences were reverse-scored to align the direction of all items. All items were then used to create reliable, independent scales to assess the multiple dimensions of attitude towards diversity among faculty in higher education.

In addition to questions about attitudes towards diversity, a number of demographic questions such as gender, ethnicity, academic discipline, and rank were included on the instrument for the purpose of comparing groups after scales were identified. The purpose of these analyses was to help validate the dimensional structure of the items, given prior research that shows there are important differences among respondents based on race, ethnicity and gender (Conley & Hyer, 1999; D'Augelli & Herschberger, 1993; Hurtado, 1992; Kossek & Zonia, 1993) as well as time in rank and disciplinary background (Austin, 1990, 1994; Somers et al., 1998). After the creation of scales, a series of demographic analyses were conducted to statistically compare group mean differences across scale scores. These comparisons were conducted using Multivariate Analysis of Variance using the four subscale scores as dependent variables, with appropriate follow-up pairwise comparisons.

Results

Preliminary Analyses

To determine the factorability of the inter-item correlation matrix, Bartlett's (1954) test of sphericity and the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy (Kaiser, 1974) were calculated. Bartlett's test of sphericity is a chi-squared statistic which tests the null hypothesis that the population inter-item correlation matrix is an identity matrix (a square matrix with 1s for the diagonal elements, and 0s for all off-diagonal elements). If a correlation matrix is not statistically different than an identity matrix, it indicates that the variables are not substantially interrelated. This null hypothesis was rejected ($\chi 2 = 7459.94$, df=1770, $p < 0.01$). The KMO measures the extent to which the items measure a common component or components by determining their shared variance after accounting for their partial correlations. Results of this analysis indicated a very high degree of shared variability (KMO = 0.87), indicating that a factor analysis would account for a large portion of the overall

variability in the data. It was therefore determined that an exploratory factor analysis was appropriate and would provide meaningful results.

Exploratory Factor Analysis

Results of the initial solution from Principal Components Analysis indicated that four components would be appropriate for explaining the observed data. Based on this criterion, and the previously mentioned criteria of removing items with strong factor loadings on multiple dimensions, or without any strong factor loadings, a final set of 60 from the original 100 pilot items was retained. Of the 40 items removed, six item had factor loadings less than 0.3 on all four factors, and 32 others were removed for cross-loading 0.3 or greater on two or more factors. The results of the scree test for the final set of 60 items demonstrated an "elbow" after the fourth eigenvalue, indicating that these data could be efficiently summarized by four components. These first four components collectively explained 48% of the variance in observed item responses. Adding a fifth component only explained an additional 3% of the variance and made the final solution much less interpretable. Examination of the pattern of factor loadings from the exploratory factor analysis indicated a clear pattern for simple interpretation based on the four-component solution.

Questions were included to consider not only whether faculty members support the university's diversity goals in principle, but also how they enact those commitments in various aspects of their actual individual and programmatic work.

The four components were identified as (a) Attitude towards diversity (containing general attitude questions), (b) Career activities (containing research, teaching, service, and some professional norms questions), (c) Environment (containing perceptions of institutional climate for diversity), and (d) Social interactions with diverse groups (containing questions about inter-personal relationships and several items from the teaching, research and service

Table 1

Factor Loadings, Communalities, and Descriptive Statistics for Items on the Attitude Scale

Item	R.S.	Factor Loadings				Com.	Descriptive Statistics			
		A	C	E	S		M	SD	Sk.	Ku.
Hiring a more diverse faculty should be a priority at my university.		**0.869**	0.209	-0.085	0.038	0.807	3.97	1.01	-0.89	0.32
A more diverse faculty would enhance my university.		**0.844**	0.123	-0.100	0.107	0.750	4.17	0.90	-1.22	1.71
Hiring a more diverse staff should be a priority at my university.		**0.837**	0.196	-0.060	-0.034	0.744	3.85	1.00	-0.70	0.04
Creating a diverse campus environment should be a priority at my university.		**0.825**	0.186	-0.072	0.078	0.726	4.17	0.85	-1.07	1.25
Recruiting a more diverse student body should be a priority at my university.		**0.822**	0.152	-0.086	0.071	0.711	4.13	0.86	-1.08	1.43
A diverse student body enhances the educational experience of all students.		**0.803**	0.076	-0.046	0.024	0.654	4.43	0.71	-1.67	4.69
The institutional mission of my university should include an explicit statement about its commitment to diversity.		**0.760**	0.149	-0.041	0.005	0.602	4.03	0.96	-0.95	0.65
Diversity should be a factor considered in student admissions to my university.		**0.717**	0.195	-0.038	0.054	0.557	3.69	1.08	-0.92	0.34
The promotion of gender equity among faculty should be a priority at my university.		**0.711**	0.158	-0.028	0.189	0.567	4.02	0.96	-0.89	0.40
Discriminatory practices still exist in American higher education because they have been institutionalized.		**0.689**	0.035	-0.064	0.102	0.491	4.48	0.70	-1.66	4.16
The leadership of my university should be representative of the racial and ethnic diversity of the United States.		**0.652**	0.273	-0.114	0.069	0.517	3.56	1.05	-0.48	-0.20
Gender discrimination is a major contemporary problem.		**0.647**	0.119	-0.263	0.155	0.526	3.72	1.07	-0.67	-0.18
Improving access to higher education for racial and ethnic minorities is important to compensate for the historical legacy of discrimination.		**0.610**	0.210	-0.153	-0.046	0.442	3.77	1.09	-0.72	-0.24
Racial discrimination is a major contemporary problem.		**0.579**	0.164	-0.253	0.071	0.432	4.00	0.99	-1.08	0.82
Too much attention on diversity can divide the campus community.	X	**-0.554**	0.222	-0.062	0.007	0.360	3.33	1.09	-0.18	-0.75
Diversity is relevant to the future professional lives of my students.		**0.533**	0.292	0.033	-0.036	0.372	4.22	0.73	-0.92	1.73
Efforts should be made to ensure my university is welcoming of people from all backgrounds.		**0.510**	0.021	-0.022	0.179	0.293	4.58	0.65	-2.15	7.64
Regardless of students' background characteristics, everyone in the U.S. should have an equal opportunity to attend college.		**0.492**	-0.059	0.004	0.005	0.246	4.31	0.93	-1.41	1.58
Female faculty members are given preferential treatment at my university.	X	**-0.477**	0.233	-0.062	0.034	0.286	3.86	0.92	-0.73	0.45
Racial and ethnic minority faculty members are given preferential treatment at my university.	X	**-0.436**	0.022	-0.254	-0.020	0.255	3.55	1.00	-0.39	-0.37
I am sensitive to the existence of institutionalized racism.		**0.409**	0.287	-0.230	0.045	0.304	3.81	0.84	-0.79	1.01
It is important that female faculty members serve as leaders in my university and field.		**0.365**	0.203	0.073	0.284	0.261	4.10	0.77	-0.58	-0.04
The university's goal to achieve greater diversity on this campus is a responsibility shared equally by all faculty members.		**0.357**	0.122	0.025	0.095	0.152	3.61	1.16	-0.45	-0.80
I get frustrated when I cannot understand what non-native English speakers are saying.	X	**-0.331**	0.105	0.057	0.066	0.128	3.53	1.09	-0.36	-0.88

Note. R.S. = Items marked with an "X" were reverse-scored before scale scores were calculated.
Com. = Communality, Sk. = skewness, Ku. = kurtosis.

Table 2

Factor Loadings, Communalities, and Descriptive Statistics for Items on the Career Scale

Item	R.S.	Factor Loadings				Com.	Descriptive Statistics			
		A	C	E	S		M	SD	Sk.	Ku.
Racial and ethnic diversity is represented in the curriculum of my courses.		0.174	**0.838**	0.046	0.001	0.735	3.44	1.20	-0.48	-0.64
There are frequent discussions about diversity in the classes I teach.		0.213	**0.792**	-0.112	-0.043	0.688	2.89	1.30	0.08	-1.13
I strive to expand students' knowledge of racial and ethnic minority groups.		0.269	**0.763**	-0.022	-0.059	0.658	3.61	1.12	-0.40	-0.75
I explore questions related to gender in my research.		0.123	**0.730**	-0.130	0.111	0.578	2.94	1.35	0.01	-1.24
I explore questions related to race and ethnicity in my research.		0.177	**0.709**	-0.200	0.188	0.610	2.89	1.40	0.09	-1.28
Women are represented in the curriculum of my courses.		0.075	**0.688**	0.051	0.190	0.518	3.81	1.01	-0.77	0.15
Diversity is irrelevant to my research interests.	X	0.228	**-0.684**	-0.139	0.122	0.555	3.41	1.31	-0.35	-1.04
Diversity is a central component of my research agenda.		0.269	**0.682**	-0.230	0.214	0.636	2.64	1.30	0.45	-0.91
Issues of diversity are unrelated to the content of my courses.	X	0.253	**-0.633**	-0.053	-0.054	0.470	3.41	1.33	-0.43	-1.07
I regularly participate in professional development activities related to diversity on campus.		0.196	**0.544**	-0.198	0.178	0.405	2.50	1.02	0.49	-0.37
I am familiar with resources to assist in revising my curriculum so it is more inclusive of diverse perspectives.		0.129	**0.543**	-0.094	0.077	0.327	3.05	1.13	0.10	-0.86
My faculty colleagues routinely consider issues of race, ethnicity, and gender in their work.		0.162	**0.490**	0.295	-0.100	0.364	3.01	1.14	0.00	-0.88
Accrediting bodies in my field state that diversity is a priority.		0.182	**0.472**	0.180	0.148	0.310	3.65	0.92	-0.43	0.24
Increasing the participation of people from diverse backgrounds is a priority in my field.		0.266	**0.469**	0.197	0.155	0.353	3.66	0.95	-0.48	-0.17
I serve on committees that promote racial and ethnic diversity at my university.		0.099	**0.433**	-0.049	0.186	0.234	2.75	1.15	0.34	-0.87
Funding agencies in my field support research related to diversity.		-0.024	**0.338**	0.179	0.056	0.150	3.38	1.04	-0.50	-0.19

Note. R.S. = Items marked with an "X" were reverse-scored before scale scores were calculated.
Com. = Communality, Sk. = skewness, Ku. = kurtosis.

Table 3

Factor Loadings, Communalities, and Descriptive Statistics for Items on the Environmental Scale

Item	R.S.	Factor Loadings				Com.	Descriptive Statistics			
		A	C	E	S		M	SD	Sk.	Ku.
My university sets a high priority on diversity.		-0.030	-0.011	**-0.784**	0.098	0.625	3.40	0.88	-0.36	-0.04
My university supports the professional needs of racial and ethnic minority faculty members.		-0.123	0.019	**-0.775**	0.032	0.618	3.27	0.84	-0.20	-0.11
Faculty members of different races and ethnicities are treated unfairly at my university.	X	-0.263	-0.124	**0.723**	0.067	0.612	3.57	0.92	-0.53	0.47
My faculty peers are receptive to diversity issues.		0.019	0.153	**-0.714**	-0.193	0.571	3.60	0.96	-0.85	0.51
There is a lot of rhetoric about diversity at my university, but not enough action.	X	-0.279	-0.168	**0.668**	-0.061	0.557	2.74	1.02	-0.03	-0.71
Faculty members from other countries are treated unfairly at my university.	X	-0.124	-0.181	**0.653**	0.022	0.474	3.57	0.89	-0.43	0.27
My faculty colleagues are ambivalent about the importance of diversity.	X	-0.070	0.116	**0.650**	-0.156	0.464	3.21	1.06	-0.31	-0.70
My university supports the professional needs of faculty members from other countries.		0.060	-0.106	**-0.631**	0.076	0.419	3.29	0.77	-0.07	0.64
My university upholds respect for the expression of diverse perspectives.		0.081	0.042	**-0.626**	0.007	0.400	3.76	0.82	-0.96	1.50
There is a great deal of racial tension on this campus.	X	-0.129	-0.279	**0.623**	-0.037	0.485	3.80	0.85	-0.72	1.00
My university supports the professional needs of female faculty members.		-0.247	-0.094	**-0.615**	0.005	0.448	3.30	1.01	-0.47	-0.34
Faculty members in my department support the use of strategic hiring to promote diversity.		0.111	0.287	**-0.594**	-0.165	0.474	3.38	1.03	-0.49	-0.32
Female faculty members are treated unfairly at my university.	X	-0.219	-0.131	**0.552**	0.004	0.370	3.44	0.98	-0.41	-0.08
Committees to address diversity issues exist, but they get very little done.	X	0.019	0.136	**0.359**	-0.196	0.186	2.78	0.89	-0.19	-0.01

Note. R.S. = Items marked with an "X" were reverse-scored before scale scores were calculated.
Com. = Communality, Sk. = skewness, Ku. = kurtosis.

Table 4

Factor Loadings, Communalities, and Descriptive Statistics for Items on the Social Scale

Item	Factor Loadings				Com.	Descriptive Statistics			
	A	C	E	S		M	SD	Sk.	Ku.
Mentoring female students in research is an important part of my work.	0.133	0.121	-0.063	**0.796**	0.669	3.89	1.05	-0.70	-0.28
Mentoring racial or ethnic minority students in research is an important part of my work.	0.142	0.229	-0.128	**0.760**	0.666	3.54	1.15	-0.38	-0.75
Mentoring international students in research is an important part of my work.	0.066	-0.009	0.016	**0.754**	0.573	3.62	1.15	-0.47	-0.70
I assist in the recruitment of prospective female students to my academic program.	0.046	0.177	0.002	**0.711**	0.539	3.69	1.12	-0.74	-0.16
I assist in the recruitment of prospective students from racial and ethnic minority backgrounds to my academic program.	0.087	0.170	-0.049	**0.647**	0.458	3.58	1.18	-0.60	-0.58
I collaborate on research with people who are a different race or ethnicity than I am.	0.142	0.121	-0.052	**0.530**	0.319	3.69	1.18	-0.77	-0.29

Com. = Communality, Sk. = skewness, Ku. = kurtosis.

Table 5
Scale Descriptive Statistics (N=235)

Scale	Sample Item	N of Items	M	SD	Sk.	Ku.	α
Attitude towards diversity	Hiring a more diverse faculty should be a priority at my university.	24	3.98	0.59	-0.72	0.62	0.94
Career activities related to diversity	Racial and ethnic diversity is represented in the curriculum of my courses.	16	3.20	0.78	-0.05	-0.61	0.91
Environment of diversity	My university sets a high priority on diversity.	14	3.37	0.59	-0.59	1.21	0.89
Social interaction with diverse groups	Mentoring female students in research is an important part of my work.	6	3.66	0.83	-0.39	-0.25	0.82

Table 6
Correlations among Scales (N=220)

	Attitude	Career	Environment	Social
Attitude	1.00			
Career	0.50*	1.00		
Environment	-0.26*	-0.12	1.00	
Social	0.28*	0.33*	-0.14	1.00

* $p \leq .01$.

categories focused on relationship building). Given these resulting components, we refer to the instrument as ACES. Tables 1–4 contain factor loadings, communalities after rotation, and descriptive statistics for every item on the Attitude, Career, Environment, and Social scales, respectively.

Those holding a positive Attitude towards diversity goals tended to be female, untenured, and at their institution for less than 15 years.

Descriptive statistics for each of the four scales, including an example item, number of items, mean across scale items, SD, skewness, kurtosis, and internal reliability estimates (coefficient alpha) are contained in Table 5. Table 6 presents a pattern of moderate to low correlations among the four scales.

Construct Validity Analysis

A series of statistical analyses were conducted to explore whether scores on any scale were related to particular demographic characteristics of faculty. Descriptive statistics, obtained values and effect sizes are shown for the statistically significant analyses in Table 7. All results significant at the 0.05 level are shown, but due to the large number of statistical significance tests conducted, only those analyses with p-values less than or equal to 0.001 should be considered. It should be further noted that some of these factors may represent overlapping sources of variability (that is, the results of some significance tests may be confounded with others). Effect sizes are reported and interpreted using Cohen's d and eta-squared (e.g., Keppel & Wickens, 2004). Levene's Test for Homogeneity of Variance was conducted for all analyses and the assumption of equal variance was upheld.

A number of readily-interpretable findings resulted from these analyses. Those holding a positive Attitude towards diversity goals tended to be female, untenured, and at their institution for less than 15 years. Respondents who believed their teaching or research activities reflected issues of diversity (Career scale) were more likely to be female, new to their university, and specializing in the humanities, not in the sciences. Statements that their institution promoted diversity (Environment scale) were more likely to be endorsed

Table 7
Group Comparisons by Demographic Variables

	N	Attitude		Career		Environment		Social	
Faculty		*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Female	109	4.20	0.47	3.51	0.71	3.19	0.60		
Male	120	3.79	0.61	2.95	0.74	3.52	0.53		
t		5.62***		5.77***		-4.46***			
d		0.74		0.75		-0.58			
Non-white	50					3.14	0.76		
White	185					3.43	0.52		
t						-2.50*			
d						-0.40			
Tenured	158	3.93	0.61			3.45	0.53		
Not Tenured	66	4.12	0.51			3.18	0.69		
t		2.24*				2.85**			
d		0.33				0.42			
Full	85					3.49	0.51		
Associate	86					3.38	0.55		
Assistant	58					3.14	0.69		
F						4.24**			
h^2						0.05			
0-5 Years at this University	72	4.10	0.52	3.38	0.76	3.25	0.68		
6-10 Years	39	4.18	0.49	3.29	0.81	3.28	0.63		
11-15 Years	26	4.06	0.51	3.18	0.72	3.34	0.49		
More than 15 Years	92	3.81	0.62	3.02	0.76	3.50	0.49		
F		5.79***		3.18*		2.88*			
h^2		0.07		0.04		0.04			
0-5 Years in Higher Education Overall	41					3.31	0.62		
6-10 Years	46					3.16	0.63		
11-15 Years	24					3.23	0.58		
More than 15 Years	121					3.49	0.53		
F						4.46**			
h^2						0.06			
Sciences	55			2.53	0.63			3.96	0.76
Social Sciences	44			3.40	0.63			3.65	0.72
Humanities	56			3.62	0.67			3.77	0.84
Professional Schools	55			3.30	0.73			3.43	0.82
F				27.65***				4.68**	
h^2				0.29				0.06	

* $p \leq .05$, ** $p \leq .01$, *** $p \leq .001$.
Note. The effect size of Cohen's *d* is typically interpreted as: .2, small, .5, medium, .8, large (Keppel & Wickens, 2004). The effect size of eta-squared (η^2) is typically interpreted as: .01, small, .06, medium, .14, large. The *tenured-not tenured* comparison included only faculty in a tenure-track. The comparison by discipline defined disciplines in this way: *Sciences* included engineering, pharmacy and the natural sciences; *Humanities* included Fine Arts; *Professional Schools* included architecture, business, education, social welfare, journalism and law. A second analysis which did not include architecture, journalism or law in the analysis found similar results.

Statements that their institution promoted diversity (Environment scale) were more likely to be endorsed by males, white faculty and staff, tenured faculty, veteran faculty, and those who had spent more time in higher education overall.

by males, white faculty and staff, tenured faculty, veteran faculty, and those who had spent more time in higher education overall. Those who reported that they interacted with diverse populations as part of their working activities (Social scale) were most likely to be in the sciences and least likely to be in a professional school. In addition to the comparisons shown, we compared faculty born in the United States with faculty born outside the United States, and we compared administrators with non-administrators. In both analyses we found no statistically significant differences.

Table 8 provides comparisons between faculty who taught courses or published research on issues of diversity and those who did not. As would be expected, those who taught or conducted research in areas relevant to diversity issues scored higher on the Career scale than those who did not. They also tended to have more positive attitudes toward institutional diversity goals (Attitude scale). They also scored highly on the Social scale. Additionally, those who had not written about or conducted research in areas of race, ethnicity or gender were less likely to believe that their institution promoted diversity (Environment scale).

Discussion

The central objective of this study was to create a valid and reliable instrument with which to assess faculty support of diversity goals in higher education. In the process of development, we investigated preliminary findings and formulated key questions of interest that warrant further consideration. The instrument presented here is relevant to future research and policy considerations of diversity in higher education as well.

Most existing instruments of institutional diversity focus on attitudes or perceptions of campus climate (see AACU, 2005; Shenkle et al., 1998; Smith et al., 1994). The scales that

Table 8
Group Comparisons by Teaching and Research Experience

		Attitude		Career		Environment		Social	
Faculty	N	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Taught course on global issues	72			3.55	0.71			3.88	0.78
Has not	161			3.03	0.74			3.58	0.82
t				4.99***				2.64**	
d				0.71				0.37	
Taught course on racial/ethnic issues	58	4.20	0.62	3.96	0.47			4.02	0.66
Has not	171	3.90	0.55	2.93	0.68			3.55	0.84
t		3.48***		10.74***				3.98***	
d		0.54		1.65				0.61	
Taught course on women/gender issues	50	4.26	0.61	4.01	0.54			4.08	0.73
Has not	180	3.91	0.55	2.97	0.68			3.55	0.82
t		3.84***		10.23***				4.15***	
d		0.61		1.64				0.66	
Researched global issues	85	4.08	0.61	3.50	0.71				
Has not	145	3.92	0.56	3.01	0.75				
t		2.03*		5.01***					
d		0.28		0.69					
Researched racial/ethnic issues	80	4.19	0.57	3.81	0.60	3.22	0.69	3.86	0.80
Has not	152	3.88	0.56	2.89	0.67	3.43	0.51	3.56	0.83
t		4.03***		10.25***		-2.42*		2.70**	
d		0.55		1.42		-0.36		0.37	
Researched women/gender issues	74	4.23	0.55	3.79	0.66	3.24	0.68	3.90	0.84
Has not	156	3.86	0.56	2.92	0.67	3.41	0.52	3.55	0.81
t		4.72***		9.16***		-2.12*		3.02**	
d		0.67		1.29		-0.30		0.43	

* $p \leq .05$, ** $p \leq .01$, *** $p \leq .001$.
Note. The effect size of Cohen's *d* is typically interpreted as: .2, small, .5, medium, .8, large (Keppel & Wickens, 2004).

most closely resemble these two areas of interest are the Attitudes component, which measures general views about racial/ethnic and gender diversity, and the Environment component, which assesses faculty perceptions of how well the institution is doing relative to its diversity goals. The ACES, however, includes two other important dimensions to the consideration of faculty views on diversity – namely a Social component measuring interaction with people different from oneself and a Career component related to faculty efforts in teaching, research, and service. By having four separate components, the present instrument allows for institutions and researchers to look not only at broad aspects of faculty attitudes and perceptions of their diversity environment but also at the more nuanced and essential translation of these perspectives into action.

By having four separate components, the present instrument allows for institutions and researchers to look not only at broad aspects of faculty attitudes and perceptions of their diversity environment but also at the more nuanced and essential translation of these perspectives into action.

While we initially sought to produce an instrument that could be adjusted to reflect the varied teaching, research, and service loads of faculty at different institutional types, the data collected in this study indicate that participant responses to questions about teaching, research, and service coalesce together. In other words, at least in the context of the single university from which we collected data, faculty members' engagement in research about diversity was highly correlated with the likelihood that they taught or performed service related to diversity as well. Rather than creating separate scales for each component of faculty work, we developed a single Career scale instead. This initial finding has interesting implications for thinking about how faculty can and do shape institutional diversity climates by integrating the primary aspects of their work.

Our results suggest that faculty demographics are also important to consider when assessing diversity. It is important not to conceive of the faculty body as one monolithic group (Somers et al., 1998). For example, we found that women, people of color, newer faculty, and not yet tenured faculty were more likely to have positive attitudes about the importance of diversity (Attitude scale), be engaged in diversity related work (Career scale), and be more critical of their institution's existing diversity climate (Environment scale) than their male, white, and more senior counterparts. These patterns are consistent with other studies comparing the views of racial minority and majority students (D'Augelli & Hershberger,

1993; Hurtado, 1992) as well as faculty perspectives across race, ethnicity, and gender in higher education (Conley & Hyer, 1999; Kossek & Zonia, 1993). Notably, however, faculty who researched and taught about diversity issues—regardless of their individual demographic characteristics—were more likely to have positive attitudes about diversity and positive social interactions with people different from themselves than their peers who were not engaged in such work. And similarly, faculty who researched issues of gender or race and ethnicity were more likely to have critical views about the campus commitment to diversity irrespective of their own demographic backgrounds. This is a new finding and deserves future exploration.

Given the varied scholarly pursuits of faculty across the university, one might also expect disciplinary differences with regard to diversity. Indeed, we found that faculty in the sciences were the least likely to engage in research and teaching about diversity (Career scale) but the most likely to have social interactions with people who are different from themselves (Social scale). These findings make sense considering the nature of science and what is studied on the one hand, and the internationalization of the faculty and graduate students in many science fields on the other. As recent publications have documented, international faculty constitute nearly one third of all new faculty hires in math, science, and engineering fields (Institute for International Education, 2006; Nelson & Rogers, 2005), and they are disproportionately found at research universities (National Science Board, 2003). This study shows that faculty attitudes, perceptions, and behaviors vary by characteristics such as demographics and academic discipline. Thus, future uses of the ACES instrument should be accompanied with information that captures these important differences among respondents.

Limitations

This study is based upon data gathered from faculty at just one research university in the United States with about a 25% response rate and a limited sample size. The representativeness of the results for this particular institution is not known, nor is it known how well results generalize to other colleges and universities. Administering the ACES to a wider array of institutions would help determine if variables such as control (public/private), size, selectivity, resources, geographic region, or even composition of the institution (in terms of representation of diverse students and/or faculty) lead to different results. Also unknown is the generalizability of the psychometric characteristics of the instrument for other populations. It would be useful to administer the instrument to faculty at other comparable universities and further establish evidence of its validity and reliability across institutions. A cross-validation study using confirmatory factor analysis would be valuable to judge the stability of the factor analytic solution to other institutions. Gathering data from other populations would explore whether the ACES scales are suitable in their current form or need modification for different institutional contexts.

An important distinction between these narrow and broader forms of diversity is the extent to which an institution's policies go beyond simply ensuring fair treatment of all but are affirmative in their efforts to expressly recruit, represent, and support these types of diversity.

A second study limitation affects our conclusion that ACES is valid and reliable. While the evidence we collected is supportive of validity and reliability, there exists a broad range of strategies for estimating the reliability of a measure and for developing a validity argument. Our reliability conclusion is based on a coefficient alpha analysis of the internal reliability of our subscales. Other aspects of reliability, such as test-retest (stability across time), were not examined in this study. Our belief that the ACES scores are a valid indicator of attitude or support for institutional diversity goals is based on an initial decision to match items to a theoretical framework, the ease of interpretation of clean factor analysis results, and predictable relationships between ACES scores and demographic (and other descriptive) variables. This study produced only limited or no evidence from other accepted validity sources, such as correlations with other measures of attitude toward diversity or diversity goals, evidence of how the construct measured by these scales is distinct from similar constructs, or how items might be tied to aspects of diversity support which would be identified by a more formal concept analysis.

A third limitation of this study is the relatively narrow definition of diversity we chose. An instrument more inclusive of diversity classifications beyond race, ethnicity, national origin and gender, such as religion, disability, and sexual orientation, for example, might lead to different conclusions. An important distinction between these narrow and broader forms of diversity is the extent to which an institution's policies go beyond simply ensuring fair treatment of all but are affirmative in their efforts to expressly recruit, represent, and support

these types of diversity. If different forms of diversity are in fact treated differently, then there are practical and ethical questions that must be addressed. Future studies using a revised ACES instrument could help determine whether these additional forms of diversity fit well into the existing format, or whether they generate new scales and categories. And, the results of such studies should be accompanied by institutional reflection and discussion about what it means to value diversity in its many manifestations.

Conclusions

There are many benefits to having a valid and reliable instrument for the assessment of faculty support for diversity. Institutions can establish baselines for themselves over time and compare these measures against the effects of diversity related initiatives before and after their implementation. The use of such an instrument can also standardize measures across different institutions so that more meaningful comparisons, collaborations, and modeling might be fostered than previously possible. Utilization by researchers could include determining how the ACES scales are linked to important outcome variables like faculty performance (research productivity or teaching ratings, for example) and faculty satisfaction. Further, by pairing the ACES instrument with other institutional data, one could determine the extent to which faculty views about diversity and institutional climate are linked to student outcomes at the institution such as retention, engagement, or overall satisfaction. This linkage of faculty support for institutional diversity goals to core institutional outcomes would make an important addition to the research literature.

The ability of a university to realize its diversity goals depends significantly upon those individuals who carry out its mission. Thus, it is important to understand how faculty who work in higher education share their institution's stated commitment to diversity and consider how these varied beliefs might be expressed in particular aspects of faculty work. While institutions often develop surveys internally to assess such issues, a review of existing instruments underscores a problematic lack of attention to developing evidence for the validity and reliability of the instruments themselves. Our study addresses the need for such an instrument so that future studies of diversity might be conducted in a more disciplined manner of inquiry.

References

Association of American Colleges and Universities. (2005). *Campus diversity initiative evaluation project resource kit*. Retrieved from www.aacu.org/irvinediveval

Austin, A. (1990). Faculty cultures, faculty values. *New Directions for Institutional Research, 84*, 47-64.

Austin, A. (1994). Understanding and assessing faculty cultures and climates. *New Directions for Institutional Research, 68*, 61-74.

Baird, L. (1990). Campus climate: Using surveys for policy making and understanding. *New Directions for Institutional Research, 68*, 35-45.

Bartlett, M. S. (1954). A note on multiplying factors for various chi-squared approximations. *Journal of the Royal Statistical Society, Series B, 16*, 296-298.

Bensimon, E.M. (2004). The diversity scorecard: A learning approach to institutional change. *Change, 36*(1), 44-52.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*(2), 245-276.

Clark, B. (1987). *The academic life: Small worlds, different worlds*. Princeton, NJ: Carnegie Foundation for the Advancement of Teaching.

Conley, V. M., & Hyer, P. B. (1999). *A faculty assessment of the campus climate for diversity*. Annual Meeting of the Association for the Study of Higher Education, San Antonio, TX.

D'Augelli, A.R., & Hershberger, S.L. (1993). African American undergraduates on a predominantly white campus: Academic factors, social networks, and campus climate. *Journal of Negro Education, 62*(1), 67-81.

Edelman, L.B., Fuller, S.R., & Mara-Drita, I. (2001). Diversity rhetoric and the managerialization of law. American *Journal of Sociology, 106*(6), 1589-1641.

Hurtado, S. (1992). The campus racial climate: Contexts of conflict. *Journal of Higher Education, 63*(5), 539-569.

Institute for International Education. (2006). *Open doors: Report on international educational exchange*. Washington D.C. Bureau of Educational and Cultural Affairs of the U.S. Department of State.

Kaiser, H.F. (1974). An index of factorial simplicity. *Pschometrika, 39*, 31-36.

Keppel, G., & Wickens, T.D. (2004). *Design and analysis: A researcher's handbook (4th ed.)*. Upper Saddle River, NJ: Prentice Hall.

Kossek, E.E., & Zonia, S.C. (1993). Assessing diversity climate: A field study of reactions to employer efforts to promote diversity. *Journal of Organizational Behavior, 14*(1), 61-81.

McGowan, M.O. (1996). Diversity of what? *Representations, 55*, 129-138.

Moses, M.S., & Chang, M.J. (2006). Toward a deeper understanding of the diversity rationale. *Educational Researcher, 35*, 1, 6-11.

National Science Board. (2003). *The science and engineering workforce: Realizing America's potential*. Washington, DC: National Science Foundation.

Nelson, D. J., & Rogers, D.C. (2005). *A national analysis of diversity in science and engineering faculties at research universities*. Norman, OK: University of Oklahoma.

Pohan, C.A., & Aguilar, T.E. (2001). Measuring educators' beliefs about diversity in personal and professional contexts. *American Educational Research Journal, 38*(1), 159-182.

Shenkle, C.W., Snyder, R.S., & Bauer, K. (1998). Measures of campus climate. *New Directions for Institutional Research, 98*, 81-99.

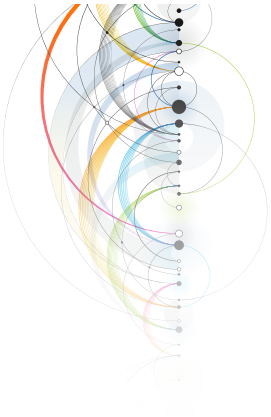
Smith, D. G. (2009). *Diversity's promise for higher education: Making it work*. Baltimore, MD: Johns Hopkins University Press.

Smith, D.G., Wolf-Wendel, L.E., & Levitan, T. (1994). Studying diversity in higher education. *New Directions for Institutional Research, 81*.

Somers, P., Cofer, J., Austin, J., Inman, D., Martin, T., Rook, S., Stokes, T., & Wilkinson, L. (1998). Faculty and staff: The weather radar of campus climate. *New Directions for Institutional Research, 98*, 35-52.

Terenzini, P.T., Cabrera, A.F., Colbeck, C.L., Bjorklund, S.A., & Parente, J.M. (2001). Racial and ethnic diversity in the classroom: Does it promote student learning? *The Journal of Higher Education, 72*(5), 509-531.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. Washington, DC: American Psychological Association.



Abstract

The goals of a Liberal Studies education are designed to prepare citizens to live responsible, productive, and creative lives in a changing world. Ideally, a liberal education fosters well-grounded intellectuals with dispositions toward learning and an acceptance of responsibility regarding their ideas and actions. To measure the efficacy of a Liberal Studies education, a Midwestern regional university developed a systematic, rubric-guided assessment based on nationally recognized science principles and inquiry processes to evaluate student work in undergraduate science laboratory courses relative to a liberal education. The rubric presented a direct measure of student understandings of science inquiry processes. The assessment procedure used stratified random sampling at confidence levels of 95% to select student work, maintained anonymity of students and faculty, addressed concerns of university faculty, and completed a continuous improvement feedback loop by informing faculty of assessment results to assess and refine science-inquiry processes of course content. The procedure resulted in an assessment system for benchmarking science inquiry processes evident in student work and offered insights into the effect of undergraduate science laboratory courses on student knowledge and understanding of science inquiry. The current assessment was without additional burdening of faculty or supplementary testing of students.

AUTHORS

Judith Puncochar, Ph.D.
Northern Michigan University

Mitchell Klett, Ph.D.
Northern Michigan University

A Model for Outcomes Assessment of Undergraduate Science Knowledge and Inquiry Processes

Liberal education is an approach to college learning that empowers individuals and prepares them to deal with complexity, diversity, and change inherent in a democracy (Carson, 1997). This approach emphasizes broad knowledge of science, culture, and society (Pingree, 2007). A liberal education is posited to help students develop an intellectual foundation to recognize real world issues and a sense of social responsibility to hone practical skills for solving problems in real-world settings (Schneider, 2008). The Association of American Colleges and Universities conducted a survey in 2013 and found 74 percent of employers would recommend a liberal education approach to college-bound students (Hart Research Associates, 2013). "What employers clearly want and need are liberally educated professionals" (Humphreys, 2013, para. 8). A commitment to advancing and improving liberal education must be measured and assessed to determine how well the liberal education approach meets the intended outcomes.

Assessment of science knowledge and learning is centuries old and initially used processes such as the Socratic Method. More recently, an upsurge of standardized testing has influenced assessment of science knowledge, but standardized tests do not offer a process by which to improve science inquiry processes and learning outcomes of natural sciences courses in higher education (Steedle, Kugelmass, & Nemeth, 2010). Standardized testing methods rarely assess student learning experiences, account for individual differences in learning needs, or assess the ability of students to think analytically, understand big picture concepts, or apply specific science details to the real world.

The American Association for the Advancement of Science (AAAS; 2013), the National Research Council (NRC; Shavelson & Towne, 2002), and National Science Teachers Association (NSTA; 2011) agree that scientific inquiry is a powerful way for students to understand science content. Assessing student understanding of science inquiry knowledge and processes has been a challenge to several assessment approaches. Students must learn

CORRESPONDENCE

Email
jpuncoch@nmu.edu

how to ask science questions and use evidence to answer these questions. In the process of acquiring strategies of scientific inquiry, students learn to conduct an investigation, collect evidence from a variety of sources using evidence-based methodologies (Faust, 2000), develop an explanation from the data, and communicate and defend their conclusions. Scientific inquiry refers to these "activities through which students develop knowledge and understanding of scientific ideas, as well as an understanding of how scientists study the natural world" (National Academy of Sciences, 2013, p. 23).

Standardized testing methods rarely assess student learning experiences, account for individual differences in learning needs, or assess the ability of students to think analytically, understand big picture concepts, or apply specific science details to the real world.

Our research posits the use of a rubric based on guidelines from the AAAS, NRC, NSTA, and the National Numeracy Network (NNN) to assess undergraduates' learning of science inquiry knowledge and processes in their science laboratory courses. We also propose an easy to implement, easy to replicate, and generalizable method of data collection of student work in undergraduate science laboratory classes.

Previous science program assessments may not have been based on science standards and their data collection methods may not have been easily transferable to other programs. For example, various methods to measure student knowledge of science inquiry and to assess natural science outcomes have included attitudinal surveys, interviews, journaling, performance assessments, portfolios, conceptual level tests, and rubrics (Ellis, Mathieu, & Brissenden, 2003). Traditional testing of students in science laboratory courses has shown little value in guiding student learning or in course or program improvement (Rennie, 1994). Evaluations of science laboratory instruction have lacked feedback on student learning outcomes (Seymour, Wiese, & Hunter, 2003). Alternative assessments of the influence of a science program on science literacy have included an internal evaluation conducted with teams of students in capstone courses to explore student perceptions of science learning (Augeri et al., 2011), an examination of the relationship between science knowledge and creating argumentation (Hakyolu & Ogan-Bekiroglu, 2011), and quasi-experimental comparisons of student achievement of inquiry-based science knowledge under conditions of the presence or absence of traditional tests and quizzes (Taylor, 2000) and student-centered versus teacher-centered instruction (Lord, Travis, Magill, & King, 2005)

Web-based science assessment tools are available for science program evaluations (e.g., Assessment Tools in Informal Science, 2011). Several web-based computer models of active science processes offer additional approaches to help students understand science inquiry (Kastens & Rivet, 2008). Inquiry Science Environment (WISE) web-based modules provide visualizations of thermodynamics, electrostatics, and plate tectonics to guide students to connect scientific ideas when conducting inquiry investigations (Resnick & Zurawsky, 2007). The Student Assessment of Learning Gains (SALG) offers powerful individualized statistical analysis of science learning from a student's perspective to help with immediate formative course evaluation (Seymour, Wiese, Hunter, & Daffinrud, 2000). However, student self-assessments are best used by instructors who seek to improve their courses or by students who can take responsibility for their own learning improvement. Such self-report methods are unlikely to be reliable or generalizable for science program improvement.

An evaluation of science inquiry processes ideally would include performance assessments of student understanding of tools and processes for addressing scientific relationships within the real world, which could be difficult to implement on a large scale (Buxton & Provenzo, 2011). Moreover, procedures for performing an evaluation of a postsecondary science program have political and educational importance. Results must be reliable, unbiased, meaningful, and based on the strength of evidence, but such program evaluations are few in number (Slavin, 2008).

For a process to be useful in measuring student knowledge and understanding of science inquiry, an assessment must focus on student learning, be useful for program and course improvement, employ replicable methods to assess student work, and have a process in place to act on the findings. The university implemented these criteria and followed steps in Wright's (2003) assessment loop:

1. Setting goals or asking questions about student learning and development;
2. Gathering evidence that will show whether these goals are being met;
3. Interpreting the evidence to see what can be discovered about students' s trengths and weaknesses;
4. Using those discoveries to change the learning environment so that student performance will improve.

The cycle was repeated to include improved interventions for student learning based on assessment data. Our assessment goal was to determine the extent of undergraduate science inquiry abilities and understandings as evidenced by student work in science laboratory courses.

An evaluation of science inquiry processes ideally would include performance assessments of student understanding of tools and processes for addressing scientific relationships within the real world, which could be difficult to implement on a large scale.

Laboratories have opportunities for students to design and conduct investigations. Students can collect evidence needed to answer a variety of questions, draw conclusions, and think critically and logically to create explanations based on evidence. In science laboratories, students have a setting to communicate and defend their results to peers and others.This study is limited to an examination of student knowledge and understanding of science inquiry processes within science laboratory courses taught during Spring Semester 2010.

The university's bulletin has a description of core competencies expected of students in science laboratory courses. Students in science laboratory courses are expected to be active in learning the processes and strategies of scientific inquiry. Students also are expected to demonstrate knowledge of science and abilities, design and conduct investigations, collect evidence from a variety of sources, develop an explanation from the data, and communicate and defend their inferences from data to conclusions. Student work in science laboratory courses should provide evidence not only of studied scientific knowledge, but also of the nature of scientific inquiry processes. Scientific, analytical, and logical processes should transcend particular course knowledge to provide students with greater talents and abilities to solve problems and reason rationally.

Foundations of Assessment Process

The American Association for Higher Education and Accreditation (AAHEA) placed assessment as an ongoing process aimed at understanding and improving student learning (AAHEA, 2013). The goal of the current assessment process was to report results to faculty to implement appropriate curricular and instructional changes to support and improve student learning.

The Liberal Studies Committee (LSC) is a standing committee of the university's Academic Senate. The LSC has oversight and responsibilities to review, evaluate, and recommend changes or improvements of the Liberal Studies Program based on assessed effectiveness of undergraduates to develop knowledge, skills, and perspectives while progressing through their liberal studies education.

The LSC evaluated one of six different divisions each academic year for assessment purposes. In 2010, the LSC selected Division III Natural Science and Mathematics courses as the assessment focus. Mathematics courses were assessed separately. The current study reports only assessment of student work in Liberal Studies Natural Science courses with a laboratory component.

The LSC developed a plan using Wright's (2003) assessment procedure. First, the LSC identified instructors and science courses within Division III during Spring Semester 2010. Second, the Director of Institutional Research generated a randomly selected student sample from a list of science laboratory courses. Instructors were contacted and provided instructions regarding how to maintain student anonymity when submitting the requested sampling of student work. Meanwhile, the LSC created scoring criteria for assessing science abilities and understandings following guidelines of the AAAA, NNN, NRC, and NSTA, established reporting procedures, and identified an independent evaluation team of three faculty members to score student work with rubric. Each step of the assessment process is described more completely below.

The LSC determined courses in Astronomy, Biology, Chemistry, Environmental Science, Geography, Physics, and Psychology met criteria of Division III Foundations of Natural Sciences during spring 2010. All courses were designed to introduce students to quantitative reasoning and scientific understanding of current views of the natural world. Nearly all courses were introductory courses. Since no courses taught in spring 2010 were approved for advanced Liberal Studies credit, an assessment of the influence of advanced level courses on science inquiry knowledge and understanding was not conducted.

In March 2010, the Chair of the LSC met with the Director of Institutional Research to identify science faculty and instructors of Division III science laboratory courses taught during spring 2010 and create a list of randomly selected students for each identified science laboratory course. Each science instructor received a letter and an email. Natural Science laboratory instructors were also provided a list of Liberal Studies guidelines for Division III math and science courses (see Appendix A).

Stratified Random Sampling

The LSC and the Office of Institutional Research compiled the population of students enrolled in all Division III Natural Science courses and determined a stratified random sampling of 350 students would provide a confidence level of 95%, which is the confidence level used by the LSC in previous assessments. The Office of Institutional Research's list of randomly selected students represented 8%–10% of students enrolled in each course. Since the list was generated prior to the drop date, some students had dropped the course before collection of student work occurred, which contributed to a return rate of less than 100% of requested student work.

Students were selected by stratified random sampling, which produces an allocated proportion of the total population. For example, if the population consisted of 60% women and 40% men, then three women and two men would reflect proportions of the sample. The LSC reviewed a random sample of about 9% of student work in the Division III Natural Science laboratory courses.

Individual science laboratory courses sometimes consisted of both lecture and lab or lab only. Faculty and instructors decided whether to submit student work from both lecture and lab or lab only because individual students could enroll in lecture and lab concurrently or separately. Many of the selected students in the sample were enrolled only in laboratory sections of a course. All work submitted for each randomly selected student counted as one set of student work or artifact. The Chair of the LSC collected student work and artifacts after finals week. All identifying features of students were removed from their work and artifacts. Student anonymity was maintained.

Instrument Design

The LRC formed a subcommittee of three members to create a rubric to score criteria for assessing science abilities and understandings based on guidelines from the AAAS, NRC, NSTA, and NNN. Rubrics have long been used to assess student performance using criteria to focus an evaluation with a set of objective external scoring criteria and point-values associated with the criteria by level of performance (Schmoker, 2006). Data from rubrics are used for summative program assessment to compare worthiness of student performance and expected outcomes against external standards (Ebert-May, 2003). Rubrics provide faculty a readily accessible way to quantitatively assess student achievement based on the sum of a range of criteria determined by looking directly at student work (Dodge & Pickette, 2001).

Rubrics provide faculty a readily accessible way to quantitatively assess student achievement based on the sum of a range of criteria determined by looking directly at student work.

At the outset, subcommittee members read the AAAS, NRC, NSTA, and NNN guidelines. Each member arrived at the next meeting with an attempt to distill common core competencies into learning outcomes. The subcommittee discussion was facilitated by the Chair of the LSC, who was a professor of earth and space science. The subcommittee reached consensus on five separate learning outcomes based on AAAS, NRC, and NSTA guidelines (see Figure 1). Notably, rubric development did not start with the goals and objectives of the Liberal Studies Division III Natural Science courses. The rubric used science competencies and concepts based on nationally recognized science principles to assess student knowledge of science and scientific inquiry processes. The numerical scoring format was based on recommendations of the NNN

for "Advancing Assessment of Scientific and Quantitative Reasoning," which was a National Science Foundation funded project (DUE 0618599) to "further the development of collegiate scientific and quantitative reasoning assessment tools and procedures" (Sundre, Murphy, & Handley, 2009, para. 1).

The subcommittee developed a rubric using a five-point ordinal scale to reflect nuances within the Liberal Studies abilities and understandings of scientific concepts, recognition and use of scientific reasoning methods, understanding and discussion of general scientific articles, and use of mathematics in scientific reasoning and/or problem resolutions. A score of 0 meant the student work completely lacked evidence that the learning outcome was met (e.g., all evidence for the learning outcome was missing). A score of 1 indicated the student work was lacking sufficient evidence to meet the learning outcome (e.g., sporadic, patchy evidence and unfinished or imperfect responses). A score of 2 indicated the rater neither agreed nor disagreed that the outcomes were met and served as a neutral response for cases where a rater could not decide whether the student work did nor did not meet the learning outcome (e.g., perhaps a good start but lacking solid evidence). A score of 3 indicated the rater agreed the learning outcome was met by consistent, sufficient evidence provided by the student work (e.g., recognizes various forms of evidence and uses knowledge of natural phenomena and the physical world). A score of 4 indicated the rater strongly agreed that the student work provided quality evidence that exceeded expectations for the level of the course (e.g., synthesizes well-structured, articulated inquiry processes of natural phenomena and the physical world).

Changing raw scores to percentages revealed that 27% of student work had evidence to exceed expectations (Strongly Agree) and 32% of student work had evidence to meet expectations (Agree), resulting in 59% of the student work meeting or exceeding expectations.

Each of the five criteria in the rubric addressed specific scientific processes as defined by the AAAS, NRC, and NSTA. Understanding and use of scientific concepts referred to evidence of use of science knowledge as information in student work. Applying knowledge of science to everyday experiences referred to evidence of the ability to apply science outside of the laboratory to experiences in the natural world. Recognizing and use of scientific reasoning referred to evidence of scientific inquiry process and reasoning skills, which are distinct from the scientific procedures, observations, or concepts. Understanding and discussing general scientific articles required evidence of citations, references, or referrals to science articles, research, or researchers in student work. Use of mathematics in scientific reasoning and/or problem resolutions required evidence of credible use of scientific and mathematical information in scientific developments and public policy issues. Construction of the rubric used "sound assessment methods and practices" (Sundre et al., 2009, para. 2). After creating the rubric, the subcommittee selected three faculty members to form an assessment team to score student work in science laboratory classes.

Assessment Team

The assessment team was selected using the following criteria: (a) at least one member must teach courses in the Liberal Studies Division III Natural Sciences, (b) at least one member must not teach in Division III, and (c) a third member who may or may not teach in the Division III. Faculty members from Psychology, Biology, and Chemistry formed the assessment team. Team members consisted of voluntary faculty volunteers from departments that offered undergraduate science courses. Members were chosen based on their experience in teaching math and science courses and on their expertise in science knowledge, assessment, and evaluation. Each faculty volunteer received a stipend to work on the assessment team. Assessment team members attended a training session to practice scoring samples of laboratory science work not included in this study. Reviewers completed their review of student work from 350 students in five to eight hours. The LSC chair acted as the coordinator of the assessment team.

Results

As with earlier collections of student work in other Liberal Studies Divisions, faculty who submitted student samples did so in a timely fashion. Science laboratory faculty and instructors were 79% in compliance with submitting student work, which represents the highest percentage of compliance within the six Liberal Studies divisions. The LSC commended efforts of the Office of Institutional Research staff and of faculty who submitted student work for assessment. All departments represented within Division III turned in student work for the assessment process.

The three raters scored student work from entry-level science laboratory courses using the rubric. Numbers with the symbol # in Table 1 refer to the following rubric criteria:

1. Understanding and use of scientific concepts
2. Application of knowledge of science to everyday experience
3. Recognition and use of scientific reasoning methods
4. Understanding and discussion of general scientific articles
5. Use of mathematics in scientific reasoning and/or problem resolutions

Table 1

Frequency of Raters' Scores Using Five-point Likert Scaling to Assess Science Abilities and Understandings

	Rater 1					**Rater 2**					**Rater 3**				
	Science Abilities and Understandings					Science Abilities and Understandings					Science Abilities and Understandings				
Rating	#1	#2	#3	#4	#5	#1	#2	#3	#4	#5	#1	#2	#3	#4	#5
SD	2	0	129	309	111	21	24	21	66	95	0	0	1	54	24
D	18	36	47	12	116	13	8	8	5	5	17	24	140	163	201
ND NA	238	141	70	2	21	20	16	16	2	14	21	171	102	12	28
A	124	201	116	69	90	79	27	90	30	14	252	93	51	71	22
SA	14	17	34	4	54	158	216	156	58	153	14	16	10	4	25
Total	396	395	396	396	392	291	291	291	161	281	304	304	304	304	300

Note. Abbreviations are as follows: SD = Strongly Disagree, D = Disagree, ND NA = Neither Disagree Nor Agree, A = Agree, and SA = Strongly Agree. The symbol # followed by a number refers to the order of rubric criteria for assessing abilities and understandings of science inquiry processes.

Cohen's Kappa statistic was used to assess the degree to which two or more raters who examine the same ordinal data agree when assigning data to rubric categories. Kappa is a "chance-corrected proportional agreement" (Stawicki, 2010, para. 2) with possible values ranging from +1 (perfect agreement) to 0 (no agreement above that expected by chance) to -1 (complete disagreement). Kappa values were statistically significantly different from 0, suggesting that ratings between coders were largely similar. Table 2 includes the 15 Kappa ratings comparing raters with one another across the five rubric categories. Nine of the 15 Kappa ratings were in the substantial range (0.61–0.80), three were in the moderate (0.41–0.60) range, and three were in the fair range (0.21–0.40; see Landis & Koch, 1977). Cross tabulation reaffirmed that coders largely agreed.

Higher inter-rater agreement occurred in ratings associated with understanding and use of scientific concepts, recognition and use of scientific reasoning methods, understanding and discussion of general scientific articles, and use of mathematics in scientific reasoning and/or problem resolutions (see Table 2). Lower inter-rater agreements occurred in ratings associated with understanding multiple problem-solving perspectives.

Ratings of Student Work

Summing ratings by similar levels of the rubric (e.g., Strongly Agree) showed higher ratings on student work with evidence of an understanding of current views of natural phenomena, specifically through "Understanding and Use of Scientific Concepts" and "Application of Knowledge of Science to Everyday Experience." Lower ratings occurred on student work with evidence of an "Understanding and Discussion of General Scientific Articles" and "Use of Mathematics in Reasoning and Problem Solving." Coders used all five points of the rubric (see Figure 1).

Table 2
Kappa Calculations between Three Raters on their Assessments of Students' Abilities and Understandings of Science Inquiry Processes

n = 145	R1 R2	R1 R3	R2 R3
1. Understanding and Use of Scientific Concepts	0.66	0.33	0.64
2. Application of Knowledge of Science to Everyday Experience	0.54	0.23	0.38
3. Recognition and Use of Scientific Reasoning Methods	0.65	0.57	0.65
4. Understanding and Discussion of General Scientific Articles	0.66	0.53	0.65
5. Use of Mathematics in Scientific Reasoning and/or Problem Resolutions	0.77	0.66	0.78

Note. Abbreviations R1, R2, and R3 refer to Rater 1, Rater 2, and Rater 3 respectively. The symbol *n* designates number of scores in a randomly selected, limited portion of the total sample of 10,596 scores.

Changing raw scores to percentages revealed that 27% of student work had evidence to exceed expectations (Strongly Agree) and 32% of student work had evidence to meet expectations (Agree), resulting in 59% of the student work meeting or exceeding expectations. Forty-one percent of student work did not provide evidence to meet expectations (i.e., 19% of student work lacked evidence for the criteria and 22% of student work had equivocal evidence).

Likert Scaling	Number of Scores	Percent of Total Scores
Strongly Disagree	857	8%
Disagree	1173	11%
Neither Agree Nor Disagree	2284	22%
Agree	3339	32%
Strongly Agree	2943	27%
	10596	100%

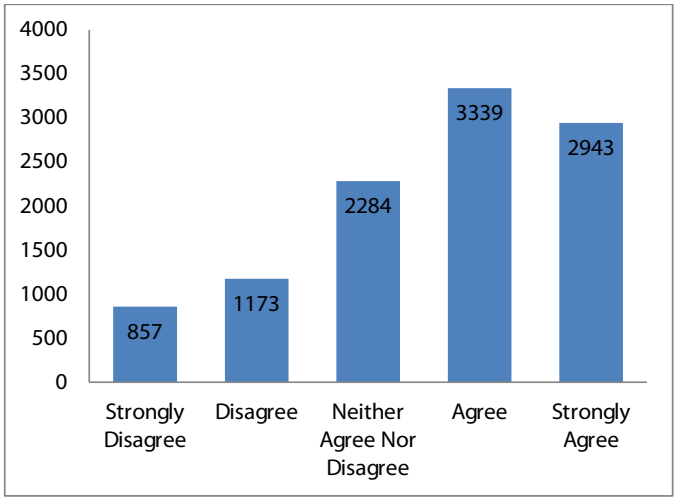


Figure 1. Ratings of student work submitted by instructory science labratory courses as evidence of student knowledge and understanding of science inquiry in Liberal Studies. Results are displayed from left to right acccording to rubric-guided Likert scaling by number of scores, percentage of total scores, and bar graphs with a display of raw scores.

Discussion

Liberal education is an approach posited to prepare students to deal with complexity, diversity, and change (Carson, 1997). An assessment of outcomes of a liberal education establishes a baseline to measure practical skills for solving problems in real-world settings (Schneider, 2008) and science competence of all students taking Liberal Studies science laboratory classes, including students typically underserved by the undergraduate learning experience (Seymour, 2002). Creating a baseline of undergraduate knowledge and inquiry processes helps to determine how well the liberal education approach meets its intended outcomes.

Information garnered from the assessment of students' understanding and use of scientific concepts, recognition and use of scientific reasoning methods, understanding and discussion of general scientific articles, and use of mathematics in scientific reasoning and/or problem resolutions taught us at least three important lessons to enhance future practices

in liberal education outcomes assessment. These lessons include maintaining excellent communication practices, developing a valid and reliable rubric for the assessment, and using internal experts to conduct the assessment.

First, maintaining transparency in communications about the process was imperative to gain faculty cooperation with the assessment process. Initially, we used email and phone calls to communicate with faculty about collection procedures of student work and development of the assessment rubric. A faculty-led discussion on the assessment process at an Academic Senate meeting was helpful in garnering faculty support.

Second, science laboratory courses are well suited for performance-based assessments. Students and faculty are familiar with inquiry-based assessments and external science standards allowed the development of a robust rubric based on valid criteria for the assessment process. A concise rubric scoring scale helped to avoid scoring bias and unreliability. The validity and reliability of the process provided a vigorous, easily defensible assessment process.

Third, a committee comprised of faculty from all colleges developed the evaluation rubric, and an assessment team of faculty from diverse science backgrounds conducted the assessment process, both of which added credibility and included an explicit process to avoid scoring bias.

Creating a baseline of undergraduate knowledge and inquiry processes helps to determine how well the liberal education approach meets its intended outcomes.

With assessment results in hand, we looked for ways to "close the loop" on how these results are being used to improve student outcomes (see the National Institute for Learning Outcomes Assessment, 2013). Using assessment evidence at department, program, and course levels to make actual improvements in student learning and inform curriculum decisions is challenging (Bailie, Marion, & Whitfield, 2010; Banta & Blaich, 2011). At first, assessment results went directly to department heads to share with faculty of science laboratory courses and the LRC assessment report was posted on the university's assessment website. No formal reporting mechanism was initially in place to follow whether or how faculty and instructors used assessment information to improve their science laboratory courses or student learning of science inquiry processes. In 2012, a process was initiated as an Academic Quality Improvement Program initiative to have faculty from all disciplines work together in small groups to develop learning outcomes for their syllabi (see Hammock & Richardson, 2011, for a similar process). Science laboratory faculty who attended the workshops developed inquiry-based learning outcomes to provide student data for a continuous improvement feedback loop to assess and refine science-inquiry processes of course content.

The next phase of science laboratory course assessments is slated for 2016. Links to national science standards, the rubric, and a report about the assessment process are on the university assessment website. Discussions are underway to explore the benefits of creating a "connections" type of science laboratory course with a focus on applying/integrating science inquiry processes. Presently, the Liberal Studies Natural Science Division III has two course levels (i.e., 100–200 [Emerging aka "lower division"] and 300–400 [Innovating aka "upper division"]).

Summary

Evidence of assessment and evaluation are critical to a university's accreditation processes. We recommend selecting a non-intrusive, statistically defensible, stratified random sampling of student artifacts for the assessment and evaluation process. The method of data collection worked well and met the usual goal of sampling, which is to produce a representative sample. Occasionally, faculty would inquire whether they could submit the "best examples of student work," rather than submitting the work of randomly selected students. The LSC insisted on conforming to accepted statistical practices on the collection of student artifacts from stratified random samples.

After the assessment, raters gave their feedback on the assessment and evaluation process. They suggested more training on initial ratings of student work samples to hone their skills to automaticity with the scoring rubric.

We advise giving clarifying information to faculty and instructors on how to select examples of student work and artifacts to submit. For example, laboratory reports, papers, essays, and even short answer problem-based items were excellent artifacts for assessing

Laboratory reports, papers, essays, and even short answer problem-based items were excellent artifacts for assessing science understanding and use of scientific concepts, recognition and use of scientific reasoning methods, understanding and discussion of general scientific articles, and use of mathematics in scientific reasoning and/or problem resolutions.

science understanding and use of scientific concepts, recognition and use of scientific reasoning methods, understanding and discussion of general scientific articles, and use of mathematics in scientific reasoning and/or problem resolutions. Submitting student grades was of no value to raters for assessing science abilities or understandings and resulted in a rating of zero.

Student work for this assessment was gathered from science laboratory courses taught in Spring Semester 2010 in entry-level science courses. No student work came from advanced courses. Consequently, finding only 28% of student work exceeded expectations is not surprising on an assessment of science knowledge in introductory courses. Comparing our findings to a baseline of TIMSS 2007 results offered insights into trends in student knowledge of science and science processes. When compared to the international median, about 38% of U.S. eighth-graders performed at a high benchmark (28%) or above the advanced benchmark in science (10%; TIMSS 2007). In comparison, 59% of the study's undergraduates performed at expectations in entry-level undergraduate Division III Natural Science courses. For lower performing students, TIMSS 2007 results had 29% of U.S. eighth-graders performing at or below the low benchmark in science. Our raters determined 19% of undergraduates performed below expectations and 22% were approaching expectations.

Our research offers a feasible, systematic, outcomes assessment approach to evaluation of undergraduate science programs. We have honored Wright's (2003) outline of the assessment process and met criteria outlined by Slavin (2008) for a reliable, rigorous, unbiased, and meaningful assessment based on the strength of evidence.

Next steps include using the assessment results of student knowledge and understanding of science inquiry processes to improve teaching and learning in Division III Natural Science courses and to invite other postsecondary institutions to use the rubric to assess student knowledge and understanding of science inquiry processes. The assessment process provides a meaningful measurement and documentation of undergraduates' science learning and offers an opportunity for faculty and instructors to bridge the gap between undergraduate science teaching and student learning of science theory and practice.

Goals for a liberal education include intellectual development and attainment of intellectual skills, broad knowledge, social responsibility, integrative learning, and demonstrated ability to use one's knowledge in real-world contexts (Schneider, 2008). To assess whether the goals of a liberal education have been achieved, college faculty members have a responsibility to evaluate science inquiry learning outcomes of a general education that academic institutions seek to impart to students. The fundamental worth of our assessment method is use of a generalizable stratified random sampling assessment method of student work and an easy to implement and replicate rubric based on nationally recognized science standards and inquiry processes, which strives to rise above the studied scientific knowledge to assess student understanding of the nature of scientific inquiry processes. Such understanding of scientific inquiry processes should transcend particular course knowledge to provide students with greater talents and abilities to solve problems, reason logically, and live rationally.

References

American Association for Higher Education and Accreditation. (2013). *AAHE/AAHEA*. Retrieved from http://www.aahea.org/aahea/

American Association for the Advancement of Science. (2013). *AAAS homepage*. Retrieved from http://www.aaas.org/

Assessment Tools in Informal Science. (2011). *ATIS homepage*. Retrieved from http://www.pearweb.org/atis/

Augeri, M., Brents, M., Christiansen, E., Etzenhouser, B., Fox, S., Giese, K. … van Andel, M. (2011). *Assessment of student learning outcomes at an institutional level*. Retrieved from www.uiowa.edu/~outcomes/documents/FocusGroups-FINAL.pdf

Bailie F., Marion B., & Whitfield, D. (2010). How rubrics that measure outcomes can complete the assessment loop. *Journal of Computing Sciences in Colleges, 25*(6), 15-28.

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

Buxton, C., & Provenzo, E. F. (2011). "Natural Philosophy" as a foundation for science education in an age of high-stakes accountability. *School Science and Mathematics, 111*(2), 47-55. doi:10.1111/j.1949-8594.2010.00060.x

Carson, R. N. (1997). Science and the ideals of liberal education. *Science & Education, 6*(3), 225-238. doi:10.1080/00091383.2011.538642

Dodge, B., & Pickette, N. (2001). *Rubrics for web lessons*. Retrieved from http://webquest.sdsu.edu/rubrics/weblessons.htm

Ebert-May, D. (2003). *Classroom assessment techniques: Scoring rubrics*. Retrieved from http://www.flaguide.org/

Ellis, A., Mathieu, B., & Brissenden, G. (2003). *Field-tested Learning Assessment Guide (FLAG)*. Retrieved from http://www.flaguide.org/

Faust, D. (2000). The concept of evidence. *International Journal of Intelligent Systems, 15*, 477-493.

Hakyolu, H., & Ogan-Bekiroglu, F. (2011). Assessment of students' science knowledge levels and their involvement with argumentation. *International Journal for Cross-Disciplinary Subjects in Education, 2*, 264-270.

Hammock, G., & Richardson, D. (2011). *Closing the loop: Linking assessment with course design*. SoTL Commons: A Conference for the scholarship of teaching & learning. Abstract retrieved from http://eaglescholar.georgiasouthern.edu:8080/jspui/handle/ 10518/3833

Hart Research Associates. (2013, April 10). It takes more than a major: Employer priorities for college learning and student success. Retrieved from http://www.aacu.org/leap/documents/2013_EmployerSurvey.pdf

Humphreys, D. (2013). Success after college: What students, parents, and educators need to know and do. *Liberal Education, 99*(2).

Kastens, K. A., & Rivet, A. (2008). Multiple modes of inquiry in earth science: Helping students understand the scientific process beyond laboratory experimentation. *The Science Teacher, 75*(1), 26-31.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174. doi:10.2307/2529310

Lord, T., Travis, H., Magill, B., & King, L. (2005). Comparing student-centered and teacher-centered instruction in college biology labs. Retrieved from http://stemtec.org/pathways/Proceedings/Papers/Lord-p.doc

National Institute for Learning Outcomes Assessment. (2013). *NILOA homepage*. Retrieved from www.learningoutcomeassessment.org/

National Academy of Sciences. (2013). *National science education standards*. Washington, DC: National Academy Press. Retrieved from http://www.nap.edu/openbook.php?record_id=4962

National Science Teachers Association. (2011). *Positions: Official NSTA positions on a range of issues*. Retrieved from http://www.nsta.org/about/positions.aspx?lid=abt

Pingree, S. E. (2007). Bringing theory to practice & liberal education. *Liberal Education*. Retrieved from http://www.aacu.org/liberaleducation/le-wi07/Le-wi07_feature4.cfm

Rennie, L. J. (1994). Measuring affective outcomes from a visit to a Science Education Centre. *Research in Science Education, 24*(1), 261-269. doi:10.1007/BF02356352

Resnick, L. B., & Zurawsky, C. (2007). Science education that makes sense. *American Educational Research Association (AERA): Research Points, 5*(1), 1-4.

Schmoker, M. (2006). *Results NOW: How we can achieve unprecedented improvement in teaching and learning*. Washington, DC: ASCD.

Schneider, C. G. (2008). Liberal education takes a new turn. *The NEA 2008 Almanac of Higher Education, 2008*, 29-40. Retrieved from www.nea.org/assets/img/PubAlmanac/ ALM_08_03.pdf

Seymour, E. (2002). Tracking the processes of change in US undergraduate education in science, mathematics, engineering, and technology. *Science Education, 86*(1), 79-105. doi:10.1002/sce.1044

Seymour, E., Wiese, D., & Hunter, A. (2003). *Classroom assessment techniques: Student assessment of learning gains*. Retrieved from http://www.flaguide.org/

Seymour, E., Wiese, D., Hunter, A., & Daffinrud, S. M. (2000). *Creating a better mousetrap: Online student assessment of their learning gains*. Paper presentation at the National Meeting of the American Chemical Society, San Francisco, CA. Retrieved from http://www.salgsite.org/docs/SALGPaperPresentationAtACS.pdf

Shavelson, R. I., & Towne, L. (Eds.). (2002). *Scientific research in education*. Washington, DC: National Academies. Retrieved from http://www.nap.edu

Slavin, R. E. (2008). Perspectives on evidence-based research in education--What works? Issues in synthesizing educational program evaluations. *Educational Researcher, 37*, 5-14. doi:10.3102/0013189X08314117

Stawicki, S. P. (2010). Kappa – Agreement – Statistic. *Scientific Training and Research*. Retrieved from http://www.medschoolwiki.org/msw/main/index.php/Kappa_-_Agreement_-_Statistic

Steedle, J., Kugelmass, H., & Nemeth, A. (2010). What do they measure? Comparing three learning outcomes assessments. *Change: The Magazine of Higher Learning, 42*(4), 33-37. doi:10.1080/00091383.2010.490491

Sundre, D. L., Murphy, C., & Handley, M. (2009). *Advancing assessment of scientific and quantitative reasoning*. The National Numeracy Network. Retrieved from http://serc.carleton.edu/nnn/numeracyprojects/examples/32007.html

Taylor, A. (2000). The effect of traditional classroom assessment on science learning and understanding of the processes of science. *Journal of Elementary Science Education, 12*(1), 19-32. doi:10.1007/BF03176895

Trends in International Mathematics and Science Study (TIMSS). (2007). Retrieved from http://timss.bc.edu/timss2007/index.html

Wright, B. D. (2003). *More art than science: The postsecondary assessment movement today*. Retrieved from http://www.apsanet.org/media/WordFiles/MoreArtThanScience.doc

Appendix A

Email to Science Faculty and Instructors of Science Laboratory Courses

As a professor who teaches a course listed as Division III, you have been selected to be part of the outcomes assessment evaluation. The Liberal Studies Committee will be evaluating your students' work as a part of a programmatic evaluation of liberal studies program. Please provide a sample of your students' work, making sure the sample best demonstrates the liberal studies skills and abilities that students have achieved in your course. Additionally, we need an explanation of how you have assessed your students' work. Examples of students' work could include written papers or essays, projects, tests or final exams. The Liberal Studies Committee decided on this option as possibly the least intrusive method of collection of student work samples. This effort was modeled after successful collection of student work samples from the Division I Humanities, 2006, Upper level Divisions II and IV, 2008 evaluations, and Division V Formal Communications, 2009.

1. How does this course enhance the students' ability recognize and understand the scientific processes?
2. Ability to evaluate various forms of evidence and knowledge
3. Ability to engage in analytical reasoning and
4. How does this course enhance the students' ability to understand and use scientific concepts?
5. How does this course enhance the students' ability to understand and discuss general scientific articles?
6. How does this course enhance the students' ability to apply their knowledge of science to everyday experience?
7. Are the division goals and objectives included as part of the course syllabus?
8. Ability to engage in argumentation and quantitative analysis
9. Ability to engage in scientific inquiry and processes
10. Ability to see across disciplinary boundaries
11. Understanding natural phenomena and the physical world
12. Understanding multiple problem-solving perspectives

Appendix B

Natural Science Rubric

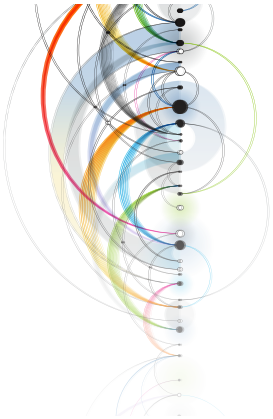
Division III Natural Science description is "These courses primarily focus on scientific and quantitative reasoning and understanding the natural world."

To the Reviewer: Indicate your level of agreement regarding the demonstration of the following components per the learning outcome artifacts reviewed as related to Division III.

	Strongly Disagree	Disagree	Neither Agree Nor Disagree	Agree	Strongly Agree
The learning outcome artifacts of this course (class?) demonstrate **Understanding and Use of Scientific Concepts**	0	1	2	3	4
The learning outcome artifacts of this course (class?) demonstrate **Application of Knowledge of Science to Everyday Experience**.	0	1	2	3	4
The learning outcome artifacts of this course (class?) demonstrate **Recognition and Use of Scientific Reasoning Methods.**	0	1	2	3	4
The learning outcome artifacts of this course (class?) demonstrate **Understanding and Discussion of General Scientific Articles.**	0	1	2	3	4
The learning outcome artifacts of this course (class?) demonstrate **Use of Mathematics in Scientific Reasoning and/or Problem Resolutions**	0	1	2	3	4

• Ability to write and communicate clearly and effectively • Ability to evaluate various forms of evidence and knowledge • Ability to engage in analytical reasoning and argumentation • Ability to engage in quantitative analysis • Ability to engage in scientific inquiry and processes • Ability to see across disciplinary boundaries	• Understanding cultural diversity within the United States • Understanding the world as a diverse and interrelated community • Understanding the relationship of the individual to society and its culture and institutions • Understanding the role of the fine and performing arts and the humanities in shaping and expressing a culture's values and ideals • Understanding natural phenomena and the physical world • Understanding multiple problem-solving perspectives

RESEARCH & PRACTICE IN ASSESSMENT



Abstract

Employment within student affairs divisions offers environments in which students can apply the knowledge they have gained, as well as acquire new competencies, helping them to build solid foundations for their futures. Researchers used an online survey to assess the outcomes associated with part-time student employment within the student affairs division at a large Midwest university. Results show duration of employment, rank, sense of community, civic engagement, and cultural awareness to be strong predictors of student development in preparation for their futures.

AUTHORS

Christina Athas, M.P.H.
The Ohio State University

D'Arcy John Oaks, Ph.D.
The Ohio State University

Lance Kennedy-Phillips, Ph.D.
The Ohio State University

Student Employee Development in Student Affairs

Research about college student development suggests that cognitive, moral, and psychosocial development takes place largely within the academic and social arenas of the institution (Pascarella, 1985). Astin's (1984) student involvement theory illustrates the many connections between student involvement (e.g., studying, time on campus, participation in student organizations) and outcomes, and stresses the importance of focusing pedagogy on the intended outcomes of specific disciplines or programs. Astin proposed two types of college student outcomes: cognitive (e.g., knowledge, decision-making, or critical thinking) and affective (e.g., attitudes, values, or self-concept; Astin, 1984). Outcomes vary, depending upon the type of involvement.

As holistic and life-long learning ideologies are emphasized more strongly in higher education (American College Personnel Association, 1996; Chickering & Reisser, 1993; Dirkxs, 1998; National Association of Student Personnel Administrators & American College Personnel Association, 2004), outcomes associated with college students must encompass a greater breadth of learning and developmental competencies that include not only skills, but personal qualities and attributes that enhance employability, such as those related to self-regulation, critical thinking, and global awareness (Barnett, 2004; Bridgstock, 2009; Brungardt, 2011; Harvey, 2000; Fallows & Steven, 2000; Muldoon, 2009; Pitman & Broomhall, 2009). The university under study refers to these broad skills as *transferable* skills.

CORRESPONDENCE

Email
athas.1@osu.edu

Student affairs divisions are well-positioned to align with such a direction, as they have both a learning-orientation and physical practice spaces. The potential for learning within student affairs divisions can take many different forms; the overarching goal is to provide students with learning opportunities that prepare them for their futures. Thus, an intentional focus on co-curricular learning is important (ACPA, 1996; Kuh, 2009).

As holistic and life-long learning ideologies are emphasized more strongly in higher education, outcomes associated with college students must encompass a greater breadth of learning and developmental competencies that include not only skills, but personal qualities and attributes that enhance employability, such as those related to self-regulation, critical thinking, and global awareness.

The student affairs division at the university under study is committed to engaging in co-curricular learning, encouraging the acquisition of twenty-first century transferable skills and competencies, and continuing a direct and symbiotic relationship with the academic side of the university. Employment within student affairs divisions is a logical setting in which to apply lessons learned in the classroom and foster students' sense of efficacy related to transferable skills. Yet, there remains much to be explored regarding what types of skills and competencies student affairs may help to develop or foster in its student employees. Conceptually, this study of student employees was developed to understand how the work environment created by student affairs professionals influenced student outcomes, namely in the form of transferable skills.

Many studies highlight positive associations between part-time student employment and social and academic outcomes, suggesting that keeping students connected to the university through employment opportunities may in fact improve their performance academically (Brint & Cantwell, 2010; Cheng & Alcántara, 2007; Dundes & Marx, 2006; Fjortoft, 1995; Kulm & Cramer, 2006; Pascarella & Terenzini, 2005), as well as provide opportunities for increased engagement that bridge both academic and "real world" preparation (Fjortoft, 1995; Kuh, 2009; Pascarella & Terenzini, 2005; Shaw & Ogilvie, 2010). In one study, students felt inclined to take on more hours to make their work more meaningful or complete, and felt their work fostered motivation as a result of on-the-job learning, access to a world beyond the immediate campus, and opportunities to interact and network; students also felt that they gained real world experiences and confidence in working with others, as well as insight into the job market (Cheng & Alcántara, 2007). This is in contrast to research regarding off-campus part-time work, which may negatively affect students' connection to campus and their academic success, especially when hours reach or exceed 20 hours per week (Dundes & Marx, 2006; Ehrenberg & Sherman, 1987; Furr & Elling, 2000; Lundberg, 2004). Off-campus employment may also fall short in terms of student growth and development in comparison to on-campus work (Brint & Cantwell, 2010; Kuh, 2009).

Employment within student affairs divisions offers environments in which students can apply the knowledge they have gained, as well as acquire new information, skills, and competencies, helping them to build solid foundations for their futures. University courses are oriented toward particular content; these may not provide clear connections to day-to-day life experiences, while student employment that is external to the university may not provide intentional learning through practical application of previously-acquired classroom knowledge.

The student affairs division within the large, Midwest public university under study employs roughly 4,000 undergraduate and graduate students as student employees during the regular school year. During their tenure as employees, students develop valuable twenty-first century transferable skills and competencies. Those emphasized by the division range from critical thinking, to oral and written communication, time management, and dependability.

In 2007, the student affairs division at this university began a learner initiative, which continues today. The initiative describes common goals for co-curricular student learning; among them are holistic learning for holistic learners, increased intentionality in programming, teaching twenty-first century transferable skills and competencies, and providing transformative experiences to learners. The preferred pedagogy of teaching and learning in the student affairs division often takes the form of constructivism, the idea that learning takes place both individually and socially and is constructed by the meaning attributed to a certain experience (Hein, 1991). The learner initiative incorporates holistic learning, real-world problem solving, and individual contextualized meaning-making, adapted from aspects of the Social Change Model of Leadership Development (Astin & Astin, 1996). Constructs such as "consciousness of self" and "congruence" relate to students' ability to contextualize their experiences, while "commitment" [to leadership], "collaboration," recognition of "common purpose," and "controversy with civility" speak to development of problem-solving skills and competencies. The holistic view of learning incorporates these ideas and seeks to support the notion of "citizenship" within the model through constructivist methods. The phrase, "challenge and support," describes a scaffolded learning environment that incorporates instructional support through resources and appropriate professionals.

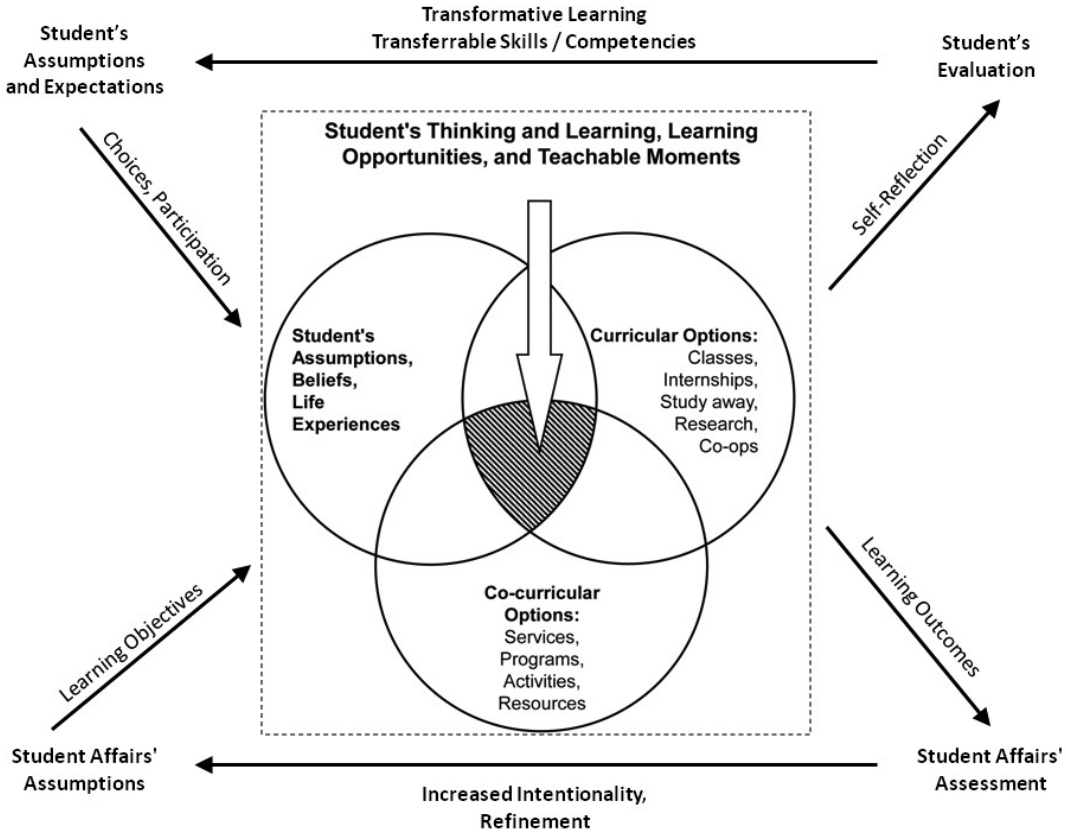


Figure 1. Learning System Map. Adapted from *Learner model & learning system: Concept maps informing practice*, by L.K. Brendon & D.J. Oaks, 2010. Copyright 2010 by the Center for the Study of Student Life.

The Student Employee Outcomes Survey explored the learning environment that the university's student affairs division has created for its student employees. The data and analyses assist the division to leverage its position within students' lives of learning. The current study focuses on the following research questions:

RQ 1: How does the student employee experience provided by the student affairs division foster student development?

RQ 2: What sorts of transferable skills and competencies predict student success related to preparation for the future?

Keeping students connected to the university through employment opportunities may in fact improve their performance academically, as well as provide opportunities for increased engagement that bridge both academic and "real world" preparation.

Method

Instrument Development

At the university under study, the student affairs divisional approach to learning is grounded in a holistic learner model that integrates learning outcomes, wellness dimensions, and social domains (Brendon & Oaks, 2010). Eleven *learner dimensions* represent the aspects of the "whole learner," while four *learner domains* illustrate the areas in which a learner operates (self, others, community, change/society). Double-sided arrows (Figure 2) on each of the eleven dimensions represent development, and the movement between the domains demonstrates the interconnectedness of a particular learning area and the learning dimensions. These dimensions and domains are placed within larger *contexts of learning*, specifically university general education outcomes and student affairs learning outcomes. Two "environments" for learning, curricular and co-curricular initiatives, exist within the institutional context (Brendon & Oaks, 2010).

The Student Employee Outcomes Survey was premised on the merging of two "environments" for learning, the curricular and co-curricular environments. This merging

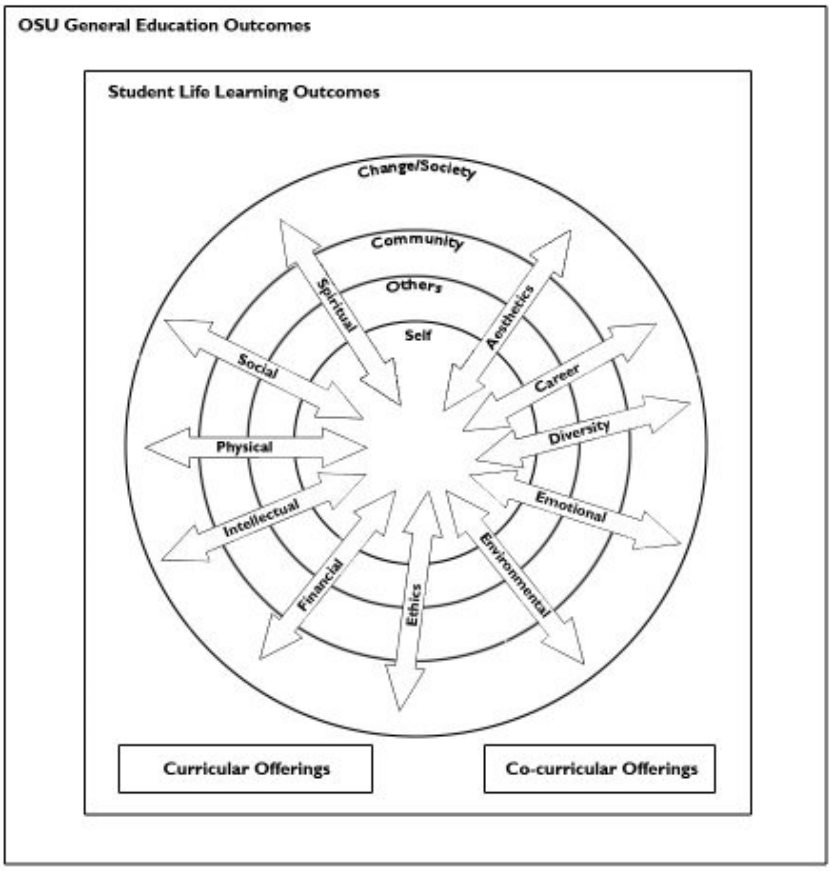


Figure 2. Holistic Learner Model. Adapted from *Learner model & learning system: Concept maps informing practice*, by L.K. Brendon & D.J. Oaks, 2010. Copyright 2010 by the Center for the Study of Student Life.

The initiative describes common goals for co-curricular student learning; among them are holistic learning for holistic learners, increased intentionality in programming, teaching twenty-first century transferable skills and competencies, and providing transformative experiences to learners.

implies that certain kinds of learning take place in the academic (curricular) realm, and certain kinds of learning take place in the co-curricular realm (with overlap). Knowledge and skills acquired from the curriculum can then be applied and practiced through interaction/involvement with the (co-curricular) student affairs realm. In a co-curricular environment, students may apply what was learned in a classroom, cultivate those skills, and may acquire and practice new skills and competencies in a practical setting.

We used two conceptual frameworks to guide the survey items: the Council for the Advancement of Standards (CAS) in Higher Education's *Book of Professional Standards for Higher Education* (2003), and a set of transferable skills developed by the university's student affairs career office. The Council for the Advancement of Standards is comprised of professional organizations consisting of practitioners in higher education student affairs. The council develops and promotes standards that serve as guidelines for student affairs programming and services, and are designed to enhance student development through intentional program improvement. The transferable skills developed by the university's student affairs career office were grounded in the CAS standards and in career services literature. The survey assessed student employees' perceived influence of their employment experience on various skills and attributes. Items were intended to reflect core aspects of higher education learning, as evidenced by the CAS standards and the division's transferable skills of focus, and were reviewed to ensure that the instrument met its intended goal. Survey items related to intrinsic/personal development, self-regulation, leadership/career skills, and career exploration. Each item was measured on a Likert-type scale of 1 to 6 in order to assess perceived influence using the stem "my experience as a student employee has…" to keep responses specific to the experience and minimize the possibility of confounding by maturation. The six-point scale (*Not at All to Greatly*) was used to assess the extent to which working as a student employee influenced the development of attributes and the acquisition of certain transferable skills.

The survey, consisting of 65 items, was reviewed for face validity by an expert panel, which included professionals in career exploration and preparation, higher education research, counseling, student wellness, and human resources.

Participants and Procedure

All full-time undergraduate and graduate students who were employed within student affairs (N=4,092) were invited to take the Student Employee Outcomes Survey; this group of students accounts for approximately 10% of the university population and included part-time paid student employees, work-study employees, paid interns, and unpaid interns. No exclusions were made beyond employment within the student affairs division. The survey was administered through a secure, web-based server. Students were identified via a computer-generated list from the human resources database and were invited to participate via e-mail. To bolster the response rate during online data collection, participants were offered the chance to be one of six winners of a $50.00 student ID card cash deposit.

Data were collected over a four-week period, during which students received an invitation e-mail and up to three reminders (sent once per week to students who had not completed the survey). By the close of the survey, 1,415 students responded, yielding a 34.5% response rate. The authors found the sample to be representative of the overall university population. Data were analyzed using Statistical Software for the Social Sciences (SPSS) 17.0.

Analysis

The authors followed a two-step analysis advocated by Wang and Kennedy-Phillips (2013). A Principal Component Analysis (PCA) was conducted with the intention of reducing the data into manageable summated scales. The PCA analyzed 65 items on the survey that addressed student perception of growth in each area as a result of the work environment. According to Cudeck and MacCallum (2007), "An eigenvalue is the variance explained by the components in a PCA" (p. 190). Using the Kaiser criterion, only components with an eigenvalue greater than 1 were retained (Appendix A). A Varimax rotation was used in the development of the component structure. The components that emerged became the five scales chosen to represent the constructs of the measured dependent variables: interpersonal skills, personal wellness awareness, practical skill acquisition, academic self-efficacy, and self-awareness, and three predictors: community involvement, civic engagement, and cultural competencies. Scale means based on these components were then included in the regression models predicting the outcome measures of student growth in the work environment.

In a co-curricular environment, students may apply what was learned in a classroom, cultivate those skills, and may acquire and practice new skills and competencies in a practical setting.

Five separate Ordinary Least Squares (OLS) regression models were developed through a non-iterative approach to predict how a student's personal and academic growth were affected by the work environment. The five OLS models represented a test of the five independent variables resulting from the PCA. In general, a student's growth in the areas was assumed to be a function of background characteristics (gender, rank, residence, hours worked, and duration of employment) and civic involvement (community involvement, civic engagement, and cultural competencies). Models were tested to assess the relevant importance of each set of independent variables in predicting students' perception of growth in the student affairs work environment.

Dependent and Independent Variables

The dependent variables consisted of five summated scales (interpersonal skills, personal wellness awareness, practical skill acquisition, academic self-efficacy, and self-awareness) that represented the learning environment fostered by student employment within the division of student affairs. All dependent variables were derived from a PCA explained in the analytical approach section. The independent variables included the following background variables: gender (dummy variable coded in male = 0 and female = 1), rank (dummy variable coded into under-class = 0 and upper-class = 1), hours worked (dummy variable coded >10 hours = 0 and < 10 hours = 1), duration of employment in the division (dummy variable coded >3 quarters = 0 and < 3 quarters = 1), and finally, residence (dummy variable coded on-campus = 0 and off-campus = 1). In addition to the background variables, the independent variables included three measures of civic involvement: community involvement, civic engagement and cultural competencies. These, similar to the dependent variables, were mean scales derived

from a PCA. The model hypothesized that students' perceptions of community involvement, civic engagement, and cultural competence were predictors of the five summated dependent variables. Definitions of each component were derived from the individual items (see Appendix A). All independent and dependent variables were self-reported. Descriptive statistics on each variable are provided in Tables 1 and 2.

Table 1
Descriptive Statistics for the Sample

Variable	Percentage
Female	61
Upper Class	52
Hours worked <10 hours	44
Duration >3 quarters	55
Off Campus	50

Note. N=1,415

Table 2
Descriptive Statistics for the Student Employee Outcomes Survey Scaled Items

Variable	Mean	*SD*
Independent Variables		
Community Involvement	4.7	1.1
Cultural Competencies	4.7	1.1
Civic Engagement	3.6	1.4
Dependent Variables		
Interpersonal Skills	4.6	1.0
Personal Wellness Awareness	4.6	1.0
Practical Skill Acquisition	4.5	1.0
Academic Self-Efficacy	4.1	1.4
Self-Awareness	4.3	1.1

Note. Variables are measured on a scale of 1-6 with higher values indicating a greater degree

Results

Under-class students reported greater development of interpersonal skills than upper-class students.

The following summarizes the results of the regression analyses. All models accounted for at least 40% of the variance in students' perception of growth within the five areas of development (Appendix B). Component labels were developed based upon the individual items that informed the emergence of the component.

Interpersonal Skills

Model 1 summarizes the predictors of student employees' perceived growth in their interpersonal skills as a result of employment in the division of student affairs (R^2= .68, p< .05). When considering the background variables, rank was the only significant background predictor of interpersonal skill growth in the work environment. Under-class students reported greater development of interpersonal skills than upper-class students. All three civic-involvement variables were significant predictors of growth in interpersonal skills in the student affairs work environment. The more students positively identified with community involvement, cultural competencies and civic engagement, the more growth they perceived in their interpersonal skills. Community involvement was the strongest predictor.

Personal Wellness Awareness

In model 2 (R^2 = .53, p < .05), two background measures, rank and residence, significantly predicted students' perceived growth in personal wellness. As in model 1, under-class students

reported developing a higher level of personal wellness awareness in the work environment than did upper-class students. Students who lived off campus reported a higher growth of personal wellness awareness than did students who lived on campus. All three civic-involvement variables were significant predictors of growth in personal wellness awareness in the student affairs work environment. The more students perceived the work place to develop their community involvement, the higher their perceived personal wellness awareness.

Practical Skill Acquisition

In model 3 (R^2 = .57, p < .05), gender was a significant predictor of practical skill acquisition. Female students reported that they gained greater practical skill acquisition in comparison to male students. Additionally, the more students positively identified with community involvement, cultural competencies and civic engagement, the more growth they perceived in their practical skill acquisition. Civic engagement was the strongest predictor of a student's perception of skill acquisition.

Academic Self-Efficacy

In model 4 (R^2 = .49, p < .05), rank and duration of employment, two of the five background characteristics, were significant predictors of academic self-efficacy. The longer students were employed within the student affairs division, the more academically self-efficacious they reported that they were. Under-class students reported that they were more academically self-efficacious as a result of the work environment than did upper-class students. All three civic-involvement scales were significant predictors of academic self-efficacy. Civic engagement was the strongest predictor. The more socially engaged students were, the higher their perception of academic self-efficacy.

Self-Awareness

In model 5 (R^2 = .58, p < .05), none of the background characteristics significantly predicted self-awareness. As with the other four models, community involvement, cultural competencies and civic engagement were significant predictors of self-awareness. Civic engagement was the greatest predictor.

Discussion

The data suggest that students perceive their student employee experiences in this university's student affairs division to be instrumental in their skill development in a variety of areas. Rank was a predictor of interpersonal skills, personal wellness awareness, and academic self-efficacy. Regarding interpersonal skills, under-class students reported greater perceived growth than upper-class students. One reason may be that many under-class students typically participate in an on-campus lifestyle, which includes a strong climate for social engagement (Astin, 1984). This environment, coupled with engagement within the student employee experience, may help students develop a variety of interpersonal skills useful in a future career (Harvey, 2000; Muldoon, 2009). These skills can include understanding repercussions of actions, admitting mistakes, resolving conflict respectfully, communicating effectively, working as part of a team, providing constructive criticism, fostering integrity, learning patience, and becoming a more tolerant person.

The longer students were employed within the student affairs division, the more academically self-efficacious they reported that they were.

Rank was also a predictor of students' perceived growth in personal wellness awareness. Under-class students reported greater perceived growth than did upper-class students. Personal wellness awareness includes skills and competencies such as time management, productive lifestyle, self-sufficiency, work-life balance, responsibility, dependability, organization, money management and timely decisions. Findings such as those of Watts and Pickering (2000) indicate that undergraduate students expressed a great deal of importance on organization when balancing part-time work with their academic and social lives, though it is unknown whether this holds true across different undergraduate ranks. Other studies also share findings that suggest that part-time student employment fosters aspects of personal wellness, such as self-reliance, responsibility, and dependability (Curtis & Shani, 2002; Curtis & Williams, 2002).

As a predictor of perceived growth in academic self-efficacy (confidence in academic and career goals, motivation to pursue further academic endeavors), perceived gain was higher

Students reported that, as they maintained longevity in working in student affairs departments, they had higher levels of motivation to pursue education, increased motivation to work on their academic pursuits, and were better able to clarify their academic goals and solidify their career goals. The findings suggest that relationships exist between the curricular and the co-curricular realms of the university, and that students perceive a strong link between their employee experience and their academic endeavors.

for under-class students. This may be in part because for under-class students, the student employee experience helps to clarify skills and interests, and allows for the exploration of different career possibilities, as expressed in Chang and Alcántara's (2007) study.

While maturity has been cited as an outcome of part-time paid student employment (Dustmann et al., 1996), it is difficult within the context of our study to ascertain to what extent maturation or the maturation of particular skills was/were a direct result of employment as opposed to natural growth and development, over the course of students' time at the university. Though we did try to control for this phenomenon to some extent by using the stem, "My experience as a student employee has…" we can only suggest an association between the student employee experience and such development.

Duration of employment also predicted academic self-efficacy. The longer that students remained employed within the student affairs division, the greater their perceived growth related to areas of academic self-efficacy. This is in accordance with Kulm and Cramer's (2006) findings regarding the relationship between the length of employment and persistence toward a degree. Students reported that, as they maintained longevity working in student affairs departments, they had higher levels of motivation to pursue education, increased motivation to work on their academic pursuits, and were better able to clarify their academic goals and solidify their career goals. The findings suggest that relationships exist between the curricular and the co-curricular realms of the university, and that students perceive a strong link between their employee experience and their academic endeavors. These findings, supported by the literature (Brint & Cantwell, 2010; Cheng & Alcántara, 2007; Dundes & Marx, 2006; Fjortoft, 1995; Kulm & Cramer, 2006; Pascarella & Terenzini 2005), further suggest that students benefit when they choose jobs within student affairs, since these positions are tied closely to the university, and thus help keep students academically and socially engaged.

The finding that students' employment experiences helped them to solidify career goals suggests that jobs within student affairs divisions may be instrumental in helping students make decisions that affect their futures. Studies, such as that reported by Cheng and Alcántara (2007), indicate that on-campus work may play an important role in helping students shape their academic interests and career choices. This suggests that student affairs divisions should strengthen relationships with academic affairs divisions in order to intentionally create opportunities within student employee positions that connect to academic endeavors.

Residence was a predictor of personal wellness awareness. Students who reported living off campus indicated greater perceived growth than did those who lived on campus. At the university under study, first-year students reside on campus, while upper-class students tend to move off-campus. According to the data, approximately 3% of first-year, 37.4% of second-year, 64.5% of third-year, 75.1% of fourth-year, and 88.7% of fifth-year or more students lived off campus. Other options included on campus or with parent(s)/guardian(s). It may be that students who live off campus perceive greater benefit from on-campus employment due to interaction with campus that they might not normally experience as part of the off-campus lifestyle. Student development literature (e.g., Astin, 1984; Pascarella, 1985) consistently cites the learning and developmental benefits associated with on-campus interactions, and thus, greater gains might be realized as a result of the lack of this interaction. Further research is needed to fully understand this phenomenon.

Another predictor of students' perceived growth was gender as it related to skill acquisition. Females reported greater perceived growth in skill acquisition as compared to males, which could be explained by further research that explores gender differences related to perception of growth in this area. Baxter Magolda (2004) suggests that there are gender differences in intellectual development. Specifically, females tend to listen and absorb information, while males more often practice and master information, though it remains unclear how exactly this might translate to greater perceived growth.

A fourth predictor of reported growth was sense of community. Students who felt a greater sense of community (e.g., meaningful friendships, sense of belonging) reported higher levels in interpersonal skills, self-awareness, personal wellness awareness, skill acquisition, and academic self-efficacy. These findings suggest that when students feel as though their student employment experience has fostered a sense of community, this helps them feel connected to the university and provides them with a comfortable environment within which

they can exercise interpersonal skills, learn new skills, focus their academic and career goals, and improve personal wellness, which falls in line with previous research (Cheng & Alcántara, 2007; Fjortoft, 1995; Kuh, 2009; Pascarella & Terenzini, 2005; Shaw & Ogilvie, 2010), and mirrors work by Astin (1984) that documents the many developmental benefits of on-campus community. Studies such as Swanson, Broadbridge, and Karatzias' (2006) suggest that on-campus employment facilitates adjustment to the university, and cites student self-reported benefits such as perceived long-term employment benefits and the enhancement of personal development and social involvement.

The finding that students' employment experiences helped them to solidify career goals suggests that jobs within student affairs divisions may be instrumental in helping students make decisions that affect their futures.

Cultural competencies predicted students' perceived growth in multiple areas; the more exposed students were to other cultures, the greater their reported growth in interpersonal skills, self-awareness, personal wellness awareness, skill acquisition, and self-efficacy as a result of their student employment. Students who believed that their employment experience expanded their interactions with people of diverse backgrounds and increased their awareness of other cultures seemed to gain a greater benefit in other areas; students who reported that they dealt with individuals from different cultures reported that they perceived greater personal gains in developing a better understanding of themselves and their values than did students who did not report that they dealt with different types of people, which builds upon the findings of Cheng and Alcántara (2007), who suggest that students feel their horizons are broadened beyond the university scope as a result of on-campus employment.

The final predictor of students' perceived growth related to civic engagement. Students who felt that their employment experience exposed them to national and global issues and motivated them to be involved in their community reported greater perceived growth in interpersonal skills, self-awareness, personal wellness awareness, skill acquisition, and academic self-efficacy. Intertwining social and civic awareness into the student employee experience provides opportunities to bridge academic areas with co-curricular areas to provide structured, multi-dimensional learning experiences.

Perceived growth in the aforementioned areas indicates that student development takes place within the student affairs student employment experience. The regression analyses suggest that there are a number of variables that predict development and preparation for the future, indicating aspects which student affairs may be able to foster through intentional student employment practices.

When students feel as though their student employment experience has fostered a sense of community, this helps them feel connected to the university and provides them with a comfortable environment within which they can exercise interpersonal skills, learn new skills, focus their academic and career goals, and improve personal wellness.

Future Research

The topic of student development as it relates to university employment is an area of growing research, and there are a number of aspects still to be addressed. More research is needed to assess gender as a predictor of perceived growth throughout the student employment experience. Males and females reported varying degrees of skill acquisition (e.g., learning new skills, realizing a greater potential in oneself), and more research is required to examine these differences. It is also important to further explore the needs and interests of first- and second-year students, as they relate to employment. Rank was associated with a number of components related to interpersonal skills, academics, and personal wellness. Such associations require further investigation to determine the differences among ranks, as well as the aspects of development that are attributable to employment experiences rather than general maturation. Further research is also needed in regard to civic engagement and its relation to student development within the context of student employment. Knowledge in this area would help to clarify the benefits of this form of engagement, and inform potential programming designed to bridge curricular and co-curricular civic engagement experiences.

There are some larger questions that this study did not address. First, we did not address the ways in which the five components (interpersonal skills, personal wellness awareness, practical skill acquisition, academic self-efficacy, and self-awareness) interacted with each other. It is likely that there are important connections to be noted, and further analysis is necessary to delineate these associations. Second, this study did not examine development according to the type of job the student employee held. Development may vary depending upon the job type, and further study would help to illuminate differences and inform training and programming efforts to ensure that all student employment opportunities achieve well-rounded student development.

Conclusion

This study measured outcomes related to employment within student affairs at a large Midwestern university. Further research might expand beyond student affairs to include other employment opportunities both within the university and outside of the university. Such research would be an opportunity to compare learning experiences of other employment experiences to those within student affairs. This study examined a number of developmental factors related to college student development within the context of university employment. While many implications for practice can be drawn from the associations found in this study, more research is necessary to fully understand the ways in which student employment benefits students during their time at the university, as well as beyond.

Student affairs units offer places to apply lessons learned in the classroom and to acquire new skills and competencies both through programming and employment. This analysis suggests that student affairs divisions bridge curricular and co-curricular learning and shows that the variables of duration of employment, rank, community involvement, civic engagement, and cultural competencies are strong predictors of personal development within the student employee experience.

References

American College Personnel Association. (1996). The student learning imperative: Implications for student affairs. *American College Personnel Association*.

Astin, A. W. (1984). Student involvement: A developmental theory for higher education. *Journal of College Student Development, 40*(5), 518-529.

Astin, H., & Astin, A. (1996). *A social change model of leadership development: Guidebook*. Version III. Los Angeles, CA: Higher Education Research Institute.

Barnett, R. (2004). Learning for an unknown future. *Higher Education Research and Development, 23*(3), 247-260.

Baxter Magolda, M.B. (2004). Evolution of a constructivist conceptualization of epistemological reflection. *Educational Psychologist, 39*(1), 31-42.

Brendon, L.K., & Oaks, D.J. (2010, March). *Learner model & learning system: Concept maps informing practice*. Paper presented at the 2010 NASPA Conference in Chicago. PowerPoint retrieved from http://cssl.osu.edu/posts/documents/2010-naspa-learner-model-learning-system-present.pdf

Bridgstock, R. (2009). The graduate attributes we've overlooked: Enhancing graduate employability through career management skills. *Higher Education Research and Development, 28*(1), 31-44.

Brint, S., & Cantwell, A. M. (2010). Undergraduate time use and academic outcomes: Results from the University of California Undergraduate Experience Survey 2006. *Teachers College Record, 112*(9), 2441-2470.

Brungardt, C. (2011). The intersection between soft skill development and leadership education. *Journal of Leadership Education, 10*(1), 1-22.

Cheng, D. X., & Alcántara, L. (2007). Assessing working students' college experiences: A grounded theory approach. *Assessment & Evaluation in Higher Education, 32*(3), 301-311.

Chickering, A. W., & Reisser, L. (1993). *Education and identity* (2nd ed.). San Francisco, CA: Jossey-Bass.

Council for the Advancement of Standards in Higher Education. (2003). *The book of professional standards for higher education* (3rd Rev. ed.). Washington, DC: Author.

Cudeck, R., & MacCallum, R.C. (Eds.). (2007). *Factor analysis at 100: Historical developments and future directions*. Mawah, NJ: Lawrence Erlbaum Associates, Inc.

Curtis, S., & Shani, N. (2002). The effect of taking paid employment during term-time on students' academic studies. *Journal of Further and Higher Education, 26*(2), 129-138.

Curtis, S., & Williams, J. (2002). The reluctant workforce: Undergraduates' part-time employment. *Education and Training, 44*(1), 5-10.

Dirkxs, J.M. (1998). Transformative learning theory in the practice of adult education: An overview. *PAACE Journal of Lifelong Learning, 7*, 1-14.

Dundes, L., & Marx, J. (2006). Balancing work and academics in college: Why do students working 10-19 hours per week excel? *Journal of College Student Retention, 8*(1), 107-120.

Dustmann, C., Michlewright, J., Rajah, N., & Smith, S. (1996). Earning and learning: Educational policy and the growth of part-time work by full-time pupils. *Fiscal Studies, 17*(1), 79-103.

Ehrenberg, R., & Sherman, D. (1987). Employment while in college, academic achievement, and post-college outcomes: A summary of results. *Journal of Human Resources, 22*, 1–23.

Fallow, S., & Steven, C. (2000). Building employability skills into the higher education curriculum: A university-wide initiative. *Education & Training, 40*(2), 75-83.

Fjortoft, N. F. (1995). College student employment: opportunity or deterrent? Presented at the 1995 AERA Meeting, San Francisco, CA.

Furr, S., & Elling, T. W. (2000). The influence of work on college student development. *NASPA Journal, 37*(2), 454–470.

Harvey, L. (2000). New realities: The relationship between higher education and employment. *Tertiary Education and Management, 6*, 3-17.

Hein, G. E. (1991). *Constructivist learning theory*. Presented at CECA International Committee of Museum Educators Conference, Jerusalem, Israel.

Kulm, T. L., & Cramer, S. (2006). The relationship of student employment to student role, family relationship, social interactions and persistence. *College Student Journal, 40*(4), 927-938.

Kuh, G. D. (2009). What student affairs professionals need to know about student engagement. *Journal of College Student Development, 50*(6), 683-706.

Lundberg, C. A. (2004). Working and learning: The role of involvement for employed students. *NASPA Journal, 41*(2), 201–215.

Muldoon, R. (2009) Recognizing the enhancement of graduate attributes and employability through part-time work while at university. *Active Learning in Higher Education, 10*(3), 237-252.

National Association of Student Personnel Administrators & American College Personnel Association. (2004). Learning reconsidered: A campus-wide focus on the student experience. *National Association of Student Personnel Administrators; American College Personnel Association*.

Pascarella, E. T. (1985). Students' affective development within the college environment. *The Journal of Higher Education, 56*(6), 640-663.

Pascarella, E.T., & Terenzini, P.T. (2005). *How college affects students* (Vol. 2). San Francisco, CA: Jossey-Bass.

Pitman, T., & Broomhall, S. (2009). Australian universities, generic skills and lifelong learning. *International Journal of Lifelong Education, 28*(4), 439-458.

Shaw, S., & Ogilvie, C. (2010). Making a virtue out of a necessity: Part time work as a site for undergraduate work-based learning. *Journal of European Industrial Training, 34*(8/9), 805-821.

Swanson, V., Broadbridge, A., & Karatzias, A. (2006). Earning and learning: Role congruence, state/trait factors and adjustment to university life. *British Journal of Education Psychology, 76*, 895-914.

Wang, X., & Kennedy-Phillips, L. (2013). Focusing on the sophomores: Characteristics associated with the academic and social involvement of second-year college students. *Journal of College Student Development, 54*(5), 541-548.

Watts, C., & Pickering, A. (2000). Pay as you learn: Student employment and academic progress. *Education and Training, 42*, 129-134.

Appendix A

Student Employee Outcomes Survey Instrument Scales and Constructs

Interpersonal Skills	R^2 = .68	Eigenvalue = 3.15
Ability to admit mistakes	Made more approachable	
Consider repercussions of actions	Ability to take initiative	
Ability to think before acting	Ability to take direction/follow instructions	
Ability to communicate effectively	Improved critical thinking skills	
Ability to resolve conflict respectfully	Made more tolerant person	
Ability to express thoughts/opinions clearly	Ability to remain focused on individual tasks	
Ability to weigh different perspectives	Ability to provide constructive criticism	
Ability to comfortably interact with others	Increased attention to detail	
Ability to work as part of a team	Helped to learn patience	

Personal Wellness Awareness	R^2 = .53	Eigenvalue = 2.97
Ability to make timely decisions	Improved time management skills	
Transitioned into more productive lifestyle	Made more responsible in everyday actions	
Helped better manage money	More dependable person	
Made more self-sufficient	Improved organizational skills	
Improved work-life balance		

Practical Skill Acquisition	R^2 = .57	Eigenvalue = 1.72
Allowed to acquire new skills	Introduced to skills didn't know I had	
Helped to realize greater potential in self	Pushed me beyond what I thought to be my capabilities	

Academic Self-Efficacy	R^2 = .49	Eigenvalue = 1.41
Motivated pursuit of a higher level of education	Increased motivation to work on academics	
Solidify career goals	Clarify academic goals	

Self-Awareness	R^2 = .58	Eigenvalue = 1.26
Helped to solidify values	Helped add value to life	
Helped to develop a better understanding of self	Gave greater sense of purpose	

Cultural Competences	R^2 = .62	Eigenvalue = 1.21
Expanded my interactions with people of diverse backgrounds		
Increased my awareness of other cultures		

Civic Engagement	R^2 = .52	Eigenvalue = 1.12
Opened my eyes to national issues	Opened my eyes to global issues	

Community Involvement	R^2 = .79	Eigenvalue = 1.01
Motivated me to become more involved with my community		
Brought me closer to my community		

Appendix B

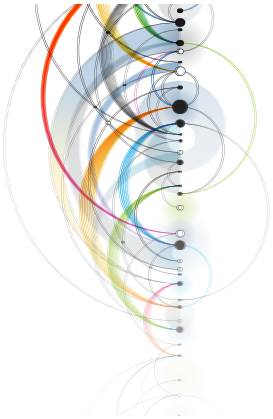
Summary of OLS Regression Analysis for Variables Predicting Students' Perception of Growth in the Student Affairs Work Environment

	Interpersonal Skills		Personal Wellness Awareness		Practical Skill Acquisition		Academic Self-Efficacy		Self-Awareness	
Independent Variable	*B (SE)*	*ß*	*B (SE)*	*ß*	*B (SE)*	*ß*	*B (SE)*	*ß*	*B (SE)*	*ß*
Female	.020 (.034)		.044 (.040)		.145 (.048)	.056**	.038 (.058)		-.007 (.044)	
Upper Class	-.087 (.039)	.043**	-.176 (.046)	.087**	-.052 (.055)		-.127 (.066)	-.045**	.079 (.050)	
Hours worked (< 10 hrs)	-.016 (.033)		-.033 (.039)		.024 (.047)		.060 (.057)		.017 (.043)	
Duration (> 3 quarters)	-.003 (.035)		-.019 (.042)		-.018 (.050)		-.123 (.060)	-.044**	-.034 (.045)	
Off-campus	.047 (.037)		.081 (.044)	.040*	.033 (.053)		.067 (.064)		-.052 (.048)	
Community Involvement	.282 (.017)	.317**	.306 (.021)	.338**	.242 (.025)	.215**	.144 (.030)	.115**	.430 (.022)	.412**
Cultural Competencies	.258 (.017)	.316**	.230 (.020)	.280**	.216 (.024)	.211**	.167 (.029)	.147**	.154 (.022)	.162**
Civic Engagement	.261 (.015)	.363**	.195 (.018)	.270**	.420 (.021)	.465**	.537 (.026)	.537**	.287 (.019)	.342**
Model summary	$R^2 = .68$ $F = 3.2$, $p < .05$		$R^2 = .53$ $F = 1.8$, $p < .05$		$R^2 = .57$ $F = 2.1$, $p < .05$		$R^2 = .49$ $F = 1.5$, $p < .05$		$R^2 = .58$ $F = 2.2$, $p < .05$	

Note. N=1,415. Betas are reported for statistically significant coefficients only.

*** $p < .01$, ** $p < .05$, * $p < .10$

RESEARCH & PRACTICE IN ASSESSMENT



Abstract

Examinee effort can impact the validity of scores on higher education assessments. Many studies of examinee effort have briefly noted gender differences, but gender differences in test-taking effort have not been a primary focus of research. This review of the literature brings together gender-related findings regarding three measures of examinee motivation: attendance at the assigned testing session, time spent on each test item, and self-reported effort. Evidence from the literature is summarized, with some new results presented. Generally, female examinees exert more effort, with differences mostly at very low levels of effort—the levels at which effort is most likely to impact test scores. Examinee effort is positively correlated with conscientiousness and agreeableness, and negatively correlated with work-avoidance. The gender differences in these constructs may account for some of the gender differences in test-taking effort. Limitations and implications for higher education assessment practice are discussed.

AUTHORS

Christine E. DeMars, Ph.D.
James Madison University

Bozhidar M. Bashkov, M.A.
James Madison University

Alan B. Socha, M.A.
James Madison University

The Role of Gender in Test-Taking Motivation under Low-Stakes Conditions

Test-taking motivation is important to many university assessment efforts, because higher education assessments often have low or no consequences for individual students but high consequences for the university. Test-taking motivation has been extensively studied. Examinees score higher when the test has some stakes for them, such as a grade in a course (Sundre, 1999; Sundre & Kitsantas, 2004; Terry, Mills, & Sollosy, 2008; Wolf & Smith, 1995; Wolf, Smith, & DiPaulo, 1996), course placement, promotion, graduation (DeMars, 2000), admissions (Cole & Osterlind, 2008), hiring decisions (Rothe, 1947), or simply knowing that faculty and employers will see the scores (Liu, Bridgeman, & Adler, 2012). Following the terminology of Wise (2009, p. 154), the phrase *low-stakes* will be used here to describe tests with no personal stakes for examinees, regardless of the stakes for institutions or instructors.

When the test has no personal stakes, examinees who report higher effort tend to score somewhat higher (Cole, Bergin, & Whittaker, 2008; Eklöf, 2007; Schiel, 1996; Sundre & Kitsantas, 2004; Wolf & Smith, 1995). As a result, examinees' levels of proficiency will likely be underestimated when examinees do not give their best effort to a low-stakes test. Specifically, lack of examinee motivation can impact the reliability and validity (e.g., increase construct-irrelevant variance) of the inferences one can make from test scores. This includes inferences about gender differences in test scores.

CORRESPONDENCE

Email
demarsce@jmu.edu

Although gender differences have seldom been the primary focus of motivation studies, many studies have briefly noted gender differences in test-taking motivation among university students on low-stakes tests. The purpose of this integrative review is to bring together a variety of evidence to illustrate ways in which these differences are revealed. When available, new data are described after each section to add to the existing evidence from the published literature. Finally, potential explanations of gender differences in test-taking motivation are examined, as well as implications for higher education assessment practice.

Lack of examinee motivation can impact the reliability and validity (e.g., increase construct-irrelevant variance) of the inferences one can make from test scores.

Absence at Test Administration

Students who are extremely unmotivated may simply not show up at the assigned test administration. Swerdzewski, Harmes, and Finney (2009) studied a group of students who failed to attend an assigned testing session but later attended a make-up session. Although scores on the test had no consequences, students could not enroll for the following semester until the assessment requirement was completed; most students eventually complied and came to one of the make-up sessions. Those who attended the regular session were labeled Attenders and those who attended the make-up session were labeled Avoiders.

Male students were less likely to attend the regular session. Aggregating data from tables and text included in Swerdzewski et al. (2009), and assuming that those who provided complete data at the make-up session were representative of the total group of Avoiders and that those in the comparison group were representative of the total group of Attenders, about 30% of male students, compared to 22% of female students, failed to attend the regular testing session.

Although some students likely missed the regular testing session due to reasons other than willful noncompliance, three pieces of evidence suggest that a large portion of the non-attendance was related to motivation. First, Avoiders scored much lower than the Attenders on a fine arts test and a science test (-0.74 and -0.77 standard deviation units, respectively). Second, only 12.7% of Avoiders tried on at least 90% of the items, compared to 47.2% of Attenders. Finally, self-reported effort was 0.42 standard deviation units higher for Attenders.

It might be argued that the Avoiders skipped the assessment and then performed poorly and exerted little effort in the make-up session simply because they had low levels of knowledge. The Avoiders did have somewhat lower average grades (2.80 compared to 3.02), but this difference does not seem large enough to explain their difference in test performance. Further, SAT scores were equivalent for the two groups (1165 compared to 1163).

Overall, it appears that male students are less likely than female students to exert even the minimal effort to show up for an assigned testing session. Not attending the testing session may represent extremely low levels of test-taking motivation.

Rapid Guessing

Another indicator of very low motivation is responding to test items without taking the time to read the question. On a partly-speeded test, rapid guessing may occur toward the end of the test, if examinees run out of time. When time limits are ample or nonexistent, extremely rapid responding is more likely to indicate that the examinee put no effort into selecting an answer. Schnipke (1995) coined the term *solution behavior* to describe responses in which the examinee attempted to choose the correct answer and *rapid guessing behavior* to describe responses in which the examinee simply rapidly chose a response. Wise and Kong (2005) proposed the *response time effort* (RTE) index, the percent of items on which an examinee engaged in solution behavior.

RTE provides an unobtrusive way to collect motivation data for each item, and thus does not rely on examinee judgments of their own motivation (Kong, Wise, & Bhole, 2007; Wise & Kong, 2005). RTE is based on the notion that examinees who are not motivated will exhibit rapid guessing behavior; that is, they will rapidly respond to items without taking the time to read or fully consider the items. With this approach, response times are a proxy for motivation. Thus, rapid guessing can decrease test score validity (Wise & DeMars, 2010). In part, this is because the correctness of answers resulting from rapid guessing behavior will be at or near chance levels, as the answers are essentially random (Wise & DeMars, 2010; Wise & Kong, 2005). RTE scores have high internal consistency, are correlated with other measures of test-taking effort, and are uncorrelated with external measures of proficiency (Wise & DeMars, 2010; Wise & Kong, 2005). Because RTE is based on response times, this method is only feasible with computer-based tests where the software permits collection of response times. RTE scores have several uses: (a) to indicate levels of examinee effort, (b) to provide information on the dynamics of examinee motivation, and (c) to supply data for motivation filtering (Wise & Kong, 2005).

To compute RTE, examinee item responses are first classified as either exhibiting rapid guessing behavior (i.e., when examinees appear to supply an answer without considering the item) or solution-based behavior (i.e., when examinees attempt to find the best answer for the item; DeMars, 2007; Kong et al., 2007; Wise, 2009; Wise & Cotton, 2009). Thus, a time threshold defining response times that are too short for an examinee to have a chance to read and consider the item must be set for each item (DeMars & Wise, 2010; Swerdzewski, Harmes, & Finney, 2011). Several methods exist for setting the time threshold (see DeMars, 2007; Kong et al., 2007; Swerdzewski et al., 2011).

An index, item solution behavior (SB_{ij}), is then assigned a value of 0 or 1 based on whether an examinee's response time is below (i.e., rapid guessing behavior) or above (i.e., solution-based behavior) the threshold (DeMars & Wise, 2010; Kong et al., 2007; Wise & Kong, 2005). Because the thresholds are based on the minimum amount of time needed to read and consider the items, this index will only identify responses which the researchers are reasonably certain are noneffortful (Kong et al., 2007). The proportion of items on which the examinee exhibited solution behavior is the examinee's RTE score (DeMars, 2007; DeMars & Wise, 2010; Kong et al., 2007; Swerdzewski et al., 2011; Wise, 2009; Wise & Cotton, 2009; Wise & DeMars, 2010; Wise & Kong, 2005).

RTE values can be used in motivation filtering (Swerdzewski et al., 2011; Wise & Cotton, 2009; Wise & Kong, 2005). In fact, motivation filtering using RTE scores has been found to be favorable compared to using self-reported measures (Wise & Kong, 2005). Motivation filtering involves removing data from unmotivated examinees. Doing so should result in higher mean test scores, lower test score standard deviations, and higher correlations between test scores and external measures (i.e., convergent validity evidence) of ability when examinee effort is not related to actual proficiency (DeMars, 2007; Wise & Cotton, 2009; Wise & Kong, 2005; Wise & DeMars, 2010). In this case, motivation filtering reduces construct-irrelevant variance (Wise & Cotton, 2009). If, however, examinee effort were related to actual proficiency, data would be filtered from the lower part of the proficiency distribution, which would artificially inflate the mean of the remaining scores (Wise & DeMars, 2010). Thus, external measures of proficiency should be used to examine whether examinee effort (i.e., RTE) is related to proficiency prior to motivation filtering. For example, Wise and Kong (2005) showed that filtering students based on RTE on a university assessment made no difference in the average SAT score before and after filtering.

Importantly, gender differences can be misestimated if examinee motivation is not taken into account.

Importantly, gender differences can be misestimated if examinee motivation is not taken into account. For example, some studies found that female students exhibit more solution-based behavior than male students (Wise & Cotton, 2009; Wise & DeMars, 2010). One study found that, when motivation differences were ignored, female students showed sizeable gains between two time periods, whereas male students showed virtually no gains (Wise & DeMars, 2010). However, when examinees with the lowest RTE were removed from the data, both male and female students showed clear gains. Thus, without taking motivation into account, observed differences in mean score changes may misrepresent the actual difference in mean changes by the degree to which there are differences in rapid guessing behavior between the groups. In a study of middle school and high school students (Wise, Kingsbury, Thomason, & Kong, 2004), only 27 out of 2,382 students had RTE scores less than .90, but 23 of these 27 students were boys. Freund and Rock (1992) studied a behavior conceptually related to rapid guessing: pattern-marking (random marking of responses or systematic strings such as ABCDABCD). On the National Assessment of Educational Progress (NAEP), pattern-marking was more common among male adolescents than female adolescents, and the gender gap was greater among high school seniors than among 8th graders.

However, not all studies have found gender differences in RTE. On a test of scientific and quantitative reasoning administered under low-stakes conditions, gender was only very slightly correlated with RTE (Wise, Pastor, & Kong, 2009).

Empirical Study

RTE data were available from a science test administered to a random sample of university students and four business tests administered to students majoring in business. The science test was administered to a random sample of university students with 45-70

Far more men than women exhibited rapid guessing behavior on over half of the items.

cumulative credit hours during the spring 2009, spring 2010, and spring 2012 semesters. The test is used to directly measure objectives of the General Education program. It is low-stakes for students. The business tests were used to assess objectives from core courses taken in the first two years of the college of business curriculum. On the business tests, students who did not complete the tests had points deducted from their class grades, but the points earned did not depend on how well they scored on the test. Additionally, this group of students was required to take nine 30-item tests, spread over four weeks, outside of class time (see DeMars, 2007, for more information on this series of tests). This testing burden likely made the tests even less motivating. Only tests administered during the last week were included because tests administered in the last week tended to invoke far more rapid guessing than tests administered in the first week.

Table 1 shows the mean gender difference in RTE. Negative differences indicate lower RTE for men. The degree of the gender gap varied, but men had somewhat lower average RTE on every test. Although RTE was consistently lower for men, what is not evident from Table 1 is that the gender gap was particularly large at the low end of the RTE distribution—far more men than women exhibited rapid guessing behavior on over half of the items. For illustration, the RTE distribution is plotted in Figure 1 for Business Test Q, the test with the greatest gender difference in RTE. Although a minority of men were at the extreme low end, there were far more men than women in this extreme group. The main graph does not include examinees with RTE = 1, because the percentages in this group were much higher than the percentages with any other value of RTE. Instead, these values are shown on a bar chart inlay; although the majority of both male and female examinees had RTE = 1, more women than men were in this extremely high group. The same pattern persisted in Business Test R, as shown in Figure 2, even though the mean gender difference in RTE was smaller for this test. Overall, the gender difference in RTE was small on all tests, but it was most noticeable at low values of RTE. This matters because it is the students exhibiting extremely low effort who are likely to score much lower than they are capable of scoring.

Table 1
Gender Differences in RTE

	N		
Test	Men	Women	Gender Difference in RTE
Science	260	446	-0.01
Business Test Q	215	178	-0.10
Business Test R	214	178	-0.04
Business Test S	208	207	-0.04
Business Test T	208	205	-0.06

Self-Reported Test-Taking Effort

Although many studies do not separate the results by gender, most studies that provide scores by gender tend to show slightly higher levels of self-reported effort for female examinees.

Not attending a required test administration or rapid guessing during the test captures only the lowest levels of test-taking motivation. Self-report scales, on the other hand, may be able to capture a wider range of motivation. In some form, these scales include questions asking the examinees how hard they tried on the test, often for the purpose of studying relationships between motivation and test performance. Hoyt (2001) found that in a sample of college students taking a low-stakes General Education test, 22% reported giving little or no effort to the mathematics subtest, 8% reported giving little or no effort to the English subtest, and 15% reported little or no effort on the critical thinking subtest. Similarly, Schiel (1996), using a larger sample of over 20,000 college and university students, found the percent reporting little or no effort varied from 4-28%, depending on the subtest.

Although many studies do not separate the results by gender, most studies that provide scores by gender tend to show slightly higher levels of self-reported effort for female examinees. Wise et al. (2009) administered a measure of assessment citizenship to university students participating in mandatory low-stakes assessment. Assessment citizenship was a concept modeled on the idea of academic citizenship; students high on this trait would agree that they

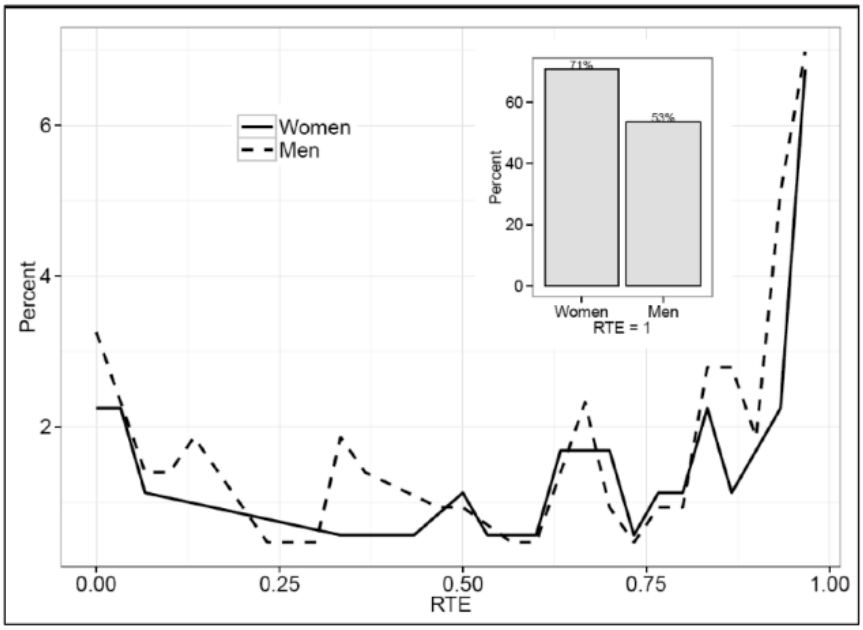


Figure 1. Distribution of RTE on Business Test Q, by gender.

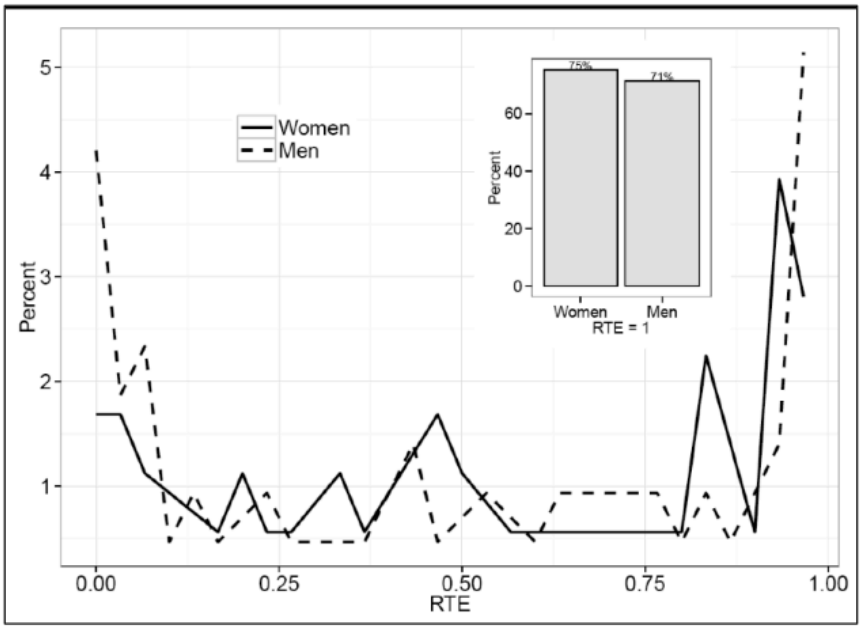


Figure 2. Distribution of RTE on Business Test R, by gender.

had a responsibility, as members of the university community, to comply with requests for participation and exert reasonable effort so that the university could collect valid data. They found a gender difference of 0.22 standard deviation units, with female students reporting more cooperativeness. Cole et al. (2008) administered four General Education tests and asked students to report effort for each test. Gender differences in effort, with negative values indicating greater effort reported by women, ranged from -0.41 standard deviations in English to 0.18 standard deviations in social studies, with intermediate standardized differences of -0.22 in math and -0.02 in science. It seems that, as with studies of RTE, the gender differences in effort vary with the subject area.

Similar results have been reported for secondary students taking low-stakes tests. Eklöf (2007) found that test-taking motivation was about 0.33 standard deviation units higher for girls than for boys among Swedish 14-15 year-olds taking the TIMSS (Trends in International

Although most men, like most women, report reasonable effort, the disproportionate gender ratio in the low range could bias estimates of gender differences in learning.

Math and Science Study) test. O'Neil, Abedi, Miyoshi, and Mastergeorge (2005) studied self-reported effort among 12th graders on released TIMSS items at low-achieving schools. Among students tested under the typical low-stakes instructions, self-reported effort was 0.21 standard deviation units lower for male students. Across many countries, 15-year-olds taking PISA (Programme for International Student Assessment) self-reported their test-taking effort as well as how hard they would have tried if the test counted toward their class grades. Butler and Adams (2007) used the difference between these values as a measure of relative effort. Girls reported slightly higher relative effort than boys. Karmos and Karmos (1984) administered a survey asking middle school students about their attitudes on standardized tests, specifically referring to a test they had recently taken. Three of the items related to effort on the test. Girls reported higher effort, with effect sizes ranging from 0.43 to 0.50 standard deviation units. Brown and Walberg (1993) found no gender differences in self-reported effort on a standardized achievement test, but they studied younger students (grades 3-8).

Empirical Study

To further examine the relationship between gender and self-reported test-taking effort, data were collected from 3,903 women and 2,345 men participating in a university assessment day in spring 2011 and 2012. To motivate students, the university's use of the results was emphasized, but the scores had no impact on student grades or other individual consequences. After completion of the 2.5 hour testing session, students reported their effort using the Student Opinion Scale (SOS; Sundre, 1997; Sundre & Moore, 2002). The scale contains five items pertaining to the student's effort during the assessments, each rated on a 5-point scale from *Strongly Disagree* to *Strongly Agree*. In previous literature, responses to this scale have shown that the item parameters are invariant across gender (Thelk, Sundre, Horst, & Finney, 2009), so comparisons of male and female examinees are reasonable.

In our data, there was very little difference in mean scores on the SOS; the mean for men was 3.62 (*SD* = 0.86) and the mean for women was 3.70 (*SD* = 0.75; Cohen's *d* = -0.10). However, male examinees' effort was more variable (variance ratio = 1.34). Figure 3 shows the distribution of effort. More men reported levels at or below the scale midpoint. More women reported levels between 3.4 and 4.8. Students at the very low end of the effort range may be the ones who could sabotage the test results. Hoyt (2001) and Schiel (1996) each found that the score gap on several tests was smallest between moderate effort and best effort; scores increased most between no effort and little effort, and again between little effort and moderate effort. In Figure 3, clearly there are more men in the problematic range. Although most men, like most women, report reasonable effort, the disproportionate gender ratio in the low range could bias estimates of gender differences in learning.

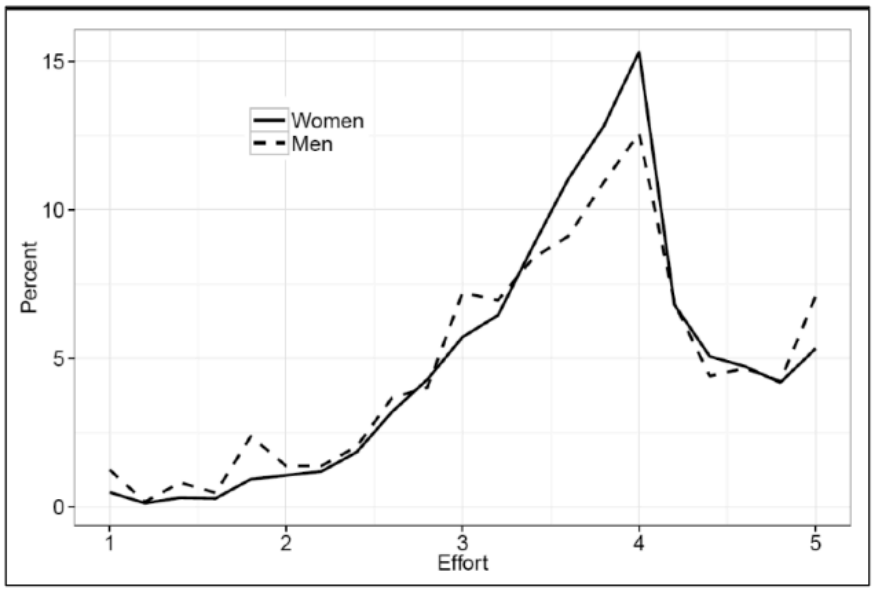


Figure 3. Distribution of self-reported effort, by gender.

As in the literature cited above, in our data self-reported effort was moderately correlated with test performance (correlations ranged from r = .22 to r = .34), but not with SAT scores (r = .05 with SAT verbal and r = .06 with SAT math). The lack of correlation between effort and SAT scores suggests that low test-taking effort yielded low test scores and not the other way around. Thus, the test scores of the subgroup of students who reported very low effort may not be representative of their knowledge. There appeared to be more male than female examinees in the very low end of the effort distribution, which may distort gender differences in test scores.

Possible Explanations of Gender Differences in Test-Taking Motivation

Two questions of interest to both researchers and practitioners might be *why* some examinees are more willing to engage in effort on low-stakes tests and why this tendency relates to gender. Some would attribute differences in examinee motivation to individual differences or personality traits. Specifically, one might expect students who are more agreeable or conscientious to also be more compliant with requests to cooperate in test-taking. Indeed, previous research has found small and not always consistent gender differences in conscientiousness, with women typically reporting being more conscientious (Feingold, 1994) and dutiful (Costa, Terracciano, & McCrae, 2001) than men. Gender differences in agreeableness have been more prominent and consistent, with women scoring higher on agreeableness than men (Costa et al., 2001; Feingold, 1994). Further, Marrs and Sigler (2012) compared study strategies for male and female college students in efforts to explain the lower academic performance of male students. They found that female students tended to employ a "deep approach" to learning, which involved engaging in the material at a deeper level, whereas male students tended to utilize a "surface approach," which involved tasks requiring minimal effort (e.g., memorization). Marrs and Sigler also found that female students were much more academically motivated than male students (d = .44). This is not surprising, given several studies have shown that work-avoidance is negatively related to motivation and achievement (Meece, Blumenfeld, & Hoyle, 1988; Meece & Jones, 1996). Given these findings, it seems reasonable to believe that male students are also more work-avoidant, in addition to being less conscientious and less agreeable than female students. These gender differences in personality may well be the key to explaining, at least in part, the gender differences in test-taking motivation.

Given conscientiousness and agreeableness are somewhat related to effort and women score higher on these attributes, it could be that the gender difference in test-taking motivation is due in part to gender differences in personality.

Empirical Study

As part of campus-wide assessment for accountability purposes in spring 2011 and 2012, students completed a battery of tests which included measures of conscientiousness, agreeableness, and work-avoidance in addition to the Student Opinion Scale. Both conscientiousness and agreeableness are subscales of the Big Five Inventory (John & Srivastava, 1999). Work-avoidance was measured via a subscale of the Achievement Goal Questionnaire (Finney, Pieper, & Barron, 2004).

Table 2
Gender Differences in Personality Traits

	Women			Men			
Trait	Mean	*SD*	*N*	Mean	*SD*	*N*	Cohen's *d*
Conscientiousness	34.03	5.47	1866	32.08	5.53	1114	.36
Agreeableness	36.21	5.35	1861	33.88	5.51	1118	.43
Work-Avoidance	11.13	4.77	3892	12.72	5.41	2333	-.32

As expected, both conscientiousness and agreeableness had about the same small positive relationship with test-taking effort (r = .22 and r = .19, respectively). Although small, both of these correlations are in the expected direction. Thus, they further support the meta-analytic findings in the literature (Costa et al., 2001; Feingold, 1994). In addition, the average conscientiousness and agreeableness scores for men and women are quite different (Table 2), with women scoring higher on both of these measures. The complete distributions of these traits are graphed in Figures 4 and 5. Unlike RTE and self-reported effort, where the gender

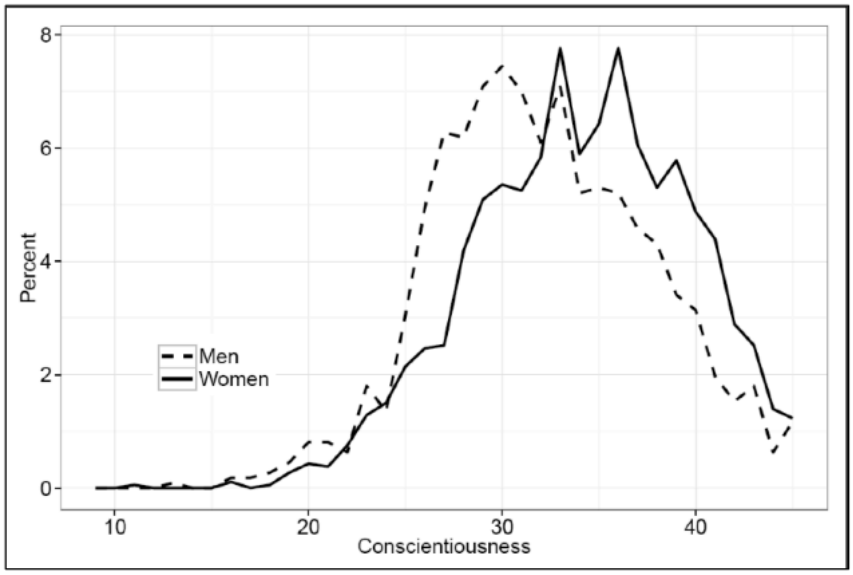


Figure 4. Distribution of Conscientiousness by gender.

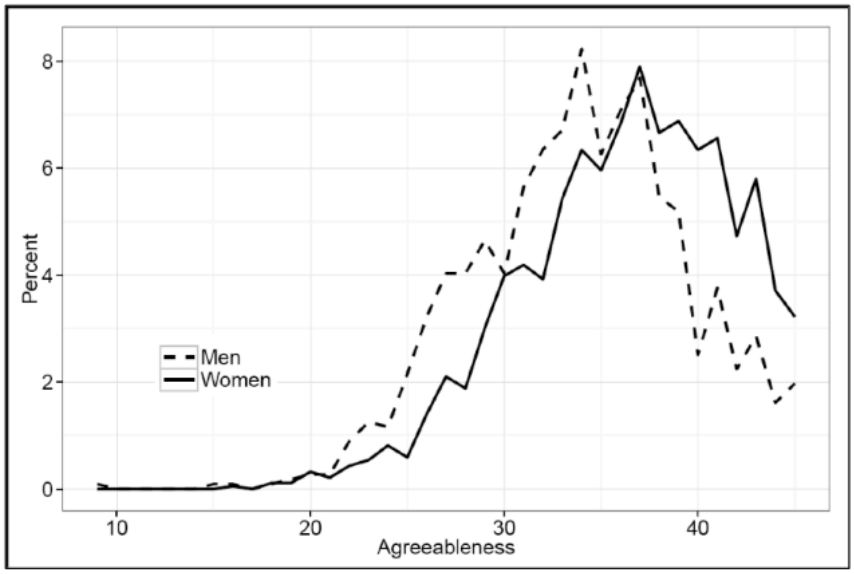


Figure 5. Distribution of Agreeablesness by gender.

differences were limited to extreme scores, these traits show fairly sizeable differences in the means. Given conscientiousness and agreeableness are somewhat related to effort and women score higher on these attributes, it could be that the gender difference in test-taking motivation is due in part to gender differences in personality. This logic is further supported by the relationship between work-avoidance and effort.

As expected, work-avoidance was negatively related to test-taking effort (r = -.23), indicating that the higher one's work-avoidance, the less effort one would likely expend on a battery of low-stakes tests. Moreover, women scored lower on this measure than men, which was not surprising based on previous research (Meece & Jones, 1996; Steinmayr & Spinath, 2008). Figure 6 shows the complete distribution of work-avoidance for men and women. Again, the negative relationship between work-avoidance and test-taking effort coupled with a noticeable gender difference in work-avoidance scores further supports the logic that personality traits may be a promising source in attempts to explain the gender gap in test-taking motivation.

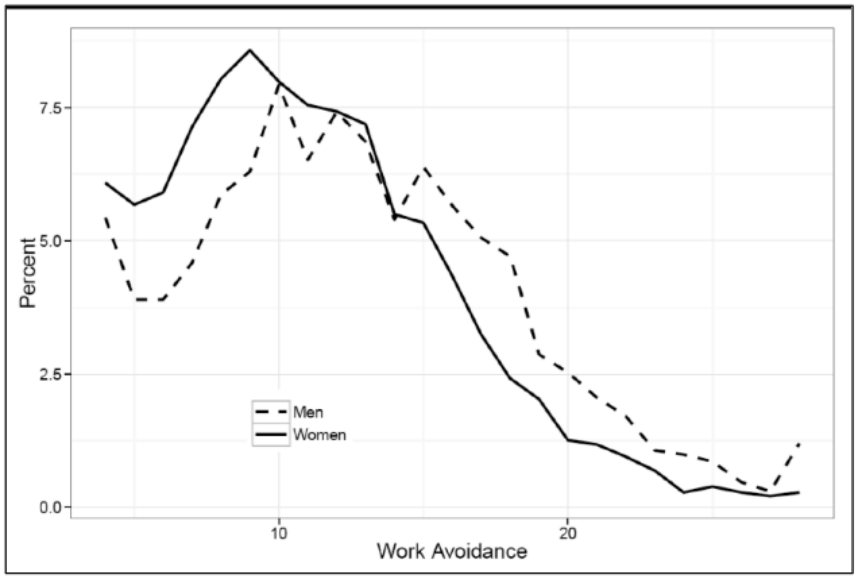


Figure 6. Distribution of Work-Avoidance by gender.

Discussion

Test-taking motivation has been the focus of considerable research in higher education assessment efforts. Previous research has linked high test-taking effort to better performance on specific tests, but not to external measures of proficiency. Thus, test-taking motivation merits close examination, in order to ensure the inferences based on low-stakes assessments are valid. In the current paper, we focused on the role of gender in test-taking motivation—an area that has received indirect attention in research but is equally important in making accurate inferences based on test scores. The purpose of the paper was to draw upon multiple sources of evidence in the existing literature reporting small but consistent gender differences in test-taking motivation and to compare these findings against our own data to investigate the phenomenon more fully. Specifically, we explored gender differences across three different indicators of test-taking motivation documented in the literature: test session attendance, rapid guessing, and self-reported test-taking effort. Where possible, we also included results from our own data, which further supported the trend of lower test-taking motivation among men than women. Finally, based on previous findings, we explored the gender gap in test-taking effort in the context of several personality traits, which we considered a possible pathway to understanding why women tend to expend more effort on low-stakes assessments.

Studies reporting examinee motivation by gender have consistently found higher self-reported effort for females than males.

We first reviewed research on the most basic level of test-taking motivation under low-stakes settings—showing up at an assigned testing session. Absence at an assigned test administration essentially indicates extremely low levels of motivation. Although only one known study has provided this type of evidence of test-taking motivation, the study revealed two compelling findings: (a) male avoiders disproportionately outnumbered their female counterparts (i.e., many more males than females failed to attend their assigned testing session); and (b) failure to attend the assigned testing session was largely related to low motivation. Thus, at the minimum level of test-taking motivation needed to show up to a testing session, males appeared to be less motivated than females. This result could be due to gender differences in personality, which we discuss later. Alternatively, it could be due to other, unmeasured variables.

Second, we examined rapid guessing via RTE, an unobtrusive indicator of test-taking motivation based on response times collected when computerized tests are administered. Specifically, an RTE score indicates the proportion of test items on which an examinee spent a minimally-adequate amount of time to read and consider the response options based on a preset time threshold for each item. Used as a proxy for motivation, RTE scores are especially useful in flagging rapid responses. Given effort is not related to proficiency in general, filtering out data from extremely unmotivated examinees (i.e., rapid responders) can reduce the construct-irrelevant variance in test scores, and thus boost the validity of inferences one

wishes to draw from responses that are now at least minimally effortful. With respect to gender, both prior research and our analyses showed that men tended to engage in rapid guessing more frequently than women. Moreover, in our data samples the gender gap was especially evident in the lower end of the distribution. These slight but fairly consistent findings across samples further support the idea that gender does indeed have a role in test-taking motivation in low-stakes conditions, with women being more motivated than men. As such, this difference should be taken into account when comparing test scores between men and women, provided item response times are available.

Next, we examined what is by far the most widely used indicator of test-taking motivation: self-report measures. Unlike the other two methods, self-report measures typically capture a wider range of examinee motivation, and thus the relationship between scores on such measures and test performance has been widely studied. Studies reporting examinee motivation by gender have consistently found higher self-reported effort for females than males. The data we analyzed also supported this trend. Specifically, we found a very small mean difference in self-reported effort (females scoring higher); however, upon examination of the distribution of effort scores, we discovered a much larger gender gap at the low end of the distribution across multiple tests, indicating men tended to report lower effort than women below the scale midpoint. Again, this gender difference in examinee motivation may appear trivial at the mean level, but it could severely bias the examination of gender differences in test scores; thus, it should be considered when making such comparisons.

Is possible to observe sizeable gender differences in performance on low-stakes assessments partly or fully due to gender differences in test-taking motivation.

Finally, the gender gap in test-taking motivation was examined in the context of personality differences in efforts to provide one plausible explanation of why such a gap in motivation exists. Both prior research and our empirical results indicated that women score higher on conscientiousness and agreeableness and lower on work-avoidance than men do. Furthermore, our analyses showed a positive relationship of effort with conscientiousness and agreeableness, and a negative relationship between effort and work-avoidance. All of these relationships were of modest magnitude but in the expected direction, based on theory and previous findings in the literature. As such, we believe these personality traits provide at least a partial explanation of why women tend to expend more test-taking effort on low-stakes assessments.

Implications for Practice

The array of findings based on prior research and new empirical data presented here clearly indicate a small but consistent gender difference in test-taking motivation under low-stakes conditions. Across a variety of measures of examinee motivation women appear to expend higher levels of effort than men. Although the size of this gender gap appears to vary across age groups and subject areas, it certainly has an impact on test scores. As demonstrated in one study (Wise & DeMars, 2010), gender differences in motivation could almost completely account for gender differences in test scores. Thus, under low-stakes testing conditions, it is of utmost importance to examine not only motivation but also the effect of gender, especially when there is interest in comparing test scores by gender. Assessment practitioners could control for effort by filtering noneffortful responses through RTE screening, when response time data are available, or by collecting other measures of test-taking motivation (e.g., self-report measures), which could then be used as covariates in the analyses.

In addition to applying statistical methods to control for effort in low-stakes conditions, researchers have proposed several approaches to enhance test-taking motivation directly. Such methods include increasing the stakes of the test (e.g., requiring a passing score or including the score in a course grade), conveying to students the importance of assessment for curricular improvement, providing valuable feedback to students regarding their performance, offering monetary incentives, or utilizing a computer-based testing environment which prompts students to expend more effort when they engage in rapid guessing behavior (Wise, 2009). For paper-and-pencil test administrations, researchers have discovered that proctors overseeing the test sessions have a significant impact on the engagement and motivation of examinees, and as a result, on their test scores (Lau, Swerdzewski, Jones, Anderson, & Markle, 2009). Any and all of these methods could be applied in practice in higher education assessment under low-stakes conditions to improve the validity of inferences drawn from test scores.

These are but a few examples of how the findings from the literature and the new empirical evidence presented here could benefit practitioners of higher education assessment. Perhaps the most important take-home message for practice is to be aware that it is possible to observe sizeable gender differences in performance on low-stakes assessments partly or fully due to gender differences in test-taking motivation. Fortunately, there are various methods to empirically investigate this possibility and control for a motivation effect moderated by gender. We presented three such methods, as well as recommendations for ways to increase test-taking motivation in efforts to combat threats to validity of score comparisons and overall test score interpretation. We encourage future research to explore the extent to which these and other motivational enhancement efforts developed in recent years are effective in narrowing the gender gap in test-taking motivation under low-stakes conditions and reducing the construct-irrelevant variance introduced by low levels of effort.

Gender is often used as a proxy variable in research for the very reason that men and women do differ on a wide range of variables that may be difficult to obtain compared to simply recording students' gender.

Limitations and Future Directions

While we identified numerous sources of evidence suggesting a consistent pattern of low test-taking motivation among men compared to women, as well as likely explanations for this pattern, our investigation was limited in several ways. First, we were unable to conduct a thorough meta-analysis of the examinee motivation literature as it pertains to gender differences simply because results are rarely broken down by gender in most published research. It is our hope that once higher education and assessment professionals become aware of the small but consistent gender differences in test-taking motivation under low-stakes conditions, more evidence will be cumulated and this phenomenon will be investigated and understood more fully.

Furthermore, we were able to explore only three personality traits that could allude to the gender gap in test-taking effort. Other important variables certainly exist that could account for gender differences. In fact, gender is often used as a proxy variable in research for the very reason that men and women do differ on a wide range of variables that may be difficult to obtain compared to simply recording students' gender (Bashkov & Finney, 2013). However, the three personality variables discussed in this study appeared essential to understanding at least in part why men and women expended different amounts of effort on the assessments. Future research should explore these and other personality traits further, in order to reach a better understanding of the role of gender in test-taking motivation under low-stakes conditions.

References

Bashkov, B. M., & Finney, S. J. (2013). Applying longitudinal mean and covariance structures (LMACS) analysis to assess construct stability over two time points: An example using psychological entitlement. *Measurement and Evaluation in Counseling and Development, 46*, 289-314. doi: 10.1177/0748175613497038

Brown, S. M., & Walberg, H. J. (1993). Motivational effects on test scores of elementary students. *Journal of Educational Research, 86*, 133-136. doi: 10.1080/00220671.1993.9941151

Butler, J., & Adams, R. J. (2007). The impact of differential investment of student effort on the outcomes of international studies. *Journal of Applied Measurement, 8*, 279-304.

Cole, J. S., Bergin, D. A., & Whittaker, T. A. (2008). Predicting student achievement for low stakes tests with effort and task value. *Contemporary Educational Psychology, 33*, 609–624. doi: 10.1016/j.cedpsych.2007.10.002

Cole, J. S., & Osterlind, S. J. (2008). Investigating differences between low- and high-stakes test performance on a general education exam. *Journal of General Education, 57*, 119-130. doi: 10.1353/jge.0.0018

Costa, P. T. Jr., Terracciano, A., & McCrae, R. (2001). Gender differences in personality traits across cultures: Robust and surprising findings. *Journal of Personality and Social Psychology, 81*, 322-331. doi: 10.1037/0022-3514.81.2.322

DeMars, C. E. (2000). Test stakes and item format interactions. *Applied Measurement in Education, 13*, 55-77. doi: 10.1207/s15324818ame1301_3

DeMars, C. E. (2007). Changes in rapid-guessing behavior over a series of assessments. *Educational Assessment, 12*, 23-45. doi: 10.1080/10627190709336946

DeMars, C. E., & Wise, S. L. (2010). Can differential rapid-guessing behavior lead to differential item functioning? *International Journal of Testing, 10*, 207-229. doi: 10.1080/15305058.2010.496347

Eklöf, H. (2007). Test-taking motivation and mathematics performance in TIMSS 2003. *International Journal of Testing, 7*, 311–326. doi: 10.1080/15305050701438074

Feingold, A. (1994). Gender differences in personality: A meta-analysis. *Psychological Bulletin, 116*, 429-456. doi: 10.1037/a0022247

Finney, S. J., Pieper, S. L., & Barron, K. E. (2004). Examining the psychometric properties of the Achievement Goal Questionnaire in a general academic context. *Educational and Psychological Measurement, 64*, 365-382. doi: 10.1177/0013164403258465

Freund, D. S., & Rock, D. A. (1992, April). *A preliminary investigation of pattern-marking in 1990 NAEP data*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA. (ERIC Document Reproduction Service No. ED 347189)

Hoyt, J. E. (2001). Performance funding in higher education: The effects of student motivation on the use of outcomes tests to measure institutional effectiveness. *Research in Higher Education, 42*, 71-85.

John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality theory and research* (Vol. 2, pp. 102–138). New York, NY: Guilford Press.

Karmos, A. H., & Karmos, J. S. (1984). Attitudes toward standardized achievement tests and their relation to achievement test performance. *Measurement and Evaluation in Counseling and Development, 17*, 56-66.

Kong, X. J., Wise, S. L., & Bhola, D. S. (2007). Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing behavior. *Psychological Measurement, 67*, 606-619. doi: 10.1177/0013164406294779

Lau, A. R., Swerdzewski, P. J., Jones, A. T., Anderson, A. D., & Markle, R. E. (2009). Proctors matter: Strategies for increasing examinee effort on general education program assessments. *Journal of General Education, 58*, 196-217. doi: 10.1353/jge.0.0045

Liu, O. L., Bridgeman, B., & Adler, R. M. (2012). Measuring learning outcomes in higher education: Motivation matters. *Educational Researcher, 41*, 352-362.

Marrs, H., & Sigler, E. A. (2012). Male academic performance in college: The possible role of study strategies. *Psychology of Men & Masculinity, 13*, 227-241. doi: 10.1037/a0022247

Meece, J. L., Blumenfeld, P. C., & Hoyle, R. H. (1988). Students' goal orientations and cognitive engagement in classroom activities. *Journal of Educational Psychology, 80*, 514-523. doi: 10.1037/0022-0663.80.4.514

Meece, J. L., & Jones, M. G. (1996). Gender differences in motivation and strategy use in science: Are girls rote learners? *Journal of Research in Science Teaching, 33*, 393-406. doi: 10.1002/(SICI)1098-2736(199604)33:4<393::AID-TEA3>3.0.CO;2-N

O'Neil, H. F., Abedi, J., Miyoshi, J., & Mastergeorge, A. (2005). Monetary incentives for low-stakes tests. *Educational Assessment, 10*, 185-208. doi: 10.1207/s15326977ea1003_3

Rothe, H. F. (1947). Distribution of test scores in industrial employees and applicants. *Journal of Applied Psychology, 31*, 480–483.

Schiel, J. (1996). *Student effort and performance on a measure of postsecondary educational development* (ACT Rep. No. 96-9). Iowa City, IA: American College Testing Program.

Schnipke, D. L. (1995, April). *Assessing speededness in computer-based tests using item response times*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco. (ERIC Document Reproduction Service No. ED383742)

Steinmayr, R., & Spinath, B. (2008). Sex differences in school achievement: What are the roles of personality and achievement motivation? *European Journal of Personality, 22*, 185-209. doi: 10.1002/per.676

Sundre, D. L. (1997, April). *Differential examinee motivation and validity: A dangerous combination*. Paper presented at the annual meeting of the American Educational Research Association. Chicago, IL.

Sundre, D.L. (1999, April). *Does examinee motivation moderate the relationship between test consequences and test performance?* Paper presented at the annual meeting of the American Educational Research Association, Montreal, Quebec, Canada. (ERIC Document Reproduction Service No. ED432588)

Sundre, D. L., & Kitsantas, A. (2004). An exploration of the psychology of the examinee: Can examinee self-regulation and test-taking motivation predict consequential and non-consequential test performance? *Contemporary Educational Psychology, 29*, 6-26. doi: 10.1016/S0361-476X(02)00063-2

Sundre, D. L., & Moore, D. L. (2002). The Student Opinion Scale: A measure of examinee motivation. *Assessment Update, 14*, 8-9.

Swerdzewski, P. J., Harmes, J. C., & Finney, S. J. (2009). Skipping the test: Using empirical evidence to inform policy related to students who avoid taking low-stakes assessments in college. *Journal of General Education, 58*, 167-195.

Swerdzewski, P. J., Harmes, J. C., & Finney, S. J. (2011). Two approaches for identifying low-motivated students in a low-stakes assessment context. *Applied Measurement in Education, 24*, 162-188. doi: 10.1080/08957347.2011.555217

Terry, N., Mills, L., & Sollosy, M. (2008). Student grade motivation as a determinant of performance on the Business Major Field ETS Exam. *Journal of College Teaching & Learning, 5*, 27-32.

Thelk, A. D., Sundre, D. L., Horst, S. J., & Finney, S. J. (2009). Motivation matters: Using the Student Opinion Scale to make valid inferences about student performance. *Journal of General Education, 58*, 129-151. doi: 10.1353/jge.0.0047

Wise, S. L. (2009). Strategies for managing the problem of unmotivated examinees in low-stakes testing programs. The *Journal of General Education, 58*, 152-166. doi: 10.1353/jge.0.0042

Wise, S. L., & Cotton, M. R. (2009). Test-taking effort and score validity: The influence of student conceptions of assessment. In D. M. McInerney, G. T. L. Brown, G. Arief, & D. Liem (Eds.), *Student perspectives on assessment: What students can tell us about assessment for learning* (pp. 187-205). Charlotte, NC: Information Age Publishing.

Wise, S. L., & DeMars, C. E. (2010). Examinee noneffort and the validity of program assessment results. *Educational Assessment, 15*, 27-41. doi: 10.1080/10627191003673216

Wise, S. L., Kingsbury, G. G., Thomason, J., & Kong, X. (2004, April). *An investigation of motivation filtering in a statewide achievement testing program*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*(2), 163-183. doi: 10.1207/s15324818ame1802_2

Wise, S. L., Pastor, D. A., & Kong, X. J. (2009). Correlates of rapid-guessing behavior in low-stakes testing: Implications for test development and measurement practice. *Applied Measurement in Education, 22*, 185-205. doi: 10.1080/08957340902754650

Wolf, L. F., & Smith, J. K. (1995). The consequence of consequence: Motivation, anxiety, and test performance. *Applied Measurement in Education, 8*, 227-242. doi: 10.1207/s15324818ame0803_3

Wolf, L. F., Smith, J. K., & DiPaulo, T. (1996, April). *The effects of test specific motivation and anxiety on test performance*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Book Review

Successful Assessment for Student Affairs: A How-to Guide
Kimberly Yousey-Elsener. Little Falls, NJ: PaperClip Communications, 2013. 147 pp. ISBN: ASSESS. Binder + CD, $358.00.

REVIEWED BY:
Nathan Lindsay, Ph.D., Dan Stroud, M.A., & Ameshia Tubbs, M.S.E.
University of Missouri - Kansas City

Successful Assessment for Student Affairs: A How-to Guide is a comprehensive toolkit for student affairs professionals that provides a wealth of guidance, resources, and learning exercises. The book was written in 2013 by Dr. Kimberly Yousey-Elsener, a well-known expert in the field who is the Coordinator of Assessment and Evaluation for the Division of Student Affairs at the University of Buffalo (NY), as well as the past-chair of ACPA's Commission for Assessment and Evaluation. The book is published by Paperclip Communications and is fairly concise (147 pages), which makes for a quick and enjoyable read. The book's headings and sub-headings clearly outline the purpose, content, and connections between the various topics, which are sequenced in a logical and helpful manner.

This review provides an overview of the book, and then highlights the book's strengths, areas for improvement, and broader implications. This analysis of the book's pros and cons, along with its utility, situate this resource in the broader literature on student affairs assessment. The workbook is primarily intended for those who are new to student affairs assessment, and the text makes assessment in student affairs more approachable and rewarding. As outlined below, the book could easily serve as the foundational curriculum for a one or two-day assessment workshop in a division of student affairs.

Assessment emphasizes good practice (instead of theory), and usually focuses on one institution (instead of broader implications for multiple contexts).

Book Overview

Successful Assessment for Student Affairs: A How-to Guide is organized as a hands-on work binder with abundant activities and ideas for engagement. The book also comes with a CD that contains all of the information in an electronic format. The author's goal is to take the reader through an assessment cycle that includes the following six components: (a) define purpose; (b) identify outcomes; (c) connect with programs or services; (d) gather data; (e) review, analyze, and interpret results; and (f) share and use information. A quote by Dan Bureau in the introduction sets the context for this work: "Assessment is a mindset, not just an activity. Student affairs professionals committed to their roles as adviser, helper, counselor, responder and advocate should also be dedicated to using assessment as a framework for practice…" (p. 1).

In the second chapter, the inclusion of Steven Covey's second habit of "Begin with the End in Mind" brings immediate purpose to the book and its lessons (Covey, 1989). As with anything written within the self-help industry, especially something that seeks to improve higher education, there is a need for a dynamic mission and clear goals to ensure the necessary follow through from its readers. This section offers just such a sufficient mix by including helpful tips from Linda Suskie (2006), as well as information to help the reader answer questions related to the "why" and "where" of assessment.

What is most impressive about this informal style of instruction is that it offers an engaging roadmap that suggests that assessment is an intriguing mix of work and play, rather than a chore.

The next sections include strategies for identifying outcomes, gathering data, and planning assessment methods. The "Gathering Data" section opens with a clear definition of assessment that distinguishes methods used in assessment from those used in research. Citing Upcraft and Schuh (1996), Yousey-Elsener explains that assessment emphasizes good practice (instead of theory), and usually focuses on one institution (instead of broader implications for multiple contexts). There is a detailed checklist, guiding questions, and brainstorming activities to help readers ensure that the assessment method they select is a reflection of the learning they are seeking to assess.

Yousey-Elsener warns against the impulse to "just do a survey," and provides a detailed analysis of the strengths and challenges of ten different assessment methods: using existing data, surveys, rubrics, focus groups or interviews, portfolios, observation, document analysis, classroom assessment techniques, visual methods, and case studies. The "Things to Consider" section poses questions to prevent potential obstacles that could be easily overlooked in planning and developing an assessment mindset, and the "Learn More" section points to additional resources about each of the assessment methods referenced. Tips for ethical assessment practices, such as "protect anonymity and confidentiality of respondents" (p. 54) are provided alongside more detailed information about institutional review boards and ethics review boards. Yousey-Elsener makes effective use of case studies as a context for ethical considerations such as the integrity of results, the impact of results and the duty to participants. She shows that assessment can be fun by infusing humor in the character names of the case study such as Walter Worry, Director of Residence Life, and Dr. Amanda Assess, Vice President. Activities such as Human BINGO and Survey Design 101 allow the reader to have multiple opportunities to move from theory to practice.

After the methods sections, the book then shifts to analyzing and interpreting qualitative and quantitative data. These sections give the reader a thorough grounding in analytical terminology, and provide a good overview for anyone new to these topics or a great refresher for anyone who has taken qualitative or quantitative research courses. The book also gives concrete examples of data to be coded or mathematically computed. The concluding chapters focus on sharing and using assessment results, as well as building a culture of assessment. There are also best practices listed from many universities, as well as suggestions for how to share data in a creative manner. The glossary at the end of the book outlines clear definitions of assessment lingo such as "internal variables," "practical significance," and "non-responsive methods." Each of these features enhances the book and should be very useful to readers.

Strengths

By far the strongest element of the book is its exhaustive, step-by-step outline of how to design, implement, and analyze effective assessments, which is done in a supportive tone and style. What is most impressive about this informal style of instruction is that it offers an engaging roadmap that suggests that assessment is an intriguing mix of work and play, rather than a chore. This approach makes assessment far less grueling for student affairs professionals who are new to this area.

The layout of the book is particularly useful—each section provides a theoretical base and then practical applications to facilitate the assessment process. The author recommends resources such as Learning Reconsidered (2004), CAS standards (2012), and the AAC&U LEAP (2012) initiative to help readers strengthen their theoretical base. These resources can help student affairs professionals to define their purpose and ground their work in the "bigger picture," in conjunction with an institutional guiding framework and foundational documents such as a strategic plan. Yousey-Elsener also includes words of wisdom from several assessment experts across the country, and these quote boxes help to illuminate the rationale behind many of the assessment principles.

These resources and tools for engagement allow for lasting and quality driven conversations about best practices in student affairs assessment.

As noted earlier, the focus on application throughout the binder is very strong. Every section offers suggestions for "Campus Connections," which are ways to collaborate with others at one's institution (e.g., with faculty, students, the Institutional Review Board), as well as resources (e.g., statistics software and Institutional Research data) that should be explored. The worksheets containing questions for application are excellent. Yet another strong feature in the book is its use of tables, figures, and tip sheet worksheets throughout. These resources and tools for engagement allow for lasting and quality driven conversations about best practices in student affairs assessment.

Areas for Improvement

It is difficult to find any significant weaknesses with this binder, but this resource could be improved with a few additions. In some situations, the process for implementing the resources could be clarified. For example, the section on designing assessment using rubrics is presented in a very tangible checklist that uses a combination of statements and questions to guide thinking; this section could be improved with the addition of a completed holistic or analytic rubric to provide an example of how verbs change the descriptors.

Higher education administrators need to be aware of the politics, pitfalls, and risks as well as the rewards that can be found in student affairs assessment.

A stronger overview of validity and reliability could have been included, and some examples of whether an instrument is measuring what it is intended to measure would be helpful. For instance, it can be difficult to assess engagement and self-esteem among students. Schuh (2009) offers some concrete ideas about how to attain and develop strong evidence in order to enhance face validity. Similarly, the consideration of responsible sample size and the reliability or error it holds in a given population would be good to strengthen future works.

It is clear that this book/manual is intended as an introductory survey of the many facets of student affairs assessment available to higher education administrators. However, there should also be room for some cautionary advice as well. Higher education administrators need to be aware of the politics, pitfalls, and risks as well as the rewards that can be found in student affairs assessment. It may be that such information might muddy the waters for young assessors, but it is probably better to start with responsible and clean assessments.

As a last suggestion, in the book's final section, it might be helpful to focus more on developing a "culture of learning" or a "culture of student development," rather than a "culture of assessment." This section speaks to the efforts to engage student affairs professionals more regularly in assessment, and this could probably be done more effectively and with more buy-in if the focus is on learning.

Implications

Successful Assessment for Student Affairs: A How-to Guide is a nice complement to other books on student affairs assessment (see Bresciani, Zelna, & Anderson, 2004; Bresciani, Moore Gardner, & Hickmott, 2009; Schuh 2009; Schuh & Upcraft, 2001; Upcraft & Schuh, 1996). As a hands-on binder, Yousey-Elsener's book is most similar to Schuh and Upcraft's (2001) *Assessment Practice in Student Affairs: An Applications Manual.* In the 12 years since that

publication, the field of student affairs assessment has made some significant strides, and this book highlights many recent innovative practices from universities across the country.

It might be helpful to focus more on developing a "culture of learning" or a "culture of student development," rather than a "culture of assessment."

In recent years, it has been fairly common for a chief student affairs officer to want to lead his or her division in a day or two of "assessment boot camp" in order to get everyone up to speed. This book could provide a wonderful curriculum for such a seminar, serving as a resource to lead staff from assessment topics A to Z.

In summary, this was a very thorough and pleasurable read. We all enjoyed reading the book, which served as a helpful review for us and taught us new principles covering a broad range of best practices in student affairs assessment. There is little doubt that such a publication should take a prominent place alongside other assessment training literature. Now that this resource exists, it would behoove leaders in the assessment community to promote its value to the coming generation of student affairs professionals who need stronger competencies in assessment.

References

Association of American Colleges and Universities. (2012). *Liberal education and America's promise* (LEAP). Retrieved from https://www.aacu.org/leap/index.cfm

Bresciani, M. J., Moore Gardner, M., & Hickmott, J. (2009). *Demonstrating student success: A practical guide to outcomes-based assessment of learning and development in student affairs.* Sterling, VA: Stylus.

Bresciani, M. J., Zelna, C. L., & Anderson, J. A. (2004). *Assessing student learning and development.* National Association of Student Personnel Administrators.

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

Covey, S. R. (1989). *The seven habits of highly effective people: Restoring the character ethic.* New York, NY: Free Press.

Keeling, R. P. (Ed.). (2004). *Learning reconsidered: A campus-wide focus on the student experience.* Washington, DC: American College Personnel Association & National Association of Student Personnel Administrators. Retrieved October 27, 2013 from http://www.myacpa.org/pub/documents/learningreconsidered.pdf

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

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

Suskie, L. (2006). Understanding and using assessment results. Retrieved on August 29, 2013 from http://www.baruch.cuny.edu/assessment/documents/2007-03StatenIslandusingresults.pdf

Upcraft, M. L., & Schuh, J. H. (1996). *Assessment in student affairs*. San Francisco, CA: Jossey-Bass.

Book Review

Learning Is Not a Sprint: Assessing and Documenting Student Leader Learning in Cocurricular Involvement. Kathy M. Collins and Darby M. Roberts (Eds.). Washington, DC: National Association of Student Personnel Administrators, 2012. 216 pp. ISBN-13: 978-0931654992. Hardcover, $34.95.

REVIEWED BY:
Lisa Endersby, M.Ed.
Seneca College of Applied Arts and Technology

The fundamental need for recognition and validity is a common feature of many student development theories. Our work in student affairs strives to provide students with meaning and purpose, while also equipping them with the tools to continue their own professional and personal journey long after their graduation ceremony. Ongoing and heightened interest in assessment can be seen as the professionals' quest for meaning making, where we heed the call for accountability while also, by necessity, raising our collective voice for justification and recognition. The forward of *Learning Is Not a Sprint* pointedly asks hard questions about the profession and our apparent role confusion, especially, "Is it because we have too long focused on what is important to us as individuals and as a profession rather than what is important and prioritized by our institutions?" (Bresciani, 2012, p. 10) Put another way, does our interest in assessment stem from what is important to us as professionals (how we are evaluated as administrators) or what is important to us as a profession (how our students can demonstrate meaningful learning and development)? *Learning Is Not a Sprint*, as an edited volume, shines the spotlight on both sides of this chicken and egg debate.

Does our interest in assessment stem from what is important to us as professionals (how we are evaluated as administrators) or what is important to us as a profession (how our students can demonstrate meaningful learning and development)?

The collected writings in *Learning Is Not a Sprint* bring together multiple perspectives on the growing popularity of assessment in higher education and balances these opinions with strategies and best practices for incorporating strong assessment practices into nearly any institutional context. Equal parts evaluation manifesto and practical textbook, the book offers real-world examples of assessment challenges and success stories, while offering lessons learned from administrators at a variety of professional levels. *Learning Is Not a Sprint* covers all of a professional's assessment bases, ranging from the why of assessment as a guiding practical philosophy to the what and how of using assessment tools to further the student affairs mission of supporting the students' academic and personal journeys.

While the book is a much newer and more modern compilation of assessment expertise, the focus on assessment itself is a much older phenomenon. Tracking, recording and evaluating student progress is not unique to post-secondary education, and the desire to compare our achievements to those of other institutions is common. This mostly unidirectional evaluation process has begun to shift, however, in higher education with students who bring a more critical eye and strong personal investment to their learning. As the book describes, this new model of learning and teaching "shifts [the] emphasis of the learning environment from being authority driven to being learner driven" (Hynes, 2012, p. 21). This learner driven environment demands we not only be accountable for learning but also to the learner themselves. Here, the emphasis is on the learner and the fundamental principles behind student affairs as a profession, including purposeful planning, a seamless learning environment and "prepare[ing] students to influence the world, to have the ability to continue to learn and develop, and to make lasting contributions to their respective fields or disciplines" (Hynes, 2012, p. 38).

This learner driven environment demands we not only be accountable for learning but also to the learner themselves.

In order to influence, learn, develop and contribute, students require well-developed skills in communication, critical thinking, and leadership skills. It is here that the emphasis on assessment shifts from the professional to the profession, much in the same way that the book begins to shift toward a critical examination of student affairs and the state of higher education. Assessment, as many of the authors who contributed to the book will attest, requires a deep understanding of these and other skills beyond a justification of their teaching. The how of student learning is seen as a process embedded in a larger mess of political, social, historical and contextual influences. These same issues that impact student learning impact the profession, to the point where it is difficult to separate the two. While some in the field may argue that students' learning is in fact synonymous with student affairs as a profession, it is our understanding of learning that can shift our focus from the profession to the professional, and from learning to the learner.

Learning Is Not a Sprint, by its title alone, focuses on the act of learning and the outside demands on this process. There are many steps and stages outlined throughout the book, providing structure to what is often a haphazard process. Here, the authors truly emphasize learning and the assessment of student development as a process, attempting to remove some of the magical thinking associated with assessment. Rather than focusing solely on the final product, in this case a neat and tidy learning outcome that has been met or a completed evaluation, the book shines a spotlight on the messy and unpredictable learning process that the profession has often tried to keep hidden. Although the steps and stages

are outlined in neatly numbered lists, each landmark signals another step on what is most often a messy and complicated process. The authors are unafraid to examine assessment with a critical eye, not for another attempt at evaluation or critique, but rather to inspire meaningful discussion about the profession itself. Too often, assessment and evaluation are end products, when they are truly impactful throughout the lifecycle of a student's learning experience. Here the emphasis on the profession is crucial, creating a shared sense of urgency and responsibility for student success.

Rather than focusing solely on the final product, in this case a neat and tidy learning outcome that has been met or a completed evaluation, the book shines a spotlight on the messy and unpredictable learning process that the profession has often tried to keep hidden.

Any emphasis on a profession, however, also demands a critical, and often closer, examination of the professionals working in the field. *Learning Is Not a Sprint* does devote attention to those responsible for creating, facilitating and assessing the exchanges and dialogues critical to positive student development. Here, the authors offer a series of practical tips and timely advice for professionals. The emphasis shifts again from learning to the learner; from the profession to the professional. A commitment to learning is described as a shared responsibility, leaving no one from coordinator, administrator, dean and principal behind. This is particularly true in assessment, as any shared responsibility for the act of assessment itself and the implementation of resulting recommendations must include the students who are impacted by the process.

Far more than other areas in student affairs, assessment acts as an underlying theme across all other projects and programs, not only evaluating student learning but, as the cycle goes, contributing directly to the learning process. As the authors note, both "the documentation and the observation by the student affairs professional … is needed to assist student leaders as they learn, develop, and grow" (Starcke & DeLoach, 2012, p. 92). This discussion is particularly appealing to those currently working in assessment as well as all professionals in student affairs, as it states the importance of assessment in the very cause most professionals chose to take up when they entered the profession. Rightly so, the chapter on "Assessing and Documenting Student Learning" concludes with the statement, "We must always keep in mind that the learning experiences and feedback that occur through interpersonal exchange and authentic mentoring have far more impact than any assessment strategy" (Starcke & DeLoach, 2012, p. 93). This is perhaps the most critical lesson of the book, highlighting the most important role of assessment; continually refining the day-to-day interactions that make up students' in class and co-curricular learning experiences.

The students discussed in *Learning Is Not a Sprint* represent a diverse array of demographic characteristics. The theories used and examples presented are based on a wide range of student experiences. While this diversity is an important attempt at mirroring the student populations many professionals currently work with, the book often conflates the example of "students" with "student leaders." At several points throughout the chapters, students are referred to as "student leaders," described by one of the authors as "students [who will] influence the world … have the ability to continue to learn and develop, and … [will] make lasting contributions to their respective fields or disciplines" (Hynes, 2012, p. 38). Several of the theories presented around student learning and development are also leadership development theories that, while certainly relevant to any and all students, are often used to explain and support the learning of those students who have taken on a defined leadership role at their institution.

While the definition noted above is inclusive of any and all students, the book goes on to describe assessing the learning and development of students participating in defined activities, programs, roles and events. It is possible that this emphasis was intentional, as the authors echo the notion that "students tend to focus on the formal academic curriculum and do not easily identify the opportunities that are available to them for learning outside the classroom or across campus" (Holzweiss, 2012, p. 61). The emphasis on those students who can be more easily evaluated in predetermined roles with more intuitive learning outcomes ignores many hidden or unintentional learning moments that happen outside of the more traditionally defined student leadership programs.

This is perhaps the most critical lesson of the book, highlighting the most important role of assessment; continually refining the day-to-day interactions that make up students' in class and co-curricular learning experiences.

While several authors discuss the need to validate and differentiate the profession from academic affairs, confining a discussion of student learning to defined roles and positions is no better than the in-class learning that we often fight to be seen apart from. In arguing to be distinct from a process that is seemingly overtaking our importance and uniqueness in higher education, the book instead draws the same lines in the sand by creating a set and elevated subsection of students to focus on. While the learning of students who are not involved in these roles outside of the classroom is certainly more difficult to assess, limiting our effectiveness as a profession to these defined, resource intensive opportunities only serves to limit the scope and scale of our potential impact on the student experience.

Learning Is Not a Sprint is poised to become a seminal work in the student affairs literature, highlighting the importance of assessment for the parallel and equally important roles of professionals in advancing student learning while striving to elevate and celebrate the contributions of the field. More than ever, student affairs must not only continue to deliver high quality programs and services but also provide concrete evidence of a very disordered process. For the good of our students, we must adopt the patient, forward thinking attitude espoused by the authors to walk

***Learning Is Not a Sprint* is poised to become a seminal work in the student affairs literature, highlighting the importance of assessment for the parallel and equally important roles of professionals in advancing student learning while striving to elevate and celebrate the contributions of the field.**

the blurry line between fast-paced, constant change and the slower, more subtle, long term transformations that can be overlooked without more deliberate attention to assessment. In treating assessment as part of, rather than apart from, our work in student affairs, the authors present a crucial shift in mindset from assessment "because we have to" and toward assessment "because we must." The combination of practical advice and a professional call to action follows a more unique approach to writing for an audience of administrators, ensuring that a call to action carries manageable, and measurable, strategies while these strategies are presented within a compelling framework denoting the importance of these methods in supporting students. Learning Is Not a Sprint is an essential and highly recommended part of a professional's library. Its call to action must be heard if we are to move forward, collectively, in answering the hard questions facing the profession.

References

Bresciani, D. (2012). The current student affairs and higher education environment. In K. M. Collins & D. M. Roberts (Eds.), *Learning is not a sprint: Assessing and documenting student leader learning in cocurricular involvement* (pp. 1-16). Washington, DC: NASPA-Student Affairs Administrators in Higher Education.

Holzweiss, P. C. (2012). Determining skills and outcomes. In K. M. Collins & D. M. Roberts (Eds.), *Learning is not a sprint: Assessing and documenting student leader learning in cocurricular involvement* (pp. 17-42). Washington, DC: NASPA-Student Affairs Administrators in Higher Education.

Hynes, S. D. (2012). Leadership and student learning. In K. M. Collins & D. M. Roberts (Eds.), *Learning is not a sprint: Assessing and documenting student leader learning in cocurricular involvement* (pp. 17-42). Washington, DC: NASPA- Student Affairs Administrators in Higher Education.

Starcke, M., & DeLoach, A. (2012). Leadership and student learning. In K. M. Collins & D. M. Roberts (Eds.), *Learning is not a sprint: Assessing and documenting student leader learning in cocurricular involvement* (pp. 73-102). Washington, DC: NASPA- Student Affairs Administrators in Higher Education.

RUMINATE: INTEGRATING THE ARTS AND ASSESSMENT

ENGAGEMENT, UNPACKED
Artist:Keith Frew
Painting untitled
collection of John Baldasare

The process of schooling in its barest form cannot be successfully studied by a scientific psychology unless that psychology is social, i.e., unless it recognizes that the processes of acquiring knowledge, of giving attention, of evaluating…must be studied in their relation to selves in a social consciousness. So far as education is concerned, the child does not become social by learning. He must be social to learn.

–George Herbert Mead, 1909

Nothing should be learned which does not in some way contribute to the life of the student – be it through a strengthening of the energy for a certain function which this learning carries, or through the farther-reaching significance which this content wins for the depth, clarity, breadth, and moral constitution of the student.

-Georg Simmel, 1922

GUIDELINES FOR SUBMISSION

All manuscripts submitted to *Research & Practice in Assessment* may be related to various higher education assessment themes, and adopt either an assessment measurement or an assessment policy/foundations framework:

Assessment Measurement:
a) instrument design, b) validity and reliability, c) advanced quantitative design, d) advanced qualitative design

Assessment Policy/Foundations:
a) accreditation, b) best practices, c) social and cultural context, d) historical and philosophical context, e) political and economic context

Article Submissions:

Articles for *Research & Practice in Assessment* should be research-based and include concrete examples of practice and results in higher education assessment. The readers of *Research & Practice in Assessment* are associated with myriad institutional types and have an understanding of basic student learning and assessment practices. Articles for publication will be selected based on their degree of relevance to the journal's mission, compliance with submission criteria, quality of research methods and procedures, and logic of research findings and conclusions. Approximately forty percent of submissions are accepted for publication.

Review Submissions:

Reviews (book, media, or software) are significant scholarly contributions to the education literature that evaluate publications in the field. Persons submitting reviews have the responsibility to summarize authors' works in a just and accurate manner. A quality review includes both description and analysis. The description should include a summary of the main argument or purpose and overview of its content, methodology, and theoretical perspective. The analysis of the book should consider how it contrasts to other works in the field and include a discussion of its strengths, weaknesses and implications. Judgments of the work are permitted, but personal attacks or distortions are not acceptable as the purpose of the review is to foster scholarly dialogue amongst members of the assessment community. The *RPA* Editor reserves the right to edit reviews received for publication and to reject or return for revision those that do not adhere to the submission guidelines.

Special Features:

Each issue of *Research & Practice in Assessment* highlights the work of a noted researcher or assessment professional in a manner that aims to extend the scholarly dialogue amongst members of the assessment community. Special Features are invited by the Board of Editors and often address the latest work of the author.

Notes in Brief:

Notes in Brief offer practitioner related content such as commentaries, reports, or professional topics associated with higher education assessment. Submissions should be of interest to the readership of the journal and are permitted to possess an applied focus. The *RPA* Editor reserves the right to edit manuscripts received for publication and to reject or return for revision those that do not adhere to the submission guidelines.

Ruminate:

Ruminate concludes each issue of *Research & Practice in Assessment* and aims to present matters related to educational assessment through artistic medium such as photography, poetry, art, and historiography, among others. Items are encouraged to display interpretive and symbolic properties. Contributions to Ruminate may be submitted electronically as either a Word document or jpg file.

Manuscript format requirements available at: www.RPAjournal.com

RPA BOOKS AVAILABLE

Bensimon, E.M., & Malcolm, L., (Eds.). (2012). *Confronting equity issues on campus: Implementing the equity scorecard in theory and practice*. Sterling, VA: Stylus Publishing. pp. 289. $35.00 (paperback).

Christensen, C., & Eyring, H. (2011). *The innovative university: Changing the DNA of higher education from the inside out.* San Francisco, CA: Jossey Bass. pp. 512. $32.95 (hardcover).

Côté, J. E., & Allahar, A. L. (2011). *Lowering higher education: The rise of corporate universities and the fall of liberal education*. Toronto, ON: University of Toronto Press Publishing. pp. 256. $24.95 (paper).

Fried, J. (2012). *Transformative learning through engagement: Student affairs practice as experiential pedagogy*. Sterling, VA: Stylus Publishing. pp. 224. $29.95 (paperback).

Gardner, M. M., Kline, K. A., & Bresciani, M. J. (2013). *Assessing student learning in the community and two-year college*. Sterling, VA: Stylus Publishing. pp. 216. $29.95 (paperback).

Gaston, P. L. (2013). *Higher education accreditation: How it's changing, why it must*. Sterling, VA: Stylus Publishing. pp. 240. $35.00 (paperback).

Hendrickson, R. M., Lane, J.E., Harris, J.T., & Dorman, R.H. (2012). *Academic leadership and governance of higher education*. Sterling, VA: Stylus Publishing. pp. 448. $45.00 (cloth).

Hutchings, P., Huber, M. T., & Ciccone, A. (2011). *The scholarship of teaching and learning reconsidered: Institutional integration and impact*. San Francisco, CA: Jossey Bass. pp. 224. $30.00 (paper).

Makela, J.P., & Rooney, G.S. (2012). *Learning outcomes assessment step-by-step: Enhancing evidence-based practice in career services*. Broken Arrow, OK: National Career Development Association. pp. 80. $35.00 (paper).

Martin, R. (2011). *Under new management: Universities, administrative labor, and the professional turn*. Philadelphia, PA: Temple University Press. pp. 253. $69.50.

Noyce, P. E., & Hickey, D. T. (Eds). (2011). *New frontiers in formative assessment*. Cambridge, MA: Harvard Education Press. pp. 260. $29.95 (paper).

Renn, K.A., & Reasond, R.D. (2012). *College students in the United States: Characteristics, experiences, outcomes*. San Francisco, CA: Jossey Bass. pp. 320. $45.00 (hardcover).

Schloss, P. J., & Cragg, K. M. (2012). *Organization and administration in higher education*. New York, NY: Routledge. pp. 350. $160.00 (hardcover).

Secolsky, C., & Denison, B. D. (Eds). (2011). *Handbook on measurement, assessment, and evaluation in higher education*. New York, NY: Routledge. pp. 704. $119.95 (paper).

Smith Morest, V. (2012). *Community college student success: From boardrooms to classrooms*. Lanham, MD: American Council on Education/ Rowman & Littlefield Publishers, Inc. pp. 132. $60.00. (hardcover).

Stone, T., & Coussons-Read, M. (2011). *Leading from the middle: A case-study approach to academic leadership for associate deans*. Lanham, MD: Rowman & Littlefield Publishers, Inc. pp. 208. $40.00 (cloth).

von Davier, M., Gonzalez, E., Kirsch, I., & Yamamoto, K. (Eds.). (2012). *The role of international large-scale assessments: Perspectives from technology, economy, and educational research*. New York, NY: Springer. pp. 160. $129.00 (hardcover).

Wills, K. V. & Rice, R. (Eds.). (2013). *ePortfolio performance support systems: Constructing, presenting, and assessing portfolios*. Anderson, SC: Parlor Press. $30.00 (paperback).