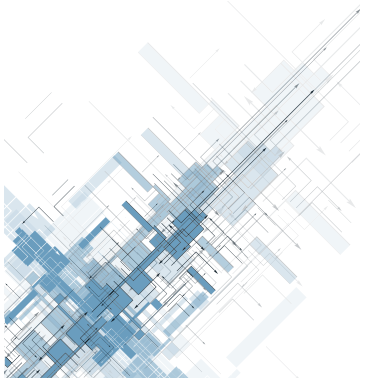
Abstract

Professional standards related to outcomes assessment call for student affairs professionals to use research to inform programming. If professionals are to rely on research to build programs that positively impact student learning outcomes, the research should be credible. We examined the quality of program effectiveness research available for programming decisions. We reviewed five years of quantitative and mixed methods program effectiveness studies published in four student affairs journals. Despite frequent assertions of program effectiveness, the research designs and analyses did not often support such claims due to plausible threats to the validity of those claims. Articles claiming that programming is effective without credible evidence to support such a claim can result in professionals offering ineffective programming and engaging in inefficient assessment efforts. To address the credibility of effectiveness claims, we call for increased training in research methods, careful review of authors' claims by editors, and assistance from assessment practitioners.



AUTHORS

S. Jeanne Horst, Ph.D.
James Madison University

Sara J. Finney, PhD.
James Madison University

Caroline O. Prendergast, MEd
James Madison University

Andrea M. Pope, PhD.
James Madison University

Morgan Crewe, MA
James Madison University

The Credibility of Inferences from Program Effectiveness Studies Published in Student Affairs Journals: Potential Impact on Programming and Assessment

Faculty and student affairs professionals strive to offer programming (e.g., activities, pedagogies, strategies) that results in students achieving intended learning and development outcomes. Professionals are then expected to assess the programming for its level of effectiveness. If effectiveness is not achieved, professionals are expected to use assessment results to inform programming changes that improve learning and development. However, there are few examples of such improvement efforts resulting in greater student learning (Banta & Blaich, 2011; Jankowski, et al., 2018). In turn, assessment practitioners have considered strategies to address this issue and increase learning improvement (e.g., Fulcher & Prendergast, 2019; Smith, et al., 2018).

One strategy is to implement evidence-based programming (Finney & Buchanan, 2021). Building programming based on evidence is also referred to as "evidence-based practice (EBP): instructional approaches shown by high-quality research to result reliably in generally improved student outcomes" (Cook et al., 2011, p. 493). Student affairs' professional competencies and standards call for programming to be intentionally built using current research that indicates what effectively impacts particular outcomes (e.g., ACPA & NASPA, 2015; Finney & Horst, 2019a, 2019b). Additionally, the *Assessment Skills Framework* indicates that the ability to identify literature domains to inform program development is necessary for high-quality assessment practice (Horst & Prendergast,

CORRESPONDENCE

Email

jeanne.horst@gmail.com

2020). When discussing the Grand Challenge in Assessment of "Driving Innovation", Singer-Freeman and Robinson (2020) noted that professionals must "identify evidence-based solutions from the research literature" (p. 5) to improve students' outcomes. In short, consulting existing research increases the probability that programming will impact intended student learning and development outcomes (Pope et al., 2019; Smith & Finney, 2020).

Carpenter (2001) likened evidence-based program development to evidence-based medicine, arguing that in the absence of rigorous evaluations of effectiveness "student affairs may be doomed to repeating past mistakes in the name of tradition and convention" (p. 302). Although often not computed, the cost of implementing ineffective programs can be quite high (Bickman & Reich, 2015). Students engaging in ineffective programs may not achieve desired outcomes, which may prompt additional programming and increased time to degree completion. Thus, professionals should strive to identify and implement effective programs that have credible evidence of impacting desired outcomes. Subsequent outcomes assessment is still necessary and is used in a confirmatory manner to assess if the evidence-based programming is effective in the specific institutional context (Finney et al., 2021). This confirmatory approach is efficient. Less time and resources are needed to improve programming because it is likely to be effective. Thus, fewer iterations of the assessment cycle are needed to inform changes to programming to obtain desired levels of student learning and development.

The Need to Evaluate the Quality of Published Effectiveness Studies

High-quality evidence of program effectiveness requires carefully designed research studies. Put simply, studies that are methodologically suspect do not provide compelling evidence for making programming and assessment decisions.

To implement evidence-based programming, professionals must locate evidence and appraise the evidence for its validity, effect size, and applicability to the population in question. This paper focuses on evidence provided by published program effectiveness studies, which are used in the development of evidence-based programs. An early definition described program effectiveness as "… the extent to which pre-established objectives are attained as a result of activity" (Deniston et al., 1968, p. 324). Analogous to the logic underlying the outcomes assessment process, program effectiveness studies are conducted to evaluate if programming impacted intended outcomes. Thus, in this paper, we refer to "program effectiveness studies" as those that explicitly state student learning or development outcomes for a program and then report findings to evaluate whether students have met those outcomes. The ideal inference from a program effectiveness study is that the program led to or "caused" student learning or development. Not all program effectiveness studies can support this inference; however, this inference is foundational to developing evidence-based programming and merits scrutiny.

High-quality evidence of program effectiveness requires carefully designed research studies. Put simply, studies that are methodologically suspect do not provide compelling evidence for making programming and assessment decisions. "[T]he methodological rigor of a piece of research dictates directly the 'credibility' (Levin, 1994; Murnane & Willett, 2011) of its evidence, or the 'trustworthiness' (Jaeger & Bond, 1996) of the research findings and associated conclusions" (Levin & Kratochwill, 2013, p. 469). Whether the evidence influences program-related decisions "depends in part on the judgements that people make of its credibility, as credibility judgements precede processes of persuasion, influence and use" (Miller, 2015, p. 41). Due to lack of training in appraising the credibility of empirical evidence (Cooper et al., 2016; Muller et al., 2018), professionals may rely on journal editors and reviewers to judge research quality (Miller, 2015). By virtue of publication in peer-reviewed journals, professionals may believe evidence is credible and, in turn, trust inferences and implications provided by the study's authors (Hilligoss & Rieh, 2008).

When consulting with student affairs colleagues who were using published program effectiveness research to inform programming, we observed variability in the credibility of evidence found in the journals they referenced. Moreover, concerns about the quality of research design and credibility of inferences have been voiced by student affairs professionals (e.g., Grace-Odeleye & Santiago, 2019; Valentine et al., 2011). These concerns prompted calls for more rigorous designs that afford trustworthy claims, thereby facilitating successful engagement in programming and assessment efforts.

Program effectiveness studies typically infer that programming *caused* or *did not cause* an outcome. Whenever causal inferences are stated, they should be held to standards and assessed for common threats to the validity of causal inferences (e.g., Shadish et al., 2002). Unjustified statements that programming "led to," "caused," or "influenced" student learning or development outcomes can lead professionals to implement ineffective programs. Moreover, if misleading causal statements are prevalent in the literature, professionals may believe these statements are justified and may offer unsubstantiated interpretations of their findings when engaging in outcomes assessment. Ultimately, when studies with poor methodological quality and incorrect inferences are routinely published, it "not only reduces the faith placed in the findings from studies examining the effectiveness of a specific intervention, but it undermines the faith that policymakers, practitioners, and the public at large place in the educational research enterprise" (Robinson et al., 2018, p. 12). Despite this reality, a formal review of the credibility of inferences about program effectiveness published in student affairs journals has not been conducted. Thus, in the current study, we systematically examined the quality of program effectiveness studies published in four student affairs journals. Using a rigorous approach, we appraised the validity of causal inferences made about program effectiveness (i.e., extent to which programming impacts intentional outcomes) given the studies' designs, data, and analyses.

It is important to note that not all assessment endeavors can be expected to support causal inferences, nor are we calling for the abandonment of outcomes assessment that does not meet the criteria of program effectiveness research.

It is important to note that not all assessment endeavors can be expected to support causal inferences, nor are we calling for the abandonment of outcomes assessment that does not meet the criteria of program effectiveness research. Instead, we are calling for honesty and transparency in the inferences made from program effectiveness studies, as these inferences may influence programming, implementation, and assessment decisions. In fact, Upcraft and Schuh (2002) noted that professionals assessing programming, particularly in published assessment studies, must describe any limitations, stating

> Failure to take this step is not only unethical, it leaves readers to assume that because the investigators did not identify limitations, they must not know them (or worse yet, they made a conscious decision to leave them out), and therefore both the investigators and the study itself lack credibility. (p. 20)

Although they show many commonalities, there are differences between "assessment" and "research", including the purpose, context, use, audience, and role of the researcher or assessment professional (Grey, 2002; Henning & Roberts, 2016; Yousey-Elsener, 2019). The intended generalizability of the findings is another distinction between assessment and research (Upcraft & Schuh, 2002). Assessment reports are intended to represent the local institution, rather than provide broadly generalizable findings. Moreover, data produced via the outcomes assessment process do not typically afford inferences about program or curriculum effectiveness. Thus, it is critical to build programming that *should* be effective based on previous research, often found in the form of program effectiveness studies (Finney et al., 2021). Recognizing this need, we focused on inferences stated in published program effectiveness studies, which student affairs professionals may read and use to build programming on their campuses.

Previous Reviews of Published Articles

Previous reviews of the methodological characteristics of research published in higher education and student affairs journals are limited. Moreover, these reviews tallied study characteristics rather than appraised the credibility of inferences given the characteristics. Common themes among the reviews were frequent use of quantitative techniques, such as regression analysis (Ferrao 2020; Hutchinson & Lovell, 2004; Johnson et al., 2016; Volkwein et al., 1988; Wells et al., 2015) and infrequent use of rigorous experimental or quasi-experimental designs (Hutchinson & Lovell, 2004; Volkwein et al., 1988; Wells et al., 2015). Non-probability sampling (Langrehr et al., 2015), descriptive research (Kuh, Bean, Bradley, & Coomes, 1986; Kuh, Bean, Bradley, Coomes, & Hunter, 1986) and cross-sectional designs (Kuh, Bean, Bradley, & Coomes, 1986; Kuh, Bean, Bradley, Coomes, & Hunter, 1986; Langrehr et al., 2015) were common in student affairs journals and journals focused on understanding college students. Moreover, one review reported that only one third of the

studies used theory to guide the research, resulting in weak to non-existent connections between the current study and prior research (Langrehr et al., 2015).

None of the reviews evaluated the credibility of the authors' inferences given the research design, sampling, and analyses. None of the reviews summarized threats to the validity of inferences (Murnane & Willett, 2011; Shadish et al., 2002). Thus, we undertook this task for published studies in several student affairs journals to provide insight into the trustworthiness of claims regarding program effectiveness. Given our aim was to inform outcomes assessment and learning improvement practice, a summary of findings, didactic explanation, and call to action follow.

Method

Position Statement

We position ourselves as assessment specialists and higher-education researchers with a primarily post-positivist research orientation. While valuing other paradigms, we acknowledge the methods and results below promote a quantitative research methods paradigm when evaluating program effectiveness. This is intentional, given historical dialogue about causality (Shadish et al., 2002). Similar to other methodologists and interventionists (e.g., Robinson et al., 2018), we believe the best evidence upon which to base recommendations for programming is that which allows for causal claims.

It is important to emphasize that, despite the choice of quantitative studies as the focus of this manuscript, we value qualitative approaches as useful, legitimate, and sound approaches to assessment (Suskie, 2018) that we also use in practice. However, although the logic underlying causality does not differ across quantitative and qualitative approaches, the way in which data are viewed and interpreted does differ (Shadish et al., 2002). Therefore, to keep the study within a manageable scope and within our personal areas of expertise, we chose to focus on quantitative studies of program effectiveness. Additional studies that review effectiveness inferences based on qualitative data would be useful but were not included in this study.

Article Sources

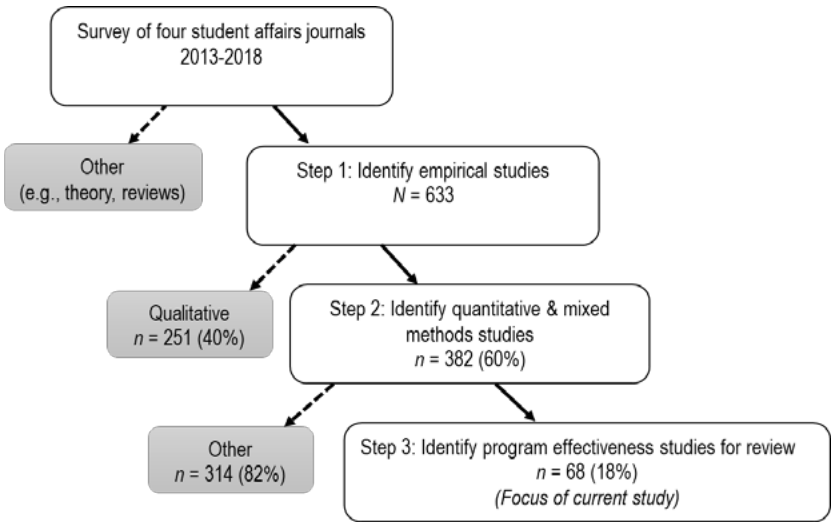
It is important to emphasize that, despite the choice of quantitative studies as the focus of this manuscript, we value qualitative approaches as useful, legitimate, and sound approaches to assessment (Suskie, 2018) that we also use in practice.

We reviewed articles published in four journals: *College Student Affairs Journal*, *Journal of College Student Development*, *Journal of Student Affairs Inquiry*, and *Journal of Student Affairs Research and Practice*. Three of the four journals are perceived as prestigious (Bray & Major, 2011), with the fourth (JSAI) being a new outlet. They are affiliated with student affairs organizations (e.g., ACPA, NASPA, and Student Affairs Assessment Leaders), have editorial boards, and conduct double-blind peer review. In our decades of experience working with student affairs colleagues, these are the journals they often reference, which aligns with studies of readership (Bray & Major, 2011). We reviewed five years (2013-2018) of articles for two reasons. First, research suggests statistical techniques tend to be stable over five years (Goodwin & Goodwin, 1985), and methodological approaches tend to be stable over 15 years (Volkwein et al., 1988). Second, the journals published issues between one and six times per year, resulting in an adequate sample of articles.

Article Selection

All 2013 to 2018 issues of the journals were examined. The process of selecting articles to review is shown in Figure 1. There were 633 published empirical studies (Step 1), comprised of 382 (60%) quantitative or mixed methods studies and 251 (40%) qualitative studies (Step 2). Of the quantitative or mixed methods studies, 68 (18%) reported an effectiveness study (Step 3). We retained quantitative studies in this step because our goal was to evaluate use of quantitative methods, analytic tools, and inferences. Although qualitative studies are essential to answering many questions about programming (e.g., implementation issues), the purpose of this study focused solely on the evaluation of quantitative program effectiveness studies.

Figure 1
Procedure for Selecting Program Effectiveness Studies for Review



The 68 program effectiveness studies were the focus of our review (references available upon request). A study was classified as an effectiveness study if it included *both* a program *and* an intentional student learning or development outcome. For example, a study of an alternative break program that evaluated program effectiveness with respect to influencing students' openness to diversity (intentional outcome) would be included in the current study. All 68 studies included a purpose statement or research question articulating that effectiveness was evaluated in terms of whether or not student learning or development outcomes were met. Some studies involved specific interventions on a single campus (n = 36, 53%), whereas others involved general interventions (e.g., alternative spring break) on multiple campuses (n = 32, 47%). Both were deemed effectiveness studies when effectiveness was considered relative to specific outcomes that were assessed. We did not review articles describing experiences (e.g., living on campus) that were not explicitly linked to intended student outcomes.

A study was classified as an effectiveness study if it included both a program *and* an intentional student learning or development outcome.

Rating Process

Five higher-education assessment professionals (two faculty members, two doctoral-level graduate students, one masters-level graduate student) rated the articles. The faculty members are formally trained in and teach quantitative methods and research design. The doctoral students each completed terminal master's degrees and multiple years of doctoral-level quantitative and research methods coursework. The masters-level graduate student completed multiple statistics and research methods courses and was completing an empirically-focused thesis. Combined, the raters have 50 years of experience in outcomes assessment.

Rating criteria (see Table 1) were based on recommendations in classic research methods texts (e.g., Shadish et al., 2002). During the initial two weeks of rating, all raters evaluated the same articles. Doing so permitted group discussion about the interpretation of rating criteria. Following the initial calibration weeks, each remaining article was evaluated by at least two raters (faculty-student or faculty-faculty pairing). Each of the raters individually rated their assigned articles and then met with another rater to adjudicate ratings, which then were combined into one spreadsheet for analysis. Quantitative analyses were conducted using SPSS 24. Study limitations noted by the 68 studies' authors and the open-ended rater comments were coded using NVivo 12 Pro.

Table 1
Rating Criteria for Published Program Effectiveness Studies

Criterion	Response Option
Citation Information	Journal, Volume, Issue, Year, Pages, Author, Title
General Information	
Type of study	Quantitative/mixed methods
Purpose of study	Description from article
What is (are) the measured outcome(s)?	Open-ended description
General intervention or specific program	General/Specific
Description of program or intervention	Open-ended description
Primary or secondary data source?	Primary/secondary
If secondary, what data were used?	Description of secondary data source
Information about Research Design	
Was there a comparison group?	Yes, No, Not clear (and open-ended description)
Was group membership self-reported?	Yes, No, Not clear
Was there random assignment to groups?	Yes, No, Not clear
Number of measurements of outcome	Number and description (e.g., pre-and post-test)
Additional details about research design	Open-ended description
What limitations did authors note?	Open-ended description
What limitations *should* be noted?	Open-ended description
Information about Sampling	
Sample size	Open-ended description
Was there random sampling?	Yes, No, Not clear
Was attrition noted?	Yes, No
Was attrition problematic? (and describe)	Yes, No, Not clear (and open-ended description)
What limitations did the authors note?	Open-ended description
What limitations *should* be noted?	Open-ended description
Information about Analyses	
Analysis	Open-ended description
Covariates	Open-ended description
Was analysis appropriate given *data collected?*	Yes, No (If no, then explanation)
Was analysis appropriate given *purpose of study?*	Yes, No (If no, then explanation)
Were inferential tests appropriately interpreted?	Yes, No (If no, then explanation)
Were effect sizes reported?	Yes, No (If yes, then description of type)
Were effect sizes appropriately interpreted?	Yes, No (if no, then explanation)
What limitations *should* be noted?	Open-ended description
Overall Conclusions	
What was the inference?	Open-ended description
Was the inference appropriate (given purpose, design, and analyses)?	Yes, No (if no, then explanation)
What was the inference?	Open-ended description
Was the inference appropriate (given purpose, design, and analyses)?	Yes, No (if no, then explanation)

Results

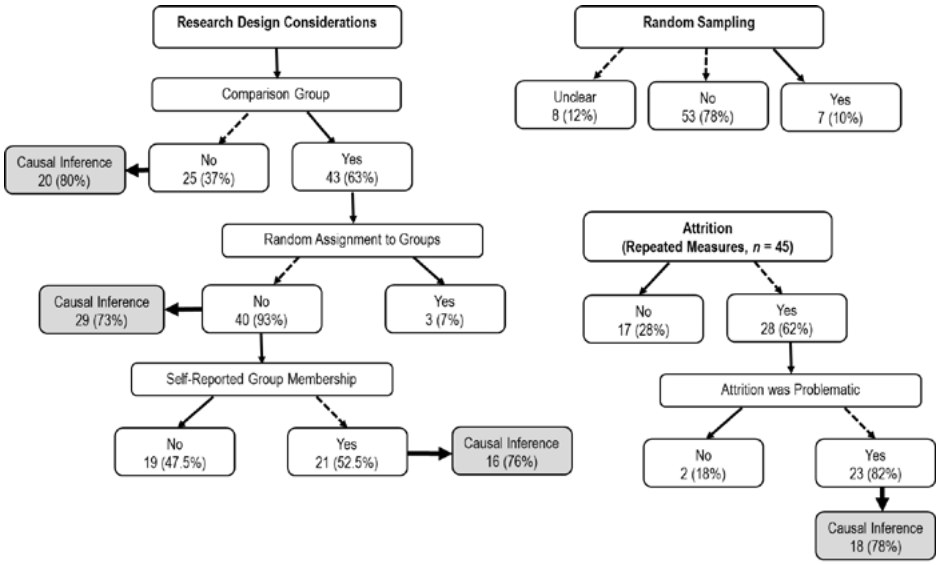
Of the 68 studies, 41 (60%) reported on data obtained from primary sources (i.e., new data collected for that particular study), whereas 27 (40%) reported on data from secondary sources (i.e., existing data collected by others).

Research Design

The sections that follow provide a summary of the research designs in the 68 reviewed studies. Figure 2 provides an overview of the findings that are described below.

Comparison group. A comparison group permits researchers to evaluate whether changes in learning or development may be attributed to causes other than the programming. In many

Figure 2
Design, Sampling, Attrition, and Causal Inferences for 68 Studies Reviewed



Note. Dashed arrows indicate threats to causal inferences. Gray boxes indicate the prevalence of unjustified causal inference.

cases, these changes may be just as feasibly due to natural development of students (i.e., maturation threat) or some other event that occurred at the same time as the intervention (i.e., history threat). Of the 68 studies, 43 (63%) included a comparison group, whereas 25 (37%) did not. Thus, for one-third of the studies, causal inferences about program effectiveness cannot be drawn as many threats to validity cannot be ruled out (see Table 2).

Acknowledging that within educational research it may not be feasible nor ethical to randomly assign students to groups, we expected the number of RCTs to be low. Of the 43 studies using a comparison group, three studies randomly assigned students to groups.

Random assignment. When interested in causal conclusions about a program's effectiveness, random assignment to groups (RCTs) is preferred, and otherwise are prone to self-selection bias (e.g., Shadish et al., 2002). For example, students who self-select into a service-learning program may be more apt to gain skills related to the outcome (e.g., cultural competence) than students who did not self-select due to pre-existing differences between groups in other variables (e.g., appreciation for diversity). Larger gains for the service-learning participants may be misinterpreted as positive program effects, when in fact the gains may have occurred with no programming.

Acknowledging that within educational research it may not be feasible nor ethical to randomly assign students to groups, we expected the number of RCTs to be low. Of the 43 studies using a comparison group, three studies randomly assigned students to groups. Of the 40 studies lacking random assignment to groups, 21 (53%) operationalized group membership through student self-report, frequently through retrospective self-reporting at post-test. If causal inferences are drawn from these non-RCT studies, they are tenuous, and it is necessary to note internal validity threats.

Number of time points. Collecting data at multiple time points permits evaluation of change over time. For program effectiveness studies, this typically means collection of data prior to and following programming, at a minimum. However, any inference that this change was due to programming is prone to validity threats associated with history, maturation, testing, and instrumentation. The addition of a comparison group aids with investigating these threats.

The number of time points for outcome measurements varied across studies. The most common design was pre-post (48%), followed by single-time point (34%) designs. The remainder included 3 (12%) or 4 time points (6%). Notably, the three RCTs included multiple time points. These RCTs, unlike single-group designs with multiple timepoints, directly address history and maturation effects.

Table 2
Appropriate Inferences Related to Specific Design Features

	Is __ a plausible threat?					
Description of Design	Selection	History	Maturation	Testing	Instrumentation	Appropriate Inference
RCT (random assignment) with a. random sampling b. no attrition c. pre- & post-test	No	Explore	Explore	Explore	Explore	Can infer cause-effect.
RCT with a. NO random sampling b. no attrition c. pre- & post-test	No	Explore	Explore	Explore	Explore	Results may not be generalizable to the population of interest given lack of random sampling. Otherwise, can infer cause-effect.
RCT with a. random sampling b. attrition is present c. pre- & post-test	Yes	Explore	Explore	Explore	Explore	If causal claims are desired, the plausibility of attrition as a threat must be considered. If attrition is non-random, characteristics of students remaining in the sample may lead to the appearance of an effect, when there is none.
RCT with a. random sampling b. no attrition c. No pre-test	No	Explore	Explore	No	No	Causal claims about the effect of the program should be made cautiously. Random assignment and control group data help to strengthen the claim, but there is no record of participants' outcomes scores prior to the program.
Two-Group Pre- & Post-Test Quasi-Experiment (same as first design, but no *random assignment*)	Yes	Explore	Explore	Explore	Explore	Variables related to selection into the program need to be considered as plausible threats to the accuracy of cause-effect claims.
One-Group Pre- & Post-Test (nothing else)	Yes	Yes	Yes	Yes	Yes	Causal claims should not be made without consideration of the plausibility of all threats. Exceptions may be when the information taught is *so specific or unusual* that the students would not have learned the information elsewhere.
One-Group Post-Test (nothing else)	Yes	Yes	Yes	No	No	Causal claims should not be made. Exceptions may be when the information taught is *so specific or unusual* that the students would not have learned the information elsewhere.

Note. "Yes" = a plausible threat. "No" = not a threat. "Explore" = threat may be plausible, but can be ruled out through group comparison. Selection = students selecting into or assigned to program differ on some variable related to outcome. History = event occurring at the same time as program may influence outcome. Maturation = students naturally develop or grow on outcome. Testing = changes in students' approach to completing an outcome measure (e.g., social desirability). Instrumentation = changes in test administration (e.g., modality, stakes).

Design limitations noted by authors. Transparency about validity threats is key to maintaining the credibility of program effectiveness studies (Levin & Kratochwill, 2013). Authors need to scrutinize causal claims and address limitations to the validity of those claims. Therefore, we recorded limitations noted in each of the studies. Of the 68 studies, 26 (38%) did not mention limitations related to research design. The remaining 42 (62%) studies' design limitations were coded into the broad themes of threats to validity, data, design, and operationalization of independent and dependent variables. The most mentioned threat to validity was selection bias (n = 14). Instrumentation, history, directionality of effect, and contamination were each mentioned once. Data limitations noted by studies' authors included lack of comprehensive sets of covariates, self-reported data, small sample size, and archival, retrospective, or secondary data. The most commonly mentioned design limitation was lack of random assignment (n = 14), cross-sectional/single-time point design (n = 10), and lack of control group (n = 5). Finally, 11 articles cautioned about the operationalization of treatment condition, particularly self-reported group membership.

Sampling and Attrition

If interested in representativeness of a population, then random sampling (or census data) without differential attrition is critical (Shadish et al., 2002). We reviewed characteristics related to sampling and attrition.

Random sampling. Of the 68 studies, 7 (10%) reported random sampling. For the remaining studies, 53 (78%) did not employ random sampling and 8 (12%) were unclear about sampling method. Some of the large secondary data sources reported initial random sampling, but data were retrieved only for specific subgroups that were not randomly sampled. Most reported high rates of attrition, which negated benefits of random sampling.

Attrition. Of the 45 repeated measures studies, 28 (62%) reported attrition. Of these, we rated 23 (82%) instances as problematic, based upon the percent of attrition and lack of acknowledgement of attrition as an issue. For example, it was common for 26% to 50% of students to provide data at time-point 1 but not time-point 2. Of the three RCT studies, one reported random sampling with non-problematic attrition.

Sample size. When tabulating sample size, we included the final sample size reported for analyses. When there were multiple samples, we recorded the size of the largest sample. Sixty-five studies reported sample sizes, ranging from a minimum of 8 participants to a maximum of 15,847. The median sample size was 436 (25th percentile = 100; 75th percentile = 1,502).

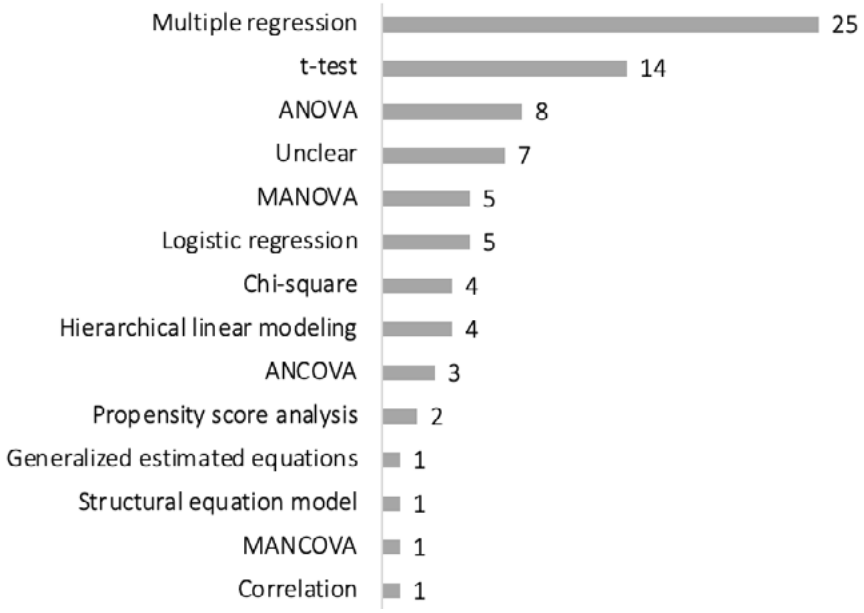
Sample limitations noted by authors. Of the 68 studies, 27 (40%) did not note limitations regarding sampling or attrition. Commonly mentioned limitations were generalizability (n = 15) or the sample composition was not representative of the population (n = 23). Eleven noted issues related to response rate or attrition, and eight noted small sample size. Other limitations included non-random sampling (n = 5), convenience sampling (n = 2), inadequate demographic information (n = 4), and clustered data (n = 1).

Notably absent for most of these studies were tests of the assumption of homogeneity of regression slopes. Without reporting of this critical assumption, we could not determine if the analyses or their interpretation were appropriate.

Analysis

Type. Types of analyses varied across the 68 studies (see Figure 3). Note, the numbers reported in Figure 3 total to greater than 68, because some studies involved more than one type of analysis. Notably absent from many studies were descriptive statistics (e.g., means, standard deviations). Of the 25 studies employing multiple regression, 23 were ANCOVA-type analyses that included a grouping variable (e.g., intervention versus comparison group) and covariates employed as "control" variables. These studies included 3 to 22 covariates (median = 8). Common choices for covariates were pre-test scores on the outcomes, demographic characteristics (e.g., gender), students' pre-college characteristics (e.g., high school involvement), and student- and institution-level college characteristics (e.g., students' major, type of college). Notably absent for most of these studies were tests of the assumption of homogeneity of regression slopes. Without reporting of this critical

Figure 3
Analyses reported in the 68 studies reviewed



assumption, we could not determine if the analyses or their interpretation were appropriate. Consequently, we assigned ratings of "inappropriate" to these analyses and interpretations of their inferential tests.

Many studies did not report descriptive statistics (e.g., distributions of scores, standard deviations). Therefore, readers cannot self-assess if the data align with statistical assumptions and whether the presence of floor or ceiling effects explain the lack of change in the outcome from pre- to post-intervention.

Appropriateness of analysis. Proper statistical analysis is necessary to achieve statistical-conclusion validity (Shadish et al., 2002). We evaluated the appropriateness of the analysis by examining if it (and its associated statistical assumptions) aligned with the type of data being modeled and if it aligned with the purpose of the study (i.e., research questions).

Of the 68 studies, 32 (47%) clearly aligned the analyses to the type of data collected. Twenty-five (37%) studies reported analyses that were misaligned to the data (e.g., ANOVA with continuous predictors requiring artificial categorization of predictors). An additional 6 studies (9%) clearly aligned some analyses to the type of data collected, while at the same time misaligning other analyses. For 5 (7%) studies, it was unclear from the studies' description whether the analyses were aligned or not.

Assumption testing was seldom reported. In addition to the lack of testing the homogeneity of regression assumption for ANCOVA, there was infrequent discussion of variability- or distribution-related assumptions. Analyses conducted to explain variability in continuous outcomes (e.g., regression, ANOVA) lack utility if there is little variability in outcomes to explain. Many studies did not report descriptive statistics (e.g., distributions of scores, standard deviations). Therefore, readers cannot self-assess if the data align with statistical assumptions and whether the presence of floor or ceiling effects explain the lack of change in the outcome from pre- to post-intervention.

Of the 68 studies, 27 (40%) appropriately aligned their statistical analyses with the purpose of the study (i.e., research question posed). Fourteen studies (16%) were unclear. However, 30 (44%) studies reported analyses misaligned with the research question (e.g., a research question about differential change in the outcome across intervention and comparison groups without testing the hypothesized interaction). Thus, for 30 studies, the results presented could not provide answers to the research questions posed.

Interpretation. Of the 68 studies, 31 (46%) appropriately interpreted the inferential statistical tests. Eight (11%) studies were unclear. Common themes among the remaining 29 (43%) included implying or misinterpreting main effects in the presence of interactions and

noting "significance" of results without conducting inferential tests. Results sections with unclear terminology and a mismatch between text and table information led to difficulty in interpretation.

Effect size (ES). Measures of ES aid in understanding the practical significance of findings. Of the 68 studies, 58 (85%) reported ES measures. Ten (15%) studies reported ES measures for some but not all analyses or did not report any ES measure.

The findings of our review are a result of curiosity about the quality of program effectiveness evidence published in student affairs journals, given expectations that professionals use research to identify programming that impacts desired outcomes.

As expected, type of ES varied by type of analysis. For 15 of the 25 multiple regression analyses that included measures of ES, 12 reported R^2 for the model, 1 reported R^2-change,10 reported standardized coefficients, and 2 reported unstandardized coefficients. When reporting t-test findings, out of the 14 studies, 1 reported Cohen's *d* and 4 reported raw mean differences. When reporting ANOVA findings, several reported eta-squared and partial eta-squared.

Although most authors provided effect size values, few authors interpreted those values for readers. Of the 68 studies, 24 (35%) both presented and explained ES values, thereby providing an interpretation of the practical significance of their findings. The remainder (65%) did not report ES, did not interpret ES, or inaccurately interpreted ES.

Causal Inferences

Given the focus of these studies, authors made inferences from results regarding program effectiveness. We evaluated the appropriateness of causal inferences. An inference was flagged for review if the discussion of, or implications from, the findings were reported with wording, such as Program X "*impacted*," "*affected*," or "*led to gains in*" outcome Y. Implications sections commonly included program suggestions informed by the study's results, implying a causal relation between programming and outcomes. If authors uncovered non-significant results and inferred a non-causal relation, we evaluated the inference regarding a lack of causality for alignment to the research design and analyses.

Of the 68 studies evaluated, one study (Thatcher, 2016) was able to make an appropriate causal inference given its design, data, and analyses. Of the remaining 67 studies, 12 (18%) drew appropriate non-causal inferences from the findings, remaining tentative about the causal impact of programming on the outcome. The remaining 55 (82%) studies included a causal claim in the results, discussion, or implication sections of the article.

To better understand when inappropriate causal inferences were made, we examined the extent to which authors drew causal inferences when employing research designs that did not support such inferences. Of the 23 single-time-point design studies, 12 (52%) made causal inferences. Of the 25 studies with no comparison group, 20 (80%) included causal inferences. Of the 40 studies with a comparison group, but non-random assignment, 29 (73%) included causal inferences. Of the 21 studies with non-random assignment and for which students self-reported group membership, 16 (76%) included causal inferences.

We recorded statements from the 68 studies' results, discussion, and implications sections that led to the rating of "inappropriate causal claim." The statements were coded for themes. The most commonly identified theme was "effect of Program X on outcome Y." Other common phrasings included "benefits of," "influenced," "improved/promoted," "result of participation in," "efficacy/effectiveness," "fosters," "successful program," "transformative," and "reduced." Another common theme was the implication of no program effect on the outcome given non-significant results (e.g., "Program X has *no impact* on outcome Y"). Nonetheless, the results were often used inappropriately to argue for or against future or additional programming to impact the particular outcome.

Discussion

When discussing peer review, Carpenter (2001) noted: "This is not a call to be critical of each other as people, but to be very critical of our work and our results. Scholars evaluate each other's work" (p. 305). The findings of our review are a result of curiosity about the quality of program effectiveness evidence published in student affairs journals, given expectations that professionals use research to identify programming that impacts

desired outcomes. Using criteria in Table 1, we examined 68 program effectiveness studies published between 2013 and 2018.

Similar to previous methodological reviews (Hutchinson & Lovell, 2004; Johnson et al., 2016; Wells et al., 2015), statistics such as multiple regression, *t*-tests, and ANOVA were the most common analyses. Unlike Wells and colleagues (2015), who noted frequent reporting of descriptive statistics, the studies we reviewed did not typically report descriptive statistics. Notably absent was reporting of assumptions testing, threatening the validity of conclusions drawn from analyses. Despite frequent claims of program effectiveness, the research design and analyses did not often support such claims due to highly-plausible threats to the validity of those claims. This finding is not new. In their review of 21 years of research, Reinhart and colleagues (2013) noted an increase in causal inferences drawn from correlation studies.

Moreover, many causal claims were unaccompanied by acknowledgment of limitations or threats to the validity of these inferences. To provide credible and trustworthy evidence of program effectiveness, at a minimum, professionals need to acknowledge plausible threats to the validity of causal claims (Levine & Kratochwill, 2013; Shadish et al., 2002). Consider a hypothetical service-learning course with the following outcome: "As a result of participation, students will demonstrate increased openness to diversity". Openness to diversity is assessed before and after the course for students who opt to participate and is found to increase. The following are plausible threats to the validity of the causal statement that the service-learning course caused (or "led to") increased openness to diversity: 1) selection bias (e.g., students interested in diversity enroll in the course), 2) attrition (e.g., uninterested students drop out of the course or skip the post-test), 3) history (e.g., another event on campus influenced the outcome), 4) maturation (e.g., students naturally develop openness to diversity), 5) testing (e.g., students respond differently to the post-test because they realize the focus on diversity), and 6) instrumentation (e.g., instructors communicate greater importance of the post-test than the pre-test, resulting in higher scores at post-test). It is essential to critically evaluate evidence and report plausible threats to the accurate interpretations of findings if program effectiveness studies are to be trustworthy (Upcraft & Schuh, 2002). Table 2 provides a concise guide to evaluate studies for threats to validity.

Call for Action

Research-to-practice efforts require being able to understand research. One course in statistics and research methods is not enough if professionals are expected to evaluate the credibility of inferences in effectiveness studies.

We understand that gathering rigorous evidence of effective programming is challenging. Random selection or random assignment of students to programs may not be feasible. The collection of pre-post data with a comparison group may not be feasible. Moreover, given the low-consensus nature of the student affairs profession (Torres et al., 2019) and higher education in general (Wells et al., 2015), limiting "evidence" to RCT studies risks over-narrowing the information available to professionals. We are *not* advocating for RCT studies as the only way to assess program effectiveness. Instead, we are advocating for professionals to 1) acknowledge threats to the validity of causal inferences, 2) draw appropriate inferences given the plausibility of threats for a specific research design, and 3) consider quasi-experimental designs that can support causal inferences in the absence of RCTs (regression discontinuity designs, interrupted time series designs, propensity score matching; Murnane & Willett, 2011). All of these support research-to-practice efforts called for in the domains of student affairs (Finney & Horst, 2019a, 2019b) and outcomes assessment (Horst & Prendergast, 2020; Singer-Freeman & Robinson, 2020).

We also echo calls for changes in graduate school training, journal review practices, and professional organization practices (Carpenter, 2001; Malaney, 2002; Wells et al., 2015). In the 68 studies reviewed, lack of clarity when describing designs and analyses suggested that some authors were not familiar with the methods they were using. Professionals must have a repertoire of research techniques not only to conduct research but to evaluate its quality (Schroeder & Pike, 2001). Research-to-practice efforts require being able to understand research. One course in statistics and research methods is not enough if professionals are expected to evaluate the credibility of inferences in effectiveness studies. Without increased training in methodology, assessment professionals will need to provide support to colleagues who are unable to independently evaluate the quality of research. With that said, assessment

professionals themselves may need to acquire additional knowledge of statistics and research methods (Curtis et al., 2020), in order to fulfill this role.

Journal editors and reviewers can also contribute to an increase in the quality of evidence. Through the double-blind peer review process, journals aim to provide content and inferences that are scrutinized and shared to improve practice (Liddell, 2019). Professionals with methodological expertise must volunteer time to the review process and hold the profession to high standards. If published studies include misinformation, the burden then falls on readers to evaluate research credibility. Rigorous review processes can reduce this burden.

Moreover, when journals require an implications section, researchers face conflicting roles, in which they need to accurately convey the limited inferences from their single study and yet are asked to speculate about broad implications for practice (Robinson et al., 2013). In doing so, the temptation is to fall into causal language. Consequently, if readers skip over methods and results sections and head straight to the discussion section, they are likely to believe the causal implications. To address this issue, Robinson and colleagues (2013) suggested the following be added to education research journal policies: "Contributors should restrict their discussion and conclusions to their data and not offer recommendations for educational practice nor speculate about the educational policy implications of their research" (p. 291). Instead, they recommended that implications from research be developed via conversation among practitioners. Professional organizations are a venue for such conversations. Organizations can also influence the quality of implications from these conversations by providing training on causal inferences.

Professionals creating programming must work to become fluent in their critiques of published literature. These skills can be developed through critical reading of published research.

Professionals creating programming must work to become fluent in their critiques of published literature. These skills can be developed through critical reading of published research. Methodological review articles, such as the current study, expose readers to the variable quality of published research. These critiques also provide useful training in identifying and understanding links between design, results, and interpretations.

Finally, assessment professionals should ask fundamental questions about program rationale when supporting colleagues engaged in outcomes assessment. Simple questions such as "What evidence supports the belief that this strategy/program will result in that student learning outcome?" may reveal that no credible evidence exists to support a programming decision (Finney & Buchanan, 2021). This awareness may provide insight into disappointing assessment results (i.e., no student learning) and struggles with learning improvement efforts. This awareness may also spur frustration for professionals who spent years implementing and assessing programming they believed would be effective given published claims. Assessment professionals can help colleagues process this frustration, frame this realization as an opportunity, and locate credible evidence of effectiveness to build should-be-effective programming.

References

American College Personnel Association & National Association of Student Personnel Administrators (2015). *ACPA/NASPA professional competency areas for student affairs educators*. Authors.

Banta, T., & Blaich, C. (2011). Closing the assessment loop. *Change: The Magazine of Higher Learning*, *43*, 22-27. https://doi.org/10.1080/00091383.2011.538642

Bickman, L. & Reich, S. (2015). Randomized controlled trials: A gold standard or gold plated? In S.I. Donaldson, C.A. Christie, & M.M. Mark (Eds.), *Credible and actionable evidence: The foundation for rigorous and influential evaluations* (2nd ed., pp. 83-113). Sage. https://dx.doi.org/10.4135/9781483385839

Bray, N., & Major, C. (2011). Status of journals in the field of higher education. *The Journal of Higher Education*, *82*, *479-503*. https://www.jstor.org/stable/29789535?seq=1

Carpenter, S. (2001). Student affairs scholarship (re?)considered: Toward a scholarship of practice. *Journal of College Student Development*, *42*, 301-318.

Cook, B. G., Smith, G. J., & Tankersley, M. (2011). Evidence-based practices in education. In K. R. Harris. S. Graham, & T. Urdan (Eds.), *APA educational psychology handbook*, volume 1 (pp. 495–528). Washington, DC: American Psychological Association. https://doi.org/10.1037/13273-017

Cooper, J., Mitchell, D., Eckerle, K., & Martin, K. (2016). Addressing perceived skill deficiencies in student affairs graduate preparation programs. *Journal of Student Affairs Research and Practice*, *53*, 107-117. https://doi.org/10.1080/19496591.2016.1121146

Curtis, N.A., Anderson, R. D., & Van Dyke, R. (2020). A field without a discipline? Mapping the uncertain and often chaotic route to becoming an assessment practitioner. *Research and Practice in Assessment*, *15*(2), 1-8. https://www.rpajournal.com/dev/wp-content/uploads/2020/08/A-Field-Without-A-Discipline.pdf

Deniston, O., Rosenstock, I., & Getting, V. (1968). Evaluation of program effectiveness. *Public health reports*, *83*(4), 323.

Ferrao M. E. (2020). Statistical methods in recent higher education research. *Journal of College Student Development*, *61*, 366-371. https://doi.org/10.1353/csd.2020.0033

Finney, S. & Buchanan, H. (2021). A more efficient path to learning improvement: Using repositories of effectiveness studies to guide evidence-informed programming. *Research & Practice in Assessment*, *16*, 36-48. https://www.rpajournal.com/a-more-efficient-path-to-learning-improvement-using-repositories-of-effectiveness-studies-to-guide-evidence-informed-programming/

Finney, S. & Horst, S. (2019a). Standards, standards, standards: Mapping professional standards for outcomes assessment to assessment practice. *Journal of Student Affairs Research and Practice*, *56*, 310-325. https://doi.org/10.1080/19496591.2018.1559171

Finney, S. & Horst, S. (2019b). The status of assessment, evaluation, and research in student affairs. In V. L. Wise & Z. Davenport (Eds.), *Student affairs assessment, evaluation, and research: A guidebook for graduate students and new professionals* (pp. 3-19). Charles C. Thomas, Publisher, Ltd. https://ebookcentral.proquest.com/lib/jmu/detail.action?docID=5722940

Finney, S., Wells, J., & Henning, G. (2021). *The need for program theory and implementation fidelity in assessment practice and standards* (Occasional Paper No. 51). Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment (NILOA). https://www.learningoutcomesassessment.org/wp-content/uploads/2021/03/Occ_Paper_51-1.pdf

Fulcher, K., & Prendergast, C. (2019). Lots of assessment, little improvement? How to fix a broken system. In S. Hundley & S. Kahn (Eds.), *Trends in assessment: Ideas, opportunities, and issues in higher education* (pp. 157-174). Stylus.

Grace-Odeleye, B., & Santiago, J. (2019). A review of some diverse models of summer bridge programs for first-generation and at-risk college students. *Administrative Issues Journal: Education, Practice & Research*, *9*, 35-47. https://doi.org/10.5929/9.1.2

Goodwin. L., & Goodwin, W. (1985). Statistical techniques in *AERJ* articles, 1979-1983: The preparation of graduate students to read the educational research literature. *Educational Researcher*, *14*, 5-11. https://www.jstor.org/stable/1174902

Gray, P. (2002). The roots of assessment: Tensions, solutions, and research directions. In T. W. Banta and Associates (Eds.) *Building a scholarship of assessment* (pp. 49-66). San Francisco, CA: Jossey Bass.

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

Hilligoss, B. & Rieh, S. (2008). Developing a unifying framework of credibility assessment: Construct, heuristics, and interaction in context. *Information Processing and Management*, *44*, 1467-1484. https://doi.org/10.1016/j.ipm.2007.10.001

Horst, S.J. & Prendergast, C. (2020). The assessment skills framework: A taxonomy of assessment knowledge, skills and attitudes. *Research & Practice in Assessment*, *15*, 1-25. https://www.rpajournal.com/dev/wp-content/uploads/2020/05/The-Assessment-Skills-Framework-RPA.pdf

Hutchinson, S. & Lovell, C. (2004). A review of methodological characteristics of research published in key journals in higher education: Implications for graduate research training. *Research in Higher Education*, *45*, 383-403. https://link.springer.com/content/pdf/10.1023/B:RIHE.0000027392.94172.d2

Jankowski, N., Timmer, J., Kinzie, J., & Kuh, G. (2018). *Assessment that matters*: *Trending toward practices that document authentic student learning*. University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment. https://files.eric.ed.gov/fulltext/ED590514.pdf

Johnson, M., Wagner, N., & Reusch, J. (2016). Publication trends in top-tier journals in higher education. *Journal of Applied Research in Higher Education*, *8*, 439-454. http://dx.doi.org/10.1108/JARHE-01-2015-0003

Kuh, G., Bean, J., Bradley, R., & Coomes, M. (1986). Contributions of student affairs journals to the literature on college students. *Journal of College Student Personnel*, *27*, 292–304.

Kuh, G., Bean, J., Bradley, R., Coomes, M., & Hunter, D. (1986). Changes in research on college students published in selected journals between 1969 and 1983. *Review of Higher Education*, *9*, 177-192.

Langrehr, K., Phillips, J., Melville, A., & Eum, K. (2015). Determinants of nontraditional student status: A methodological review of the research. *Journal of College Student Development*, *56*, 876-881. http://dx.doi.org/10.1353/csd.2015.0090

Levin, J. & Kratochwill, T. (2013). Educational/psychological intervention research circa 2012. In I. B. Weiner (Series Ed.), W. M. Reynolds & G. E. Miller (Volume Eds.), *Handbook of psychology*: *Vol. 7. Educational psychology* (2nd ed., pp. 465–492). Wiley.

Liddell, D. (2019). 60 years of scholarship. *Journal of College Student Development*, *60*, 641-644. http://dx.doi.org/10.1353/csd.2019.0059

Malaney, G. (2002). Scholarship in student affairs through teaching and research. *NASPA Journal*, *39*, 132-146. https://doi.org/10.2202/1949-6605.1795

Miller, R. (2015). How people judge the credibility of information: Lessons for evaluation from cognitive and information sciences. In S. Donaldson, C. Christie, & M. Mark (Eds.), *Credible and actionable evidence*: *The foundation for rigorous and influential evaluations* (2nd ed., pp. 83-113). Sage.

Muller, K., Grabsch, D., & Moore, L. (2018). Factors influencing student affairs professionals' attainment of professional competencies. *Journal of Student Affairs Research and Practice*, 55, 54-64. https://doi.org/10.1080/19496591.2017.1345755

Murnane, R. & Willett, J. (2011). *Methods matter*: *Improving causal inference in educational and social science research*. Oxford University Press.

Pope, A., Finney, S., & Bare, A. (2019). The essential role of program theory: Fostering theory-driven practice and high-quality outcomes assessment in student affairs. *Research & Practice in Assessment*, *14*, 5–17. https://files.eric.ed.gov/fulltext/EJ1223397.pdf

Reinhart, A., Haring, S., Levin, J., Patall, E., & Robinson, D. (2013). Models of not-so-good behavior: Yet another way to squeeze causality and recommendations for practice out of correlational data. *Journal of Educational Psychology*, *105*, 241-247. https://doi.org/10.1037/a0030368

Robinson, D., Levin, J., Graham, S., Schraw, G., Fuchs, L., & Vaughn, S. (2018). Improving the credibility of educational intervention research. In A. M. O'Donnell (Ed.), *Handbook of educational psychology*. Oxford University Press.

Robinson, D., Levin, J., Schraw, G., Patall, E., & Hunt, E. (2013). On going (way) beyond one's data: A proposal to restrict recommendations for practice in primary educational research journals. *Educational Psychology Review*, *25*, 291-302. https://doi.org/10.1007/s10648-013-9223-5

Schroeder, C., & Pike, G. (2001). The scholarship of application in student affairs. *Journal of College Student Development*, *42*, 342-355.

Shadish, W., Cook, T., & Campbell, D. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin Company.

Singer-Freeman, K., & Robinson, C. (2020). *Grand challenges in assessment*: *Collective issues in need of solutions* (Occasional Paper No. 47). University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment. https://files.eric.ed.gov/fulltext/ED612032.pdf

Smith, K., & Finney, S. (2020). Elevating program theory and implementation fidelity in higher education: Modeling the process via an ethical reasoning curriculum. *Research & Practice in Assessment*, *15*, 1-13. https://www.rpajournal.com/dev/wp-content/uploads/2020/09/Elevating-Program-Theory-and-Implementation-Fidelity-in-Higher-Education.pdf

Smith, K., Good, M., & Jankowski, N. (2018). Considerations and resources for the learning improvement facilitator. *Research & Practice in Assessment*, *13*, 20-26. http://www.rpajournal.com/dev/wp-content/uploads/2019/02/W18_A4.pdf

Suskie, L. (2018). *Assessing student learning*: *A common sense guide* (3rd ed.) San Francisco CA: Jossey-Bass.

Thatcher, W. G. (2016). FREAKS - A new program for student life and success. *College Student Affairs Journal*, *34*, 48-55. https://doi.org/10.1353/csj.2016.0002

Torres, V., Jones, S., & Renn, K. (2019). Student affairs as a low-consensus field and the evolution of student development theory as foundational knowledge. *Journal of College Student Development*, *60*, 645-658. https://doi.org/10.1353/csd.2019.0060

Upcraft, M., & Schuh, J. (2002). Assessment vs. research. *About Campus*, *7*, 16-20. https://doi.org/10.1177/108648220200700104

Valentine, J.C., Hirschy, A.S., Bremer, C.D., Novillo, W., Castellano, M., & Banister, A. (2011). Keeping at-risk students in school: A systematic review of college retention programs. *Educational Evaluation and Policy Analysis*, *33*, 214 – 234. https://doi.org/10.3102/0162373711398126

Volkwein, J., Carbone, D., & Volkwein, E. (1988). Fifteen years of scholarship. *Research in Higher Education*, *28*, 271-280. https://www.jstor.org/stable/40195866?seq=1

Wells, R., Kolek, E., Williams, E., & Saunders, D. (2015). "How we know what we know": A systematic comparison of research methods employed in higher education journals, 1996-2000 v. 2006-2010. *The Journal of Higher Education*, 86, 171-198. https://doi.org/10.1080/00221546.2015.11777361

Yousey-Elsener, K. (2019). Development of competencies in assessment, evaluation, and research, with terms and concepts. In V. L. Wise & Z. R. Davenport (Eds.), *Student affairs and assessment*, *evaluation*, *and research*: *A guidebook for graduate students and new professionals* (pp. 20-47). ProQuest Ebook Central https://ebookcentral.proquest.com