2-11-2019

Metrics and Methods Used to Compare Student Performance Data in Chemistry Education Research Articles

Michael R. Mack
University of Washington - Seattle Campus

Cory Hensen
Portland State University, chensen@pdx.edu

Jack Barbera
Portland State University, jbarbera@pdx.edu

Citation Details

Published as Michael M. Mack, Cory Hensen, and Jack Barbera, "Metrics and Methods Used to Compare Student Performance Data in Chemistry Education Research Articles," Journal of Chemical Education, 2019, 96, 3, 401-413, DOI:10.1021/acs.jchemed.8b00713.

Metrics and Methods Used to Compare Student Performance Data in Chemistry Education Research Articles

Michael R. Mack,† Cory C. Hensen,‡ and Jack Barbera‡

†Department of Chemistry, University of Washington, Seattle, Washington, 98195-5850, United States

‡Department of Chemistry, Portland State University, Portland, Oregon, 97207-0751, United States

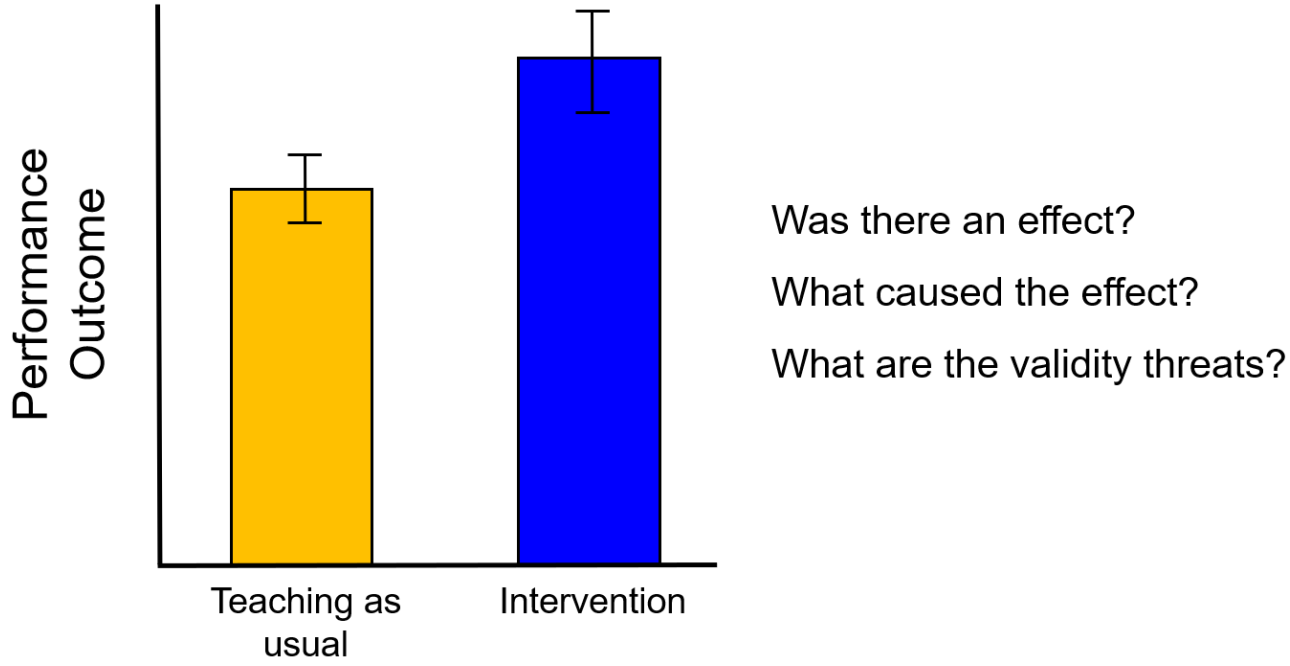
ABSTRACT

Quasi-experiments are common in studies that estimate the effect of instructional interventions on student performance outcomes. In this type of research, the nature of the experimental design, the choice in assessment, the selection of comparison groups, and the statistical methods used to analyze the comparison data dictate the validity of causal inferences. Therefore, gathering and reporting validity evidence in causal studies is of utmost importance, especially when conclusions have real policy implications for students and faculty, among other stakeholders. This review examines 24 articles that reported quantitative investigations of the effect of instructional interventions on performance-based outcomes conducted within undergraduate chemistry courses. Specifically, we examined four aspects of conducting such evaluations, including: (1) the type of quasi-experimental design used to study the relationship between interventions, students, outcomes, and settings, (2) the metrics used to measure performance outcomes, (3) the type of groups used to contrast outcomes across experimental conditions, and (4) the statistical methods used to analyze the comparison data. Through the examination of these four aspects of causal studies, together with a validity typology for quasi-experimental designs, we catalogued the metrics and methods used to compare student performance outcomes across varied instructional contexts. Recommendations for researchers and practitioners planning quasi-experimental investigations and interpreting results from causal studies in chemistry education are provided.

Reference Information:

Michael M. Mack, Cory Hensen, and Jack Barbera, “Metrics and Methods Used to Compare Student Performance Data in Chemistry Education Research Articles,” *Journal of Chemical Education*, 2019, 96, 3, 401-413, DOI:10.1021/acs.jchemed.8b00713.

GRAPHICAL ABSTRACT



KEYWORDS

Chemistry Education Research, Testing/Assessment, First-Year Undergraduate/General, Organic Chemistry

INTRODUCTION

With nearly a one hundred year history, chemistry education research (CER) has come to be a theory-driven and experimental discipline.[1,2] Constructivist learning theories have guided the exploration and elucidation of how students and teachers think, feel, and act in chemistry classrooms, and the products of this body of knowledge have challenged the theoretical underpinnings of traditional lecture-based teaching strategies,[3] which dominate Science, Technology, Engineering, and Mathematics (STEM) classrooms.[4] Consequently, one branch of CER focuses on the evaluation of pedagogies (constructivist or otherwise), often in contrast to "traditional" lecture-based instruction. Researchers and practitioners in this branch commonly conduct quasi-experiments to frame investigations of the relationship between instructional strategies and student outcomes. In these studies, students are typically distributed non-randomly across experimental conditions. For example, a study may compare the effect of an instructional intervention on student outcomes in a general chemistry course to student outcomes in the same course taught the previous year, but under a different teaching model. The shared purpose of

these studies is to compare the outcome of one group receiving an instructional intervention (or "treatment") to the outcome of one or more groups receiving either the absence of that treatment or an alternative treatment. Contrasts across groups on the outcome variable can be used to estimate the effect of the instructional intervention.

Randomized controlled trials (RCTs) share many of the structural and logical elements of quasi-experimental designs, but by contrast RCTs are uniquely characterized by the random assignment of participants to the treatment or comparison groups, and they are often touted as the "gold standard" in experimentation for many good reasons. Shadish et al.[5] explained that RCTs (1) equate groups across specified variables before a treatment begins, (2) reduce the plausibility of alternative explanations for observed effects by randomly distributing validity threats across conditions, (3) decouple treatment variables and error terms in regression analyses, and (4) separate the sources of variability in outcomes. In CER, however, RCTs are often impractical, hence a reliance on quasi-experiments for estimating the effects of instructional interventions on student outcomes. Consequently, the primary goal of research within the quasi-experimental paradigm is to rule out plausible alternative explanations for observed effects and sources of bias that emerge when participants are non-randomly distributed across experimental conditions.

Guiding Questions

The purpose of this review is to examine the experimental methods reported in CER articles published in the Journal of Chemical Education (hereinafter referred to as the *Journal*) that have been utilized when comparing student performance outcomes across different teaching conditions. We hope that our focus on the nature of experimental designs, performance metrics, comparison groups, and statistical analysis techniques will support researchers and practitioners in advancing the theory and measurement of relationships between how we teach and how students perform on assessments in chemistry education. To this end, the following questions were used to guide the analysis of articles selected for review: (1) What experimental designs have been used to study the effects of instructional interventions on student performance outcomes? (2) What metrics have been used to measure the outcomes? (3) What types of groups have been used to compare outcomes across experimental conditions? (4) What quantitative methods have been used to analyze the comparison data? With respect

to each question, we also examined the kind of validity evidence used to support or refute inferences made about treatments and the observed effects.

CONCEPTUAL FRAMEWORK

Before outlining the methodological steps for this review, we first describe a framework used to aid in our examination of research articles. Our conceptual framework used definitions and operationalizations of quasi-experiments described in the seminal works of Campbell, Stanley, Cook, and Shadish.[5–8] These models of experimentation were refined over the second half of the twentieth century and into the twenty-first century and they are heralded as some of the most influential methodological resources for designing and conducting experiments in the social sciences. The definitions and operationalizations of quasi-experiments outlined in this section are not meant to be an exhaustive review of quasi-experimental designs, therefore, we direct the reader's attention to Refs. 5-8 for additional details.

Quasi-Experimental Design

Quasi-experimental designs provide models of relationships between people, treatments, outcomes, and settings. Table 1 outlines the nine design frameworks considered for this study. For each framework, Table 1 includes a name and a label, a diagram illustrating the design elements and their relationships, and a suitable research question. The diagrams can be interpreted as follows. The one group posttest-only framework, Design #1, is represented by X – O to signify that the observation of students on an outcome (O) follows the treatment condition (X). The one group pretest-posttest framework, Design #2, is represented by O_1 – X – O_2 to signify that students are first measured on some pretest (O_1), followed by exposure to a treatment condition (X), and then measured again with the same test (O_2). Now consider Design #4, the pretest-posttest alternative treatment (or no treatment) control group design. Under this framework, two groups are measured on some pretest (O_1) under similar conditions and then one group receives treatment X while the other group receives an alternative treatment Y or the absence of that treatment. Both groups are followed up with a posttest (O_2) which occur at roughly the same time and under the same conditions. For brevity, we only consider designs with two comparison groups. However, these designs can be extended to include three or more groups.

Table 1. Matrix of Design Frameworks, Design Schematics, and Example Research Questions in the Context of Performance Outcomes Appropriate for Each Framework

Design Label	Design Framework	Diagram[a,b]	Example Research Questions
1	One Group Posttest-Only Design	X – O	What is the status of student performance outcomes following X?
2	One Group Pretest-Posttest Design	O_1 – X – O_2	How does performance on the outcome change following X?
3	Posttest-Only Alternative Treatment (or No Treatment) Control Group Design	Group 1: X – O Group 2: Y – O	What is the effect of X on the outcome compared to Y?
4	Pretest-Posttest Alternative Treatment (or No Treatment) Control Group Design	Group 1: O_1 – X – O_2 Group 2: O_1 – Y – O_2	What is the effect of X on the outcome compared to Y, given O_1?
5	Pretest-Posttest Alternative Treatment (or No Treatment) Control Group Regression Discontinuity Design[c]	Group 1: O_1 | – X – O_2 Group 2: O_1 $\|_c$ – Y – O_2	Same as 4[d]
6	Posttest-Only Alternative Treatment (or No Treatment) Control Group Design with Removed Treatment	Group 1: X – O_1 – Y – O_2 Group 2: Y – O_1 – Y – O_2	Same as 3, plus: What is the downstream effect of X on the outcome compared to Y?[e]
7	Pretest-Posttest Alternative Treatment (or No Treatment) Control Group Design with Removed Treatment	Group 1: O_1 – X – O_2 – O_3 – Y – O_4 Group 2: O_1 – Y – O_2 – O_3 – Y – O_4	Same as 4, plus: What is the downstream effect of X on the outcome compared to Y?[e]
8	Posttest-Only Alternative Treatment (or No Treatment) Control Group Design with Repeated Measures	Group 1: X – O_1 – X – O_2 Group 2: Y – O_1 – Y – O_2	Same as 3, plus: What is the sustained effect of X on the outcome compared to Y?[e,f]
9	Pretest-Posttest Alternative Treatment (or No Treatment) Control Group with Repeated Measures	Group 1: O_1 – X – O_2 – O_3 – X – O_4 Group 2: O_1 – Y – O_2 – O_3 – Y – O_4	Same as 4, plus: What is the sustained effect of X on the outcome compared to Y?[e,f]

[a]For our purposes, all diagrams assume a nonrandom distribution of participants across treatments. We do not consider randomized controlled trials. [b]We use "X" to represent an intervention of interest and "Y" to represent an alternative treatment (often TAU). We also use "Y" to represent the absence of a treatment. [c]Participants are distributed across experimental conditions based on pretest cutoff scores, C, as represented by the vertical bar and subscript C. [d]Under certain conditions, regression discontinuity designs yield unbiased estimates by fully modeling selection processes. [e]In this case, the addition of design elements affords supplementary comparisons across groups that can be used to support/refute inferences about the relationship between treatment and outcome. [f]In this case, the addition of design elements serves as the basis for demonstrating reproducibility.

To expand on the designs presented in Table 1, consider the example of a researcher interested in evaluating student performance following a Peer-Led Team Learning[9] (PLTL) model of instruction. They could assess students on the outcome variable (such as a content test or course grade) and make a judgement about their performance. This kind of study would fall under the Design #1 framework. To examine how performance on the outcome measure changes following PLTL, researchers could plan for a one group pretest-posttest design (i.e., Design #2). Under this design, students are first assessed on a pretest, subsequently participate in a PLTL model of instruction, and then are assessed again using the same instrument under the same conditions. By comparing outcomes before and after PLTL instruction,

the researcher can account for students' prior knowledge about the topics on the chosen assessment.[10] If the goal was to estimate the relative impact of PLTL on the outcome variable compared to baseline instruction, hereinafter referred to as teaching-as-usual (TAU), researchers could plan for a posttest-only alternative treatment control group design (i.e., Design #3). This design requires the researcher to select a similar group of students to serve as a comparison. By contrasting outcomes across PLTL and TAU conditions, researchers can gather evidence to estimate the effect of PLTL, controlling for students' natural development over time (e.g., transitioning from high school to college, college experience, etc.). In the case of non-randomly distributed participants, selection bias is assumed without any information about differences in students' prior knowledge, demographics, or other performance-relevant characteristics. Biases that arise when comparing non-randomly distributed participants across conditions can threaten the validity of inferences made about treatment effects. We explore these validity considerations further in the next section.

It is hard to distinguish the impact of pre-existing factors from the impact of an intervention on the outcome variable without knowing anything about the students.[10] By adding a pre-treatment assessment to both conditions, researchers will arrive at Design #4, which is intended to corroborate the hypothesis that a treatment (e.g., PLTL) is superior to an alternative treatment (e.g., TAU) at producing gains on the outcome variable. In addition, pretest data can be used to estimate the size and direction of group differences on the outcome variable before the treatment is administered. The presence of group differences may be grounds for refuting observed treatment effects; however, any such conclusions are dependent on the way the data were collected and the choice of statistical methods.[10] Pretest data can also be used to estimate group differences between students who remain in the study and those who do not (i.e., attrition bias).

Design #5 is among the family of regression discontinuity (RD) designs that, under certain conditions, can provide unbiased estimates of treatment effects by fully modeling the selection of participants into treatment and control groups.[5] In an RD design, participants are distributed across conditions based on a pretest cutoff score (or an alternative assignment variable). Consider, for example, an experiment where researchers are interested in the effect of a supplemental instruction (SI) program on grade outcomes in general chemistry. To estimate the impact of the SI program on the outcome

compared to no SI program, researchers could assign participants to experimental conditions based on a chemistry content placement exam with a predetermined cutoff score. Students scoring below the cutoff would be assigned to the SI program while students scoring above the cutoff would not be eligible. Later, all students would be tested on the outcome variable and the analysis would test for a regression discontinuity. By regression discontinuity, Shadish et al. referred to a "point-specific displacement (or discontinuity) of the regression line" that occurs exactly at the cutoff when plotting posttest scores against pretest scores (p. 212).[5] For more details on the logic and mechanics underlying RD designs, we refer readers to Ref. 5.

Designs #6 and #7 build on Designs #3 and #4, respectively, in that they assess post-treatment outcomes at multiple points in time using the same test. Designs like these are useful for demonstrating how an outcome varies with the introduction and removal of a treatment. As an example of Design #6, consider the evaluation of PLTL on course grades in the target course as well as grades in subsequent chemistry courses under the TAU model. To estimate longer-term effects of PLTL, the researcher would compare subsequent course grades of students initially receiving PLTL (i.e., O_2 for Group 1) to those initially receiving TAU in the same course (i.e., O_2 for Group 2). Designs #8 and #9 are related to Designs #3 and #4, respectively, in that they reintroduce the treatment and a final measurement of the outcome. These designs can support or refute how treatments and outcomes covary over time, and it is one way to demonstrate the reproducibility of an instructional intervention on a given outcome.

Choosing a framework that is most appropriate for a study will depend on the nature of the intervention, target population, institutional setting, and the outcome of interest. As there will always be extensive threats in any non-randomized controlled trial, researchers and practitioners intending to estimate treatment effects and publish their findings in the *Journal* can best contribute to the CER community by carefully acknowledging the measurement threats inherent to their experimental design. Next, we discuss four types of validity evidence that can be gathered to support or refute inferences about treatments and their effects within causal studies.

Validity Typology

All experimental work involves making inferences about constructs and their relationships from the samples collected within a study.[2,5,11] Important to making valid inferences about causal effects is the

clear explication of subjects, settings, treatments, and outcomes. Researchers must then take care in conducting experiments that are aligned with relevant and useful constructs and to assess potential threats due to any misalignment. As Shadish et al.[5] discussed, researchers conducting experiments are in the business of answering the following questions: How thorough is the characterization of subjects, treatments, settings, and their relationships to measured outcomes? Are treatments appropriately distributed across subjects to make inferences about treatments and their effects? How appropriate are the quantitative methods for analyzing the given data? To what extent can results be generalized across varied subjects, treatments, observations, and settings?

Following the work of Shadish et al.,[5] Table 2 lists the following four types of validity evidence that can be gathered to support or refute inferences about treatments and their effects: internal, construct, statistical conclusion, and external validity. For each validity type, Table 2 lists specific threats with labels, their descriptions, and potential evidence that could be gathered to address the threat. When referring to each type of validity threat later in the review, we signify the threat using the labels in Table 2, (e.g., Threat A1). What follows is a brief discussion on the four types of validity evidence relevant to causal studies in CER.

Internal Validity Evidence

Internal validity evidence can be gathered to support or refute inferences about the causal relationship between treatment and outcome. An important consideration for any quasi-experimental work in CER is how to distinguish observed treatment effects from students' natural development over time (i.e., maturation). Contrasting outcomes in the experimental condition to a comparison group receiving an alternative treatment or the absence of the treatment can help researchers rule out threats due to maturation.[5] Other threats to the internal validity of inferences include: differences in participant characteristics (Threat A2), external events happening parallel to the study that could also explain the observed effects (Threat A3), loss of participants (Threat A4), or a change in the instrument used to measure an outcome (Threat A5). Suggested evidence that can be gathered to address each threat is also listed in Table 2; however, this is not an exhaustive list. There are multiple ways in which to gather evidence for each source for validity threat depending on the experimental design and the context of the investigation.

Table 2. Matrix of Validity Considerations for Quasi-Experimental Designs with Selected Threats for Each Validity Type and Potential Evidence To Address the Threat

Validity Type	Threats[a] to Validity Type[b]	Label	Source of Threat	Potential Evidence to Address Threat[c]
Internal Validity	Maturation	A1	Naturally occurring changes over time could account for observed treatment effects	Comparison group(s)
	Selection	A2	Population differences confound treatment effects	Pretests, proxy pretests
	History	A3	Event(s) occurring concurrently with treatment could cause the observed effect	Sample similar groups; similar assessment protocols
	Attrition	A4	Systematic loss of respondents can produce artificial treatment effects	Qualitative or quantitative assessment of plausibility
	Instrumentation	A5	Difference in measurement/ instrumentation across groups or over time	Use the same/equivalent tests to compare across groups
Construct Validity	Inadequate explication of constructs	B1	Limited grounding of measurement variables in theoretical constructs	Multiple definitions, perspectives, and/or frameworks inform measurement and interpretation
	Construct confounding	B2	The study operation is multifaceted and not the pure representation of the intended construct	Care in initial explication of persons, settings, treatments, and outcomes within context of theory; poststudy reformulation of construct(s) where appropriate
	Confounding constructs with levels of constructs	B3	The study operation is one of many levels of construct	Use several levels or doses of the treatment
Statistical Conclusion Validity	Low statistical power	C1	Low power can lead to erroneous statistical conclusions	Power analysis; adequate sample sizes; measure and adjust for covariates
	Violated assumptions of statistical test	C2	Violations of statistical test assumptions can lead to either overestimating or underestimating treatment effects	Check statistical test assumptions within statistical software program; modify statistical test accordingly
	Unreliability of measures	C3	Measurement error threatens the accuracy of covariation between variables	Conduct psychometric testing of assessment instruments
	Inaccurate effect size estimation/no effect size estimation	C4	Some statistics systematically overestimate or underestimate the size of an effect	Report the range in which effect size estimates are accurate
External Validity	Interaction of the causal relationship with people	D1	Observed effects on one population might not hold for other populations	Test the instructional strategy across two or more settings with distinct student bodies
	Interaction of the causal relationship over treatment variations	D2	Observed effects might not hold when variants of the treatment are tested, or when the treatment is combined with other treatments	Test variants of the instructional strategy on outcomes, e.g., lecture, partially flipped, and fully flipped course
	Interaction of the causal relationship with outcomes	D3	Observed effects on one outcome may not hold if other outcomes are measured	Test the instructional strategy across two or more distinct outcomes, e.g., performance and self-efficacy
	Interaction of the causal relationship with settings	D4	Observed effects in one kind of setting may not hold in other settings	Test the instructional strategy across two or more institutional types, e.g., research-intensive, primarily undergraduate, and community college settings

[a]This is not a comprehensive list of validity threats. [b]See ref *5*. [c]There could be several options for gathering evidence to address validity threats, depending on the experimental design and context of the study.

195

Construct Validity Evidence

Three threats to the accurate characterization of treatments and the effects on measured outcomes are presented in Table 2, including inadequately defining constructs, conflating more than one construct, or not distinguishing levels of a construct.[5] For example, Chase et al.[12] described the level (or "dosage") of Process Oriented Guided Inquiry Learning (POGIL) in their treatment by contrasting their adaptation to the recommended implementation (Threat B2). As another example, Conway[13] evaluated different levels of guided inquiry in an organic-biochemistry course when they compared outcomes under three teaching conditions: non-guided inquiry, partial guided inquiry, and full guided inquiry (Threats B3 and D2). Carefully explicating constructs is important for accurately characterizing study results and assessing threats to measurement.

The formulation of construct validity presented in this study is an extension of the original use of construct validity in psychological testing, which is concerned with accurately inferring what is being measured by a given test.[11] We do not consider construct validity with respect to the psychometric data from assessments, and instead direct readers to Arjoon et al.[2] who explored this subject in more detail in the context of CER. However, we note that evidence from psychometric evaluations of assessments may provide support for the use of a particular instrument in a given study, and therefore support inferences made from statistical analyses using instrument scores. Such evidence would constitute statistical conclusion validity evidence, which is discussed next.

Statistical Conclusion Validity Evidence

Statistical conclusion validity evidence can be used to support or refute inferences about the covariation between treatments and outcomes. Table 2 lists four threats to inferences based on statistical analyses. Power, for example, is an attribute of any statistical test and can be thought of as the "ability of a test to detect relationships that exist in the population" (p. 45).[5] While, low power is often attributed to inadequate sample sizes, sample size is not the only source. The power of a statistical test can be limited when it does not adjust for covariates that would otherwise account for the observed variation in outcomes (Threat C1). For example, Theobald and Freeman[10] demonstrated how multiple linear regression techniques can distinguish the impact of student characteristics (such as SAT scores or previous course grades) on an outcome variable across groups. Other factors threatening the validity

of statistical inferences include, but are not limited to, violating assumptions of statistical tests (Threat C2), unreliably measuring latent variables (Threat C3), and inaccurately estimating effect sizes or withholding effect sizes altogether (Threat C4).[14]

External Validity Evidence

Evaluating cause-and-effect relationships over variation in subjects, settings, treatments, and outcomes may provide evidence for the external validity of inferences.[5] Table 2 describes four threats to inferences about the interaction of treatments with external sources of variation. For example, suppose an educational intervention was particularly effective at promoting performance outcomes for students in a general chemistry course for non-majors, but had relatively modest effects at promoting performance in courses for majors. Given differential effects of the treatment across settings and student populations, the results would not be generalizable (Threat D1). Such a result may prompt researchers to further explore why the intervention's effect varied across the different settings and for the different student populations. External validity evidence is also concerned with the consistency in experimental findings across variants of the treatment (Threat D2), across varied outcomes (Threat D3), and varied settings (Threat D4).

METHODS

Sampling

This study examined the experimental designs, performance metrics, comparison groups, analysis methods, and validity evidence reported in intervention studies published in the *Journal*. Initially, the scope was limited to CER articles based on the 2013 revised guidelines. The guidelines provided specific requirements for manuscripts submitted to the *Journal*, such as stating a research question, grounding analysis in a theoretical framework, and situating the research in the existing body of literature, among other requirements.[15,16] This starting point did not yield a sufficient sample from which themes could be drawn, therefore, the study period was extended by two years to January 2011. Thus, the resulting review is based on articles published between January 2011 and April 2017. While examining CER articles in the *Journal* narrowed the scope of this investigation, it limited our ability to generalize themes to the CER community more broadly. We further discuss this aspect of the review in the Limitations section.

Based on the scope outlined above, we read the title and abstract of every article in the *Journal* between January 2011 and April 2017. If the title and abstract indicated that the article fit within the scope of our review, we downloaded the article and stored it for later coding. Based on the following selection criteria, forty-five articles were initially downloaded and coded for inclusion in this study. First, a study had to report either between-group (e.g., treatment vs control) or within-group (e.g., pretest vs posttest) comparisons. Second, studies had to report measures of performance outcomes (e.g., chemistry content exam, course grades, etc.) because these are commonly used metrics for summative assessments of student learning in college chemistry courses. Third, the comparison of performance outcomes had to be accompanied by a quantitative result. Fourth, studies had to take place in college-level chemistry classroom settings. To ensure the identified articles met the criteria for inclusion, the first two authors jointly reviewed and discussed how the article did or did not meet the selection criteria. Based on an initial review of the 45 selected articles, the first two authors agreed that 24 met the selection criteria. A list of the 24 sampled articles is provided in Table S1.

Coding, Inter-Rater Agreement, and Data Displays

Based on the conceptual framework described in the previous section, the first author developed a reading protocol to provide consistency in interpretations across articles and reviewers. A copy of the protocol is provided in the Supporting Information. The first two authors independently coded each article using the reading protocol. Following several rounds of coding and discussion, raters established 100% agreement on study design elements, performance metrics, comparison groups, quantitative methods, and validity considerations.

Codes were organized into data matrices according to our conceptual framework. This data reduction technique facilitated our ability to see "what's there" and the relative prevalence of codes within each research question.[17] Table S2 displays the distribution of quasi-experimental frameworks coded in our sample of articles and Tables S3-S5 display the distributions of the most commonly coded performance metrics, types of comparisons, and the methods used for statistical inference, respectively.

RESULTS AND DISCUSSION

Question 1: What experimental designs have been used to study the effects of instructional interventions on student performance?

Sixteen articles reported posttest-only designs (i.e., Designs #1, #3, #6, and #8), making these the most prominent experimental design used to evaluate instructional interventions in the *Journal*. The remaining eight articles established pretest comparisons across experimental conditions (i.e., Designs #2, #4, #5, #7, and #9). Table S2 reports the distribution of experimental design codes for the sample of articles.

Posttest-Only Designs

Using a modified version of Design #1, Weaver and Sturtevant[18] assessed the status of performance outcomes in a first-semester general chemistry course for majors with TAU followed by outcomes in the subsequent second-semester general chemistry course with a flipped course design. The schematic for their design can be represented as $Y - O_1 - X - O_2$, where Y represents TAU in the first-semester course and X represents flipped instruction the second semester course. The authors evaluated performance using the American Chemical Society (ACS) Exams Institute Paired Question Exams for First Semester General Chemistry[19] (O_1) and the ACS Paired Question Exams for Second Semester General Chemistry[20] (O_2). The status of students following the flipped instructional design was contrasted to that following TAU based on how well students performed relative to the norming sample for each exam. Had the authors only compared O_2 to O_1 (i.e., no comparison to the norming sample), then the results would have been threatened by maturation, selection, history, and attrition biases. But if we assume students in the norming sample followed trajectories like $Y - O_1 - Y - O_2$, then effects due to students' natural development in college and in the subject of chemistry could be ruled out (Threat A1). We further discuss the use of ACS exams to measure student outcomes and validity considerations with respect to norming samples when we answer Questions 2 and 3.

The inclusion of comparison groups into study designs can help to decouple treatment effects from students' natural development over time. Under Design #3, four studies used comparison groups to evaluate the impact of instructional interventions in general chemistry courses, including: flipped instruction,[21,22] PLTL,[23] and POGIL.[12] Another six studies used comparison groups to evaluate instructional interventions in organic chemistry courses, including: POGIL,[12,24] classroom space

innovations,[25] and flipped instruction.[26–28] By adding multiple posttests to their experimental design (i.e., Design #6), Lewis[23] evaluated the downstream effects of PLTL in general chemistry courses on student persistence and grades in subsequent chemistry courses. Within Design #8, Hall et al. 2014[29] estimated the effect of the Science Advancement through Group Engagement (SAGE) program on exam performance and persistence in general and organic chemistry courses. By conducting multiple and sequential measurements, the authors were able to estimate the sustained effect of SAGE participation on performance across the curriculum.

Pretest-Posttest Designs

Pretest-posttest designs answer questions of the form: *What is the effect of X on the outcome compared to Y, given O_1?* Under Design #4, Hall et al. 2012[30] compared performance outcomes in a three-week summer intensive introductory chemistry course to outcomes from the same course taught during a normal academic semester. The authors administered and scored an in-house 24-item multiple-choice pretest to account for individual differences in chemistry content knowledge at the start of the course. Rath et al.[31] examined the impact of SI on average grades, pass rates, and graduation rates for participating students (relative to non-participants) across four chemistry courses using SAT and high school GPA data as their pre-treatment assessment. Both studies were coded as Design #4 because they utilized proxy pretests. By proxy pretests, we refer to the measurement of variables that are conceptually related and correlated with performance outcomes (i.e., O_2), but they are different metrics.[5]

Ideally, pretest-posttest designs use the same assessment instrument under the same conditions to ensure that the same constructs are being assessed before and after exposure to the treatment. However, for convenience or by necessity, authors used proxy pretests to assess pre-existing factors that may influence the outcome of the treatment. In fact, all pretest-posttest studies in our sample utilized proxy pretests because students are not typically pretested using exams or other outcomes typically reported in intervention studies, which we discuss more about when we answer Question 2. Shadish et al.[5] described two ways in which proxy pretests can enhance study designs in the absence of repeated pretest-posttest measures. First, researchers can gather evidence for selection bias by comparing groups initially on the pretest (Threat A2). To this end, proxies should adequately index group differences on the outcome measure to the extent that the proxy correlates with the posttest. Second, researchers can

gather evidence for attrition bias by conducting both between-group comparisons initially on the proxy pretest and those who remain in the study on the posttest, as well as within-group comparisons across the proxy pretest and posttest (Threat A4).

Building upon Design #4, RD designs model the selection process across groups by assigning participants to treatment conditions based on a cutoff score associated with a pretreatment assessment or an alternative assignment variable (Treat A2). Consequently, RD designs can yield unbiased estimates compared to non-randomized control trials under certain conditions.[5] Under Design #5, Shultz et al.[32] retrospectively evaluated the impact of a first-semester general chemistry course on subsequent course grades by comparing students who shared similar cutoff scores on a placement exam (i.e., a proxy pretest) but who differed in their compliance with course placement advise.

Unlike the sample of studies utilizing post-test only designs, no studies reported results from a removed treatment design with pretests (i.e., Design #7) or a repeated measures design with pretests (i.e., Design #9).

Question 2: What metrics have been used to measure student performance outcomes?

The most commonly coded performance metrics reported in the sample of articles were: instructor-authored exams (n=12), persistence/retention rates (n=11), ACS exams (n=9), letter grades (n=7), and course grade point average (n=4). Table S3 displays the distribution of articles across each code.

Instructor-Authored Exams

Instructor-authored exams were reported in half of the sampled articles, making exam scores the most frequently coded measure of student performance. Both midterm and final exams are included in this category. The ways in which researchers characterized exam properties varied. Some authors reported exam content (e.g., topics), while others reported Classical Test Theory (CTT) and/or Item-Response Theory (IRT) metrics, and some did not include exam characteristics in their report. In the context of performance measurement in intervention studies, establishing and reporting psychometric properties of tests is important for obtaining consistent scores for the same metric across treatment conditions. In terms of the validity typology for intervention studies, gathering and reporting reliability evidence for test properties reduces bias in estimates of treatment effects. For example, Ryan and Reid[33]

evaluated difficulty and discrimination indices as well as person- and item-reliability parameter estimates from Rasch analysis to judge the consistency of item functioning across groups in their sample. Evidence from their CTT and IRT analyses was used to support comparisons of exam scores across comparison groups (Threat C3).

Persistence/Retention Rates

Persistence or retention rates were the second most commonly coded performance metric in the sample of articles (n=11). This outcome included DFW (D or F grades or course withdraw) rates, FW rates, withdraw rates, and pass/fail rates. These measures typically complimented other measures of performance, such as exam scores or course grade-point averages, thereby providing a second measure of the treatment's impact. For example, by showing consistently superior effects of a "web-enhanced" first-semester general chemistry course on both course grades and withdrawal rates relative to TAU, Amaral et al.[21] gathered external validity evidence to generalize treatment effects across multiple outcomes (Threat D3).

ACS Exams

Exams developed and distributed by the ACS Examinations Institute (i.e., ACS Exams) were the third most frequently reported performance metric coded in the sample of articles. Nine studies reported performance outcomes on ACS exams, five of which utilized the exams as the sole measure of performance. Some researchers specifically cited using ACS exams for the additional comparison they afforded. As Weaver and Sturtevant[18] stated, "[u]sing these [ACS] exams allowed us to measure student performance against national norms from data made available for each exam by the ACS Exams Institute" (p. 1440).

In a study on POGIL in an organic chemistry course using the 2002 ACS Standardized Organic Exam and 2008 exam forms, Hein[24] compared outcomes both locally across comparison groups as well as to the original norming sample. The local comparison indicated a statistically significant treatment effect when controlling for students incoming college GPA. Upon binning students into three groups based on percentile rankings, the authors reported the number of students in the lowest rank decreased over each year POGIL was used. Furthermore, increased proportions of students scoring in the middle and high ranks was evident for the treated groups. By comparing outcomes both locally and to the norming

sample, the authors provided evidence for a positive treatment effect; POGIL students ranked higher in both median and mean national percentile ranking compared to comparison group.[24] However, the statistical inferences made were threatened by a change in exam from the 2002 Organic Exam form to the 2008 form. (Citations are not provided as these exam forms are no longer available.) When a variable is measured inconsistently, like when two different tests are used to measure the same construct, estimates to treatment effects may be biased without the appropriate psychometric data to support the comparison (Threat A5).

Course Grades

The fourth and fifth most frequently reported performance metrics were related to course grades. Course letter grades were reported in seven articles and course grade-point averages were reported in four articles in our sample. For example, Rath et al.[31] compared course grade point averages across students receiving SI and students not receiving SI. In addition to contrasting final exam scores, Conway[13] compared letter grade distributions across three teaching conditions: full guided inquiry, partial guided inquiry, and non-guided inquiry or TAU. In this example, course grades served as a complementary metric to report in combination with exam scores and provided additional information about student achievement that might not be captured in high-stakes exams (Threat D3).

Additional Metrics

We observed several other metrics in our sample that were less commonly used across studies. These included: retention rates in subsequent courses, change in probability of earning a passing grade, graduation rate, GPA in subsequent courses, total points earned after the first exam, percent of students who took final exam, students who received a final grade, problem set grades, and quiz grades. While some metrics were less common than others in our sample of articles, they can still be useful when they are well-aligned with the goals of the investigation.

Question 3: What types of groups have been used to compare performance outcomes?

The most common comparison groups reported in the sample of articles were: concurrent (n=12) and historical (n=9) groups, comparisons across successive assessments (n=12), comparisons across student interest groups (n=7), comparisons of ACS exam performance with nationally normed sample

(n=3), and comparisons across multiple outcomes (n=3). Table S4 reports the distribution of articles across the different comparison group categories.

Concurrent Comparison Groups

Twelve studies concurrently compared performance outcomes across treatment and comparison groups. Chase et al.[12] examined an adaptation of POGIL in three out of nine discussion sections in a general chemistry course and three of the five discussion sections in an organic chemistry course. The remaining discussion sections were TAU comparison groups. Similarly, Lewis 2011[34] examined the implementation of PLTL in eight sections of a first-semester general chemistry course in comparison to twenty-one sections of the same course under TAU over a span of four semesters. Concurrent comparison group designs are sometimes threatened by instructor and/or time-of-day effects (Threat A2). To assess the extent of such bias, Lewis 2011 evaluated outcomes across PLTL and TAU sections with the same instructor. The author found a negligible section effect bias (Threat A2); however, the method of statistical inference chosen for the analysis was not capable of accounting for both treatment and section-level effects (and pre-treatment assessment data) within one statistical model (Threat C1). We further discuss statistical methods for comparing performance outcomes when we answer Question 4.

Historical Comparison Groups

Nine studies compared performance outcomes across treatment and historical comparison groups. Amaral et al.[21] compared performance outcomes in a web-enhanced delivery of a general chemistry course to outcomes in prior years under TAU instruction. Mooring et al.[27] compared performance outcomes in a flipped first-semester organic chemistry course to TAU in the five-year period preceding the change. Crimmins and Midkiff[35] compared performance outcomes in their organic chemistry curriculum with a "high structure active learning" pedagogy in the 2013-14 academic year to outcomes in the same course taught by the same instructor in the 2002-03 academic year under a TAU model. In this study, the authors utilized propensity score matching based on SAT scores and demographic data to establish common overlap across treatment and comparison groups. Under strict assumptions, propensity score matching yields an unbiased estimate of the average treatment effect.[5,36] This is one

445 way in which historical comparison groups with large sample sizes can support powerful statistical inferences.

Historical comparisons are not without their limitations. In contrast to concurrent comparison groups, historical comparisons are threatened by an unshared history, meaning events occurring external to the study during the two time periods differ in ways that could be confused with treatment
450 effects.[5] Crimmins and Midkiff[35] discussed this threat when they stated their treatment condition was confounded by advances in technology and the external resources available to students via the Internet due to the 11-year separation between treatment and comparison group samples (Threat A3).

Comparing Groups on Successive Assessments

Twelve studies compared performance on successive assessments. By successive assessments, we
455 refer to the sequential assessment of outcomes within and/or across groups. For example, Lewis 2014[23] demonstrated the rise and fall of course grades in the presence and absence of PLTL across subsequent chemistry courses. The author reported the PLTL model in a first-semester general chemistry course "had a small effect on the grade distribution in the target class of GC1, which steadily drops off in subsequent classes" operating under the TAU model (p. 2041). In a second comparison, Lewis 2014[23]
460 examined the effect of PLTL on course grades in a second-semester general chemistry course and TAU in successive chemistry courses. Again, a rise and fall in course grades was observed. Overall, Lewis 2014[23] concluded that PLTL had a "statistically significant and small effect on enrollment in the class that directly succeeds the target class," but that "[a]s students progress through the curriculum the impact of the reform on course enrollment declines until it becomes attributable to chance" (p. 2042).

465 *Comparing Student Subpopulations*

Seven studies compared performance outcomes based on demographic variables, such as binary gender, race and ethnicity, socioeconomic status, and/or family education background. In their retrospective analysis, Shultz et al.[32] compared differential impacts of taking general chemistry on progression to subsequent chemistry courses for men and women. Rath et al.[31] compared the impact of
470 an SI program for general and organic chemistry courses for historically underrepresented racial and ethnic minorities in STEM and well represented students. Lewis 2011[34] compared the differential impact of PLTL in a first-semester general chemistry course on pass rates by binary gender and racial and

ethnic minority status separately. Testing the robustness of treatment effects across student subpopulations may confirm or refute inferences that an educational intervention had an impact. For example, the efficacy of that intervention would be challenged if it had inequitable impacts across different student groups. Comparing across student subpopulations is one method for gathering external validity evidence to support or refute inferences about treatment effects (Threat D1).

Normed Comparisons

Three studies compared performance outcomes on ACS exams to the original norming sample. One benefit of using ACS exams as a performance metric is the opportunity to compare local outcomes to the norming sample. In general, comparisons ACS exam scores made with norming samples is informative because it affords an additional comparison that is not typically available with other kinds assessments in the chemistry education community. However, comparing ACS exam scores to the original norming sample may not be strongly indicative of how the treated group would have performed without the treatment, and care should be taken to understand the similarities and differences between local and original norming samples. Without knowing the identities and characteristics (e.g., size, selectivity, region, etc.) of the institutions represented in the norming sample, population differences confound the treatment effect (Threat A2). History threats also bias the comparison because the norming sample was collected before the data from the treated sample (Threat A3).

Comparing Groups on Different Outcome Measures

Three studies in the sample of articles examined the impact of instructional interventions on both affective- and performance-based outcomes. The National Research Council[37] stated that the affective domain consists of psychological constructs including, but not limited to, anxiety, fear, motivation, attitudes, and self-efficacy, as well as sociocultural factors including values, social pressures, and stereotypes. Using the ASCIv2,[38] Mooring et al.[27] compared pre/post responses on the emotional satisfaction and intellectual accessibility scales for students in a flipped first-semester organic chemistry course and students in concurrent sections taught under the TAU model. Separately, the authors compared final grades and withdrawal rates in the experimental course to the concurrent TAU section as well as a historical TAU comparison group. Similarly, Chase et al.[12] evaluated the impact of an adaptation of POGIL on performance, attitudes towards chemistry (using the ASCI[39]), and self-efficacy

(as measured by the CAEQ[40]), separately, for both experimental groups. Examining the variation in outcomes across several measures is one way to gather external validity evidence to support or refute inferences about treatment effect estimates (Threat D3). It is possible that an effect found on one outcome might not hold true for another. However, researchers and practitioners should take care in aligning assessments with the underlying theory of the instructional intervention when collecting cognitive and affective data (Threat B1), as theory plays an essential role in the selection of variables in experimental research by specifying the relevant constructs to be measured.[5,41] We discuss this point further in the Recommendations section.

Question 4: What quantitative methods have been used to analyze the comparison data?

Authors utilized a range of statistical methods for quantifying relationships between people, treatments, outcomes, and settings. Below we discuss how descriptive statistics, univariate statistics, and multivariate statistical methods supported researchers and practitioners in gathering evidence for estimating treatment effects. Table S5 reports the distribution of articles across each analysis type.

Descriptive Statistics

Descriptive statistics provided efficient summaries of performance outcomes. Nearly every article in our sample (N=22) displayed data using tables and/or graphs. Histograms displayed letter grade distributions. Boxplots displayed exam grade distributions. Bar charts were used to visually represent average scores across course sections, academic quarters, and years. Furthermore, these kind of data displays provided readers with visual contrasts of performance outcomes across treatment and comparison groups. Tables also provided efficient summaries of performance outcomes by displaying raw counts, mean scores, standard deviations, percentages, etc., depending on the outcome. Reporting outcomes in tables and graphs is useful for summarizing outcomes and can support further inferencing using univariate and/or multivariate statistical methods.

Univariate Statistics

Researchers and practitioners interested in estimating the effects of instructional interventions generally asked questions of the form: *What is the effect of an instructional intervention on the outcome variable compared to TAU?* One approach to answering such a question is to construct falsifiable null

and alternative hypotheses about the relationship between teaching model and outcomes. Under the traditional null hypothesis significance testing framework, the alternative hypothesis would stipulate a model where the treatment condition covaries with outcomes while the null hypothesis stipulates no relationship between treatment and outcome. Once the null and alternative hypotheses are constructed and the data is collected, inferential statistical techniques can support researchers in gathering evidence to evaluate the tenability of the null hypothesis. The alternative hypothesis is never confirmed. Instead, probability statements about the likelihood of observing the data given the null hypothesis is true in the population are constructed.[42]

Nineteen of the 24 studies utilized univariate statistical analyses—e.g., analysis of variance (ANOVA), *t*-tests, and non-parametric equivalents—to test the tenability of the null hypothesis. For example, Lewis 2011[34] sampled 29 sections of a first-semester general chemistry course to test the hypothesis that PLTL had a superior effect on final exam performance relative to TAU. While no discernable difference in exam performance was observed based on a *t*-test, a larger proportion of students in the PLTL sections took the final and received passing grades. Following up on that study, Lewis 2014[23] compared the proportions of students progressing to advanced courses using chi-square test statistics and corresponding effect size estimates (Threat C4). Robert et al.[43] compared average exam scores between four sections of general chemistry under a flipped PLTL model and four sections under a TAU model and observed large effects in favor of flipped PLTL. However, the authors cautioned readers about potentially inflated effect size estimates since the data were analyzed at the class level rather than at the student level (Threat C4).

Multivariate Statistical Analyses

Researchers and practitioners were also interested in estimating the effects of instructional interventions beyond individual differences in students' prior knowledge and/or academic attainment. These studies generally asked questions of the form: *What is the effect of an instructional intervention on the outcome compared to TAU controlling for performance-related covariates?* Selection bias that arises due to non-randomly distributed participants can be addressed when covariates are included in statistical analyses (Threat A2 and C1).

Ten studies applied multivariate statistical methods for analyzing the relations between a single outcome measure and multiple explanatory variables, usually indicators of students' prior chemistry content knowledge and/or academic attainment. In some studies, authors included demographic characteristics as additional covariates. Six of the 10 studies analyzed data using multiple regression analyses. For example, Eichler and Peeples[44] applied multiple linear regression to contrast the effects of different online homework platforms on final exam performance while adjusting for SAT scores, high school GPA, and demographic variables (Threats A2 and C1). Shultz et al.[32] adjusted their linear regression model for chemistry placement exam scores when they tested the effect of "bypassing" first-semester general chemistry on performance outcomes in subsequent courses compared to students who enrolled and completed first-semester general chemistry (Threats A2 and C1). The remaining four studies utilized analysis of covariance (ANCOVA) to estimate treatment effects while simultaneously accounting for individual differences in prior knowledge or academic attainment. For example, Cook et al.[45] compared total exam points in a general chemistry course between treatment and control groups while adjusting for performance on the first exam (Threats A2 and C1). Similarly, Hall et al.[30] compared total points earned across treatment and control groups while adjusting for a chemistry content pretest, cumulative college GPA, and ACT scores (Threats A2 and C1). The authors addressed the threat of low statistical power to their ANCOVA adjusted for ACT scores due to insufficient sampling (Threat C1) by conducting the analysis twice with and without the ACT data and comparing the results.

No studies in the sample of articles used multivariate statistical modeling with multiple outcomes (e.g., multivariate analysis of variance), or both multiple predictors and multiple outcomes (e.g., Structural Equation Modeling).[46] We discuss how these statistical methods can play a role in advancing future intervention studies in CER in the Recommendations section.

RECOMMENDATIONS FOR FUTURE CAUSAL STUDIES

We reviewed the experimental designs, metrics, comparison groups, and quantitative methods used to estimate the impacts of instructional interventions on performance outcomes in 24 chemistry education research (CER) articles published in the *Journal*. We also examined the kind of validity evidence used to support or refute inferences made about treatments and the observed effects. Based on our findings, we recommend three ways in which causal studies in CER can be advanced moving

forward. We use our conceptual framework based on quasi-experimental design and the validity typology[5] to guide our recommendations. We hope that consideration of any one recommendation will aid in the future planning of intervention studies in CER.

Recommendation #1: Collection of pretreatment assessment data

The posttest-only alternative treatment control group design was the most prominent design used to evaluate instructional interventions in our sample of articles. Using this design, authors gathered evidence to contrast the impacts of different teaching models on an outcome. However, given that treatments were not randomly distributed across participants, selection bias was assumed without any information about differences in students' prior abilities and other performance-related characteristics prior to the intervention. To overcome this threat, future studies would benefit from collecting pretreatment assessment data and incorporating into their analysis.

With pretreatment assessment data, researchers and practitioners can assess the impacts of different teaching models on an outcome while considering population differences on a pretest measure. For example, by assessing students' prior academic attainment (e.g., SAT or ACT scores, grades in prerequisite courses), content knowledge (e.g., diagnostic tests), or other variables specified by theory (e.g., motivation, self-efficacy, etc.), researchers and practitioners can identify population differences between groups, which can then be accounted for using multivariate statistics.[10] Attrition bias can also be assessed with pretest data. If students who score lowest on a diagnostic pretest withdraw from the treatment condition at disproportionate rates, then treatment effects are confounded by attrition bias. We recommend that researchers and practitioners examine attrition descriptively by reporting attrition rates for every experimental condition, as well as the similarities and differences in pretest scores for students who completed the study and those who did not.

While the addition of pretests can create a more complicated approach to intervention studies published in the *Journal*, the added complexity affords researchers and practitioners the opportunity to rule out validity threats that would otherwise be unavailable in post-test only experimental designs.

Recommendation #2: Use of theory to guide variable selection

How should researchers and practitioners go about choosing which pre- and post-treatment assessment data to collect? Theory plays an essential role in the selection of variables in experimental

research by specifying the relevant constructs to be measured.[5,41] Looking back to the sample of articles utilizing pretest-posttest designs, authors were generally guided by the theory that pre-existing cognitive factors (e.g., students' prior chemistry content knowledge and/or prior academic attainment) influenced performance outcomes in chemistry courses. While it is intuitive to think of a student's performance in general and organic chemistry as a function of their prior knowledge and past performances in education, teaching interventions are typically grounded in psychological and sociocultural theories of how people learn. For example, Cook et al.[45] hypothesized that training students in metacognitive learning strategies would lead to more sophisticated study habits, which in turn would promote higher scores on course examinations. As another example, Crimmins and Midkiff[35] hypothesized that high-structure active learning would lead to a stronger sense of classroom community and social integration, which in turn would lead to improved performance on final exam scores and pass rates. Examples like these suggest that the chemistry education community is interested in further understanding the complex relations between instructional strategies and student outcomes within broader psychological and sociocultural theories. Therefore, we recommend researchers and practitioners to thoroughly explicate the theoretical constructs underpinning their instructional intervention and to use those constructs to guide the selection of pre- and post-treatment assessments. By accepting this recommendation, researchers and practitioners will use theory to guide the design and implementation of instructional interventions, as well as the selection or creation of assessment instruments.

Recommendation #3: Use of theory to guide data analysis

It was common for authors to describe the theoretical underpinnings of instructional interventions, but the measurement of theoretically specified constructs and mediated processes was a less established practice in the sample of articles. Once variables are identified (i.e., Recommendation #2) and assessment data collected (i.e., Recommendation #1), we recommend that researchers and practitioners conduct statistical analyses capable of evaluating the relations among variables postulated by the guiding conceptual framework. Structural equation modeling (SEM), multiple regression, and multivariate analysis of covariance (i.e., MANCOVA) are a few examples of statistical methods that are capable of evaluating multiple outcomes and explanatory variables under one conceptual framework.

(The mathematical formulations of these statistical methods are beyond the scope of this article, so we direct interested readers to more comprehensive introductions in other sources.[10,47–52])

By accepting these recommendations, researchers and practitioners who conduct causal studies will: (i) build and evaluate complex models of teaching and learning in college-level chemistry classrooms situated in broader psychological and sociocultural theories, (ii) measure all the constructs specified in the theory, (iii) gather inferential evidence for the extent to which relations between constructs were observed in the data, and (iv) interpret the results within the context of a conceptual framework. The point is to use theory to guide the design, implementation, and evaluation of teaching and student outcomes. Not only will theory help researchers and practitioners to make sense of data collected from causal studies, we believe an intensified use of theory will support the community in building more comprehensive understandings of teaching models and their impacts on student outcomes.

LIMITATIONS

There are five limitations to this review that we present. First, the findings and recommendations are mainly applicable to chemistry education researchers and practitioners interested in publishing future intervention studies in the *Journal*. We do not claim that the same patterns of experimental designs, performance metrics, comparison groups, and analysis techniques will be found among collections of articles from other journals during the same time frame. However, we do suggest the lessons learned from this review can be used as a resource for researchers and practitioners interested in conducting future causal studies in CER.

Second, designs #1-9 in Table 1 represent a limited sample of quasi-experimental designs for conducting causal studies of instructional interventions in CER. Although the nine designs were comprehensive for our sample of articles, Table 1 not an exhaustive list. Therefore, chemistry education researchers and practitioners interested in conducting causal intervention studies should familiarize themselves with other design frameworks, such as interrupted time-series designs, as well as the myriad ways of strengthening designs, such as pattern matching, using multiple or different outcomes, and using high quality assessments.[5]

Third, the validity typology presented in this study represents a limited sample of validity threats addressed by Shadish et al. We made a concerted attempt to include a comprehensive set of validity

threats that were relevant to our sample of articles. However, we acknowledge this was not an exhaustive list of threats to each validity type. For example, we did not discuss how regression to the mean and testing effects can threaten the internal validity of treatment effect estimates.[5] Therefore, interested researchers and practitioners would benefit from familiarizing themselves with the entire framework presented in Ref. 5 and readings therein.

Fourth, this review focused exclusively on intervention studies that evaluated performance-based outcomes. Intervention studies that exclusively examined student affect or outcomes on concept inventories were not represented in our sample of articles. Therefore, it is imperative that this work be interpreted only within the context of quantitative investigations of performance outcomes across instructional interventions.

Finally, this review did not paint the whole picture of intervention studies in the *Journal*. Focusing our review on quasi-experimental studies meant that we did not examine studies under qualitative paradigms of inquiry. Consequently, the unique contributions that qualitative research methods have for causal studies in CER were not addressed in this review.

CONCLUSION

We examined 24 articles published in the *Journal* from January 2011 to April 2017 that reported quantitative investigations of student performance outcomes across varied instructional strategies in college-level chemistry courses. The findings serve as an indicator of the state of causal studies of instructional strategies published in the *Journal*. Several quasi-experimental designs were utilized by chemistry education researchers and practitioners to support their investigations, with the most common being posttest-only designs. Student performance outcomes were most commonly contrasted using instructor-authored exams across concurrent comparison groups. Most studies analyzed data using univariate statistical methods. In some instances, inferences about the effect of an intervention were strengthened by addressing threats to at least one validity consideration. We would like to emphasize that it is not possible for a single study to address all, or even most, of the threats that can bias comparisons made within non-randomized controlled trials. Therefore, researchers and practitioners should strategically prioritize their effort at combatting validity threats. At the very least,

threats to treatment effect estimates should be reported on in detail so that future researchers and practitioners can more effectively target known gaps in prior studies.

As a result of this review, we recommended a closer alignment between theory and measurement in future casual intervention studies. We acknowledge that the inclusion of theoretical considerations at all stages of the intervention (i.e., design, data collection, and analysis) creates a more complicated approach to causal studies in CER, conceptually, statistically, and logistically. To ease this responsibility, evidence for the efficacy of an instructional model should accumulate over time using diverse experimental designs, data collection and analysis procedures, and settings. No single study can evaluate every variable and every theoretical relationship underlying an instructional model. Therefore, intervention studies should progressively build upon one another with evidence for the relationships between people, treatments, outcomes, and settings accumulating over time. We hope the information compiled in this review can encourage and facilitate this effort within the chemistry education community.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available on the ACS Publications website at DOI:

10.1021/acs.jchemed.XXXXXXX. [ACS will fill this in.]

List of Sampled Articles (Table S1); Reading Protocol; Data Matrices (Tables S2-S5); and Code Descriptions (PDF, DOCX)

AUTHOR INFORMATION

Corresponding Author

*E-mail: jbarbera@pdx.edu

ACKNOWLEDGMENTS

We would like to acknowledge the contributions to the chemistry education community made by the authors in our sample of articles. Conducting a course- or program-level intervention is no easy task and we would be remiss not to acknowledge the impressive effort put forth when refining one's teaching practice and disseminating the outcomes. MRM would like to thank Drs. Colleen Craig, Regis Komperda, Carly Schnoebelen, and Cynthia Stanich for feedback on earlier versions of the

manuscript. JB acknowledges that this material is based upon work supported by the National Science Foundation under Grant No. (1611519). Finally, the authors would like to thank the anonymous reviewers for their feedback.

REFERENCES

(1) Cooper, M. M.; Stowe, R. L. Chemistry Education Research—From Personal Empiricism to Evidence, Theory, and Informed Practice. Chem. Rev. 2018, 118, 6053–6087.

(2) Arjoon, J. A.; Xu, X.; Lewis, J. E. Understanding the State of the Art for Measurement in Chemistry Education Research: Examining the Psychometric Evidence. J. Chem. Educ. 2013, 90, 536–545.

(3) Freeman, S.; Eddy, S. L.; McDonough, M.; Smith, M. K.; Okoroafor, N.; Jordt, H.; Wenderoth, M. P. Active Learning Increases Student Performance in Science, Engineering, and Mathematics. Proc. Nat. Acad. Sci. 2014, 111, 8410–8415.

(4) Stains, M.; Harshman, J.; Barker, M. K.; Chasteen, S. V.; Cole, R.; DeChenne-Peters, S. E.; Eagan, M. K.; Esson, J. M.; Knight, J. K.; Laski, F. A.; et al. Anatomy of STEM Teaching in North American Universities. Science. 2018, 359, 1468-1470.

(5) Shadish, W. R.; Cook, T. D.; Campbell, D. T. Experimental and Quasi-Experimental Designs for Generalized Causal Inference; Houghton Mifflin: Boston, MA, 2002.

(6) Campbell, D. T.; Stanley, J. C. Experimental and Quasi-Experimental Designs for Research, 2nd ed.; Houghton Mifflin: Boston, MA, 1967.

(7) Campbell, D. T. Factors Relevant to the Validity of Experiments in Social Settings. Psychol. Bullet. 1957, 54, 297–312.

(8) Cook, T. D.; Campbell, D. T. Quasi-Experimentation: Design and Analysis Issues for Field Settings; Houghton Mifflin: Boston, MA, 1979.

(9) Wilson, S. B.; Varma-Nelson, P. Small Groups, Significant Impact: A Review of Peer-Led Team Learning Research with Implications for STEM Education Researchers and Faculty. J. Chem. Educ. 2016, 93, 1686–1702.

(10) Theobald, R.; Freeman, S. Is It the Intervention or the Students? Using Linear Regression to Control for Student Characteristics in Undergraduate STEM Education Research. CBE—Life Sci. Educ. 2014, 13, 41–48.

(11) Cronbach, L. J.; Meehl, P. E. Construct Validity in Psychological Tests. Psychol. Bullet. 1955, 52, 281–302.

(12) Chase, A.; Pakhira, D.; Stains, M. Implementing Process-Oriented, Guided-Inquiry Learning for the First Time: Adaptations and Short-Term Impacts on Students' Attitude and Performance. J. Chem. Educ. 2013, 90, 409–416.

(13) Conway, C. J. Effects of Guided Inquiry versus Lecture Instruction on Final Grade Distribution in a One-Semester Organic and Biochemistry Course. J. Chem. Educ. 2014, 91, 480–483.

(14) Wasserstein, R. L.; Lazar, N. A. The ASA's Statement on p-Values: Context, Process, and Purpose. Amer. Stat. 2016, 70, 129–133.

(15) ACS Publications. Content Requirements for Chemical Education Research Manuscripts http://pubs.acs.org/paragonplus/submission/jceda8/jceda8_CER_Guide.pdf (accessed Jan 2019).

(16) Towns, M. H. New Guidelines for Chemistry Education Research Manuscripts and Future Directions of the Field. J. Chem. Educ. 2013, 90, 1107–1108.

(17) Miles, M. B.; Huberman, A. M. Qualitative Data Analysis: A Sourcebook of New Methods; Sage Publications: Thousand Oaks, CA, 1984.

(18) Weaver, G. C.; Sturtevant, H. G. Design, Implementation, and Evaluation of a Flipped Format General Chemistry Course. J. Chem. Educ. 2015, 92, 1437–1448.

(19) ACS Examinations Institute. First Term General Chemistry Paired Questions Exam, 2005. Division of Chemical Education Examinations Institute, American Chemical Society: Washington, DC.

(20) ACS Examinations Institute. General Chemistry 2nd Term Paired Question Exam, 2007. Division of Chemical Education Examinations Institute, American Chemical Society: Washington, DC.

(21) Amaral, K. E.; Shank, J. D.; Shibley, I. A.; Shibley, L. R. Web-Enhanced General Chemistry Increases Student Completion Rates, Success, and Satisfaction. J. Chem. Educ. 2013, 90, 296–302.

(22) Hibbard, L.; Sung, S.; Wells, B. Examining the Effectiveness of a Semi-Self-Paced Flipped Learning Format in a College General Chemistry Sequence. J. Chem. Educ. 2016, 93, 24–30.

(23) Lewis, S. E. Investigating the Longitudinal Impact of a Successful Reform in General Chemistry on Student Enrollment and Academic Performance. J. Chem. Educ. 2014, 91, 2037–2044.

(24) Hein, S. M. Positive Impacts Using POGIL in Organic Chemistry. J. Chem. Educ. 2012, 89, 860–864.

(25) Muthyala, R. S.; Wei, W. Does Space Matter? Impact of Classroom Space on Student Learning in an Organic-First Curriculum. J. Chem. Educ. 2013, 90, 45–50.

(26) Christiansen, M. A.; Lambert, A. M.; Nadelson, L. S.; Dupree, K. M.; Kingsford, T. A. In-Class Versus At-Home Quizzes: Which Is Better? A Flipped Learning Study in a Two-Site Synchronously Broadcast Organic Chemistry Course. J. Chem. Educ. 2017, 94, 157–163.

(27) Mooring, S. R.; Mitchell, C. E.; Burrows, N. L. Evaluation of a Flipped, Large-Enrollment Organic Chemistry Course on Student Attitude and Achievement. J. Chem. Educ. 2016, 93, 1972–1983.

(28) Shattuck, J. C. A Parallel Controlled Study of the Effectiveness of a Partially Flipped Organic Chemistry Course on Student Performance, Perceptions, and Course Completion. J. Chem. Educ. 2016, 93, 1984–1992.

(29) Hall, D. M.; Curtin-Soydan, A. J.; Canelas, D. A. The Science Advancement through Group Engagement Program: Leveling the Playing Field and Increasing Retention in Science. J. Chem. Educ. 2014, 91, 37–47.

(30) Hall, M. V.; Wilson, L. A.; Sanger, M. J. Student Success in Intensive versus Traditional Introductory College Chemistry Courses. J. Chem. Educ. 2012, 89, 1109–1113.

(31) Rath, K. A.; Peterfreund, A.; Bayliss, F.; Runquist, E.; Simonis, U. Impact of Supplemental Instruction in Entry-Level Chemistry Courses at a Midsized Public University. J. Chem. Educ. 2012, 89, 449–455.

(32) Shultz, G. V.; Gottfried, A. C.; Winschel, G. A. Impact of General Chemistry on Student Achievement and Progression to Subsequent Chemistry Courses: A Regression Discontinuity Analysis. J. Chem. Educ. 2015, 92, 1449–1455.

(33) Ryan, M. D.; Reid, S. A. Impact of the Flipped Classroom on Student Performance and Retention: A Parallel Controlled Study in General Chemistry. J. Chem. Educ. 2016, 93, 13–23.

(34) Lewis, S. E. Retention and Reform: An Evaluation of Peer-Led Team Learning. J. Chem. Educ. 2011, 88, 703–707.

(35) Crimmins, M. T.; Midkiff, B. High Structure Active Learning Pedagogy for the Teaching of Organic Chemistry: Assessing the Impact on Academic Outcomes. J. Chem. Educ. 2017, 94, 429–438.

(36) Rosenbaum, P. R.; Rubin, D. B. The Central Role of the Propensity Score in Observational Studies for Causal Effects. Biomet. 1983, 70, 41–55.

(37) National Research Council. Discipline-Based Education Research: Understanding and Improving Learning in Undergraduate Science and Engineering; The National Academies: Washington, DC, 2012.

(38) Xu, X.; Lewis, J. E. Refinement of a Chemistry Attitude Measure for College Students. J. Chem. Educ. 2011, 88, 561–568.

(39) Bauer, C. F. Attitude toward Chemistry: A Semantic Differential Instrument for Assessing Curriculum Impacts. J. Chem. Educ. 2008, 85, 1440-1445.

(40) Dalgety, J.; Coll, R. K.; Jones, A. Development of Chemistry Attitudes and Experiences Questionnaire (CAEQ). J. Res. Sci. Teach. 2003, 40, 649–668.

(41) Hetherington, J. Role of Theory and Experimental Design in Multivariate Analysis and Mathematical Modeling. In Handbook of applied multivariate statistics and mathematical modeling; Elsevier, San Diego: CA, 2000; pp 37–63.

(42) Lang, K. M.; Sweet, S. J.; Grandfield, E. M. Getting beyond the Null: Statistical Modeling as an Alternative Framework for Inference in Developmental Science. Res. Hum. Dev. 2017, 14, 287–304.

(43) Robert, J.; Lewis, S. E.; Oueini, R.; Mapugay, A. Coordinated Implementation and Evaluation of Flipped Classes and Peer-Led Team Learning in General Chemistry. J. Chem. Educ. 2016, 93, 1993–1998.

(44) Eichler, J. F.; Peeples, J. Online Homework Put to the Test: A Report on the Impact of Two Online Learning Systems on Student Performance in General Chemistry. J. Chem. Educ. 2013, 90, 1137–1143.

(45) Cook, E.; Kennedy, E.; McGuire, S. Y. Effect of Teaching Metacognitive Learning Strategies on Performance in General Chemistry Courses. J. Chem. Educ. 2013, 90, 961–967.

(46) Tinsley, H. E.; Brown, S. D. Handbook of Applied Multivariate Statistics and Mathematical Modeling; Elsevier: San Diego, CA, 2000.

(47) Byrne, B. M. Structural Equation Modeling with AMOS: Basic Concepts, Applications, and Programming; Routledge: New York, NY, 2016.

(48) Mueller, R. O.; Hancock, G. R. In Best Practices in Quantitative Methods; SAGE Publications, Inc.: Thousand Oaks, CA, 2018, pp 458-510.

(49) Kline, R. B. In Principles and Practice of Structural Equation Modeling; Guilford: New York, NY, 2015.

(50) DiLalla, L. F. In Handbook of applied multivariate statistics and mathematical modeling; Elsevier: San Diego, CA, 2000; pp 439–464.

(51) Huberty, C. J.; Petoskey, M. D. Multivariate Analysis of Variance and Covariance. In Handbook of applied multivariate statistics and mathematical modeling; Elsevier: San Diego, CA, 2000; pp 183–208.

(52) Theobald, E. Students Are Rarely Independent: When, Why, and How to Use Random Effects in Discipline-Based Education Research. CBE—Life Sci. Educ. 2018, 17, 1-12.

Supporting Information for

Metrics and Methods Used to Evaluate Student Performance Data in Chemistry Education Research Articles

Michael R. Mack,[†] Cory C. Hensen,[‡] and Jack Barbera[‡]

[†]Department of Chemistry, University of Washington, Seattle, Washington 98195-5850, United States

[‡]Department of Chemistry, Portland State University, Portland, Oregon, 97207-0751, United States

Table of Contents

List of Sampled Articles

Table S1. List of articles reviewed in this manuscript. Labels represent how each is referred to within the manuscript and additional Supporting Information tables below.

Label	Year	Author(s)	Title
Amaral et al.	2013	Amaral, Shank, Shibley, & Shibley	Web-Enhanced General Chemistry Increases Student Completion Rates, Success, and Satisfaction
Carmel et al.	2015	Carmel,Jessa, & Yezierski	Targeting the Development of Content Knowledge and Scientific Reasoning: Reforming College-Level Chemistry for Nonscience Majors
Chase et al.	2013	Chase,Pakhira, & Stains	Implementing Process-Oriented, Guided-Inquiry Learning for the First Time: Adaptations and Short-Term Impacts on Students' Attitude and Performance
Christiansen et al.	2016	Christiansen, Lambert, Nadelson, Dupree, & Kingsford	In-Class Versus At-Home Quizzes: Which is Better? A Flipped Learning Study in a Two-Site Synchronously Broadcast Organic Chemistry Course
Conway	2014	Conway	Effects of Guided Inquiry versus Lecture Instruction on Final Grade Distribution in a One-Semester Organic and Biochemistry Course
Cook et al.	2013	Cook, Kennedy, & McGuire	Effect of Teaching Metacognitive Learning Strategies on Performance in General Chemistry Courses
Crimmins and Midkiff	2017	Crimmins and Midkiff	High Structure Active Learning Pedagogy for the Teaching of Organic Chemistry: Assessing the Impact on Academic Outcomes
Eichler and Peeples	2013	Eichler and Peeples	Online Homework Put to the Test: A Report on the Impact of Two Online Learning Systems on Student Performance in General Chemistry
Esterling and Bartles	2013	Esterling and Bartles	Atoms-First Curriculum: A Comparison of Student Success in General Chemistry
Hall et al. 2012	2012	Hall, Wilson, & Sanger	Student Success in Intensive versus Traditional Introductory College Chemistry Courses
Hall et al. 2014	2014	Hall, Curtin-Soydan, & Canelas	The Science Advancement through Group Engagement Program: Leveling the Playing Field and Increasing Retention in Science
Hein	2012	Hein	Positive Impacts Using POGIL in Organic Chemistry
Hibbard et al.	2016	Hibbard, Sung, & Wells	Examining the Effectiveness of a Semi-Self-Paced Flipped Learning Format in a College General Chemistry Sequence
Lewis 2011	2011	Lewis	Retention and Reform: An Evaluation of Peer-Led Team Learning
Lewis 2014	2014	Lewis	Investigating the Longitudinal Impact of a Successful Reform in General Chemistry on Student Enrollment and Academic Performance
Malik et al.	2014	Malik, Martinez, Romero, Schubel, & Janowicz	Mixed-Methods Study of Online and Written Organic Chemistry Homework
Mooring et al.	2016	Mooring, Mitchell, & Burrows	Evaluation of a Flipped, Large-Enrollment Organic Chemistry Course on Student Attitude and Achievement
Muthyala and Wei	2012	Muthyala and Wei	Does Space Matter? Impact of Classroom Space on Student Learning in a Organic-First Curriculum

Rath et al.	2012	Rath, Peterfreund, Bayliss, Runquist, & Simonis	Impact of Supplemental Instruction in Entry-Level Chemistry Courses at a Midsized Public University
Robert et al.	2016	Robert, Lewis, Oueini, & Mapugay	Coordinated Implementation and Evaluation of Flipped Classes and Peer-Led Team Learning in General Chemistry
Ryan and Reid	2016	Ryan and Reid	Impact of the Flipped Classroom on Student Performance and Retention: A Parallel Controlled Study in General Chemistry
Shattuck	2016	Shattuck	A Parallel Controlled Study of the Effectiveness of a Partially Flipped Organic Chemistry Course on Student Performance, Perceptions, and Course Completion
Shultz et al.	2015	Shultz, Gottfried, & Winschel	Impact of General Chemistry on Student Achievement and Progression to Subsequent Chemistry Courses: A Regression Discontinuity Analysis
Weaver and Sturtevant	2015	Weaver and Sturtevant	Design, Implementation, and Evaluation of a Flipped Format General Chemistry Course

Reading Protocol

Context

- Why are the evaluators judging students' performance in a chemistry course?
- In which settings (e.g., classes, size, institution, etc.) are students' performance being judged?
- Who are the performers?

Research questions

1. What experimental design was used to compare student performance?
2. What metrics were used to make judgements about performance?
3. What types of comparison groups were used to make judgements about treatment effects?
4. What methods were used to analyze the performance data?

Validity considerations

- (Internal) In what ways did the authors address threats to the validity of inferences about the relationship between treatments and outcomes?
- (Statistical) In what ways did the authors address threats to the validity of statistical inferences?
- (Construct) In what ways did the authors address threats to the accurate characterization of treatments and their relationships to measured outcomes?
- (External) In what ways did the authors address threats to the generalization of inferences across varied persons, settings, treatments, and outcomes?

Data Matrices

Table S2. Distribution of quasi-experimental frameworks used in the sample of articles. Article labels can be matched to the list of references in Table S2. See the 'Code Descriptions' section for more details about each code.

Design Label	Code	Count	Article Label
#1	One Group Posttest-Only Design	1	Weaver and Sturtevant
#2	One Group Pretest-Posttest Design	0	
#3	Posttest-Only Alternative Treatment (or No Treatment) Control Group Design	13	Amaral et al.; Carmel et al. Chase et al; Christiansen et al.; Conway; Hein; Hibbard et al.; Lewis 2011; Malik et al; Mooring et al; Muthyala and Wei; Robert et al.; Shattuck
#4	Pretest-Posttest Alternative Treatment (or No Treatment) Control Group Design	7	Cook et al.; Crimmins and Midkiff; Eichler and Peeples; Esterling and Bartels; Hall et al. 2012; Rath et al.; Ryan and Reid
#5	Pretest-Posttest Alternative Treatment (or No Treatment) Control Group Regression Discontinuity Design	1	Shultz et al.
#6	Posttest-Only Alternative Treatment (or No Treatment) Control Group Design with Removed Treatment	1	Lewis 2014
#7	Pretest-Posttest Alternative Treatment (or No Treatment) Control Group Design with Removed Treatment	0	
#8	Posttest-Only Alternative Treatment (or No Treatment) Control Group Design with Repeated Measures	1	Hall et al. 2014
#9	Pretest-Posttest Alternative Treatment (or No Treatment) Control Group with Repeated Measures	0	

Table S3. Distribution of the five most common performance metrics reported in the sample of articles. Article labels can be matched to the list of references in Table S2. See the 'Code Descriptions' section for more details about each code.

Code	Count	Article Label
Instructor-authored exams	12	Carmel et al.; Chase et al.; Christiansen et al.; Conway; Cook et al.; Crimmins and Midkiff; Eichler and Peeples; Hall et al. 2014; Muthyala and Wei; Robert et al.; Ryan and Reid; Shattuck
Persistence/Retention Rates	11	Amaral et al.; Chase et al.; Esterling and Bartels; Hall et al. 2014; Lewis 2011; Mooring et al.; Rath et al.; Robert et al.; Ryan and Reid; Shattuck; Shultz et al.
ACS exams	9	Hein; Hibbard et al.; Lewis 2011; Lewis 2014; Malik et al.; Mooring et al.; Robert et al.; Ryan and Reid; Weaver and Sturtevant
Letter grades	7	Amaral et al.; Conway; Crimmins and Midkiff; Hall et al. 2012; Hibbard et al.; Mooring et al.; Shattuck
Course grade-point average	4	Amaral et al.; Cook et al.; Hall et al. 2014; Rath et al.

Table S4. Distribution of five most common types of comparisons reported in the sample of articles. Article labels can be matched to the list of references in Table S2. See the 'Code Descriptions' section for more details about each code.

Code	Count	Article Label
Concurrent	12	Chase et al.; Christiansen et al.; Cook et al.; Eichler and Peeples; Hall et al. 2014; Lewis 2011; Lewis 2014; Malik et al.; Muthyala and Wei and Wei; Robert et al.; Ryan and Reid; Shattuck
Historical	9	Amaral et al.; Carmel et al.; Conway; Crimmins and Midkiff; Esterling and Bartels; Hall et al. 2014; Hein; Hibbard et al.; Mooring et al.
Successive assessments	12	Chase et al.; Christiansen et al.; Esterling and Bartels; Hall et al. 2014; Lewis 2014; Muthyala and Wei; Rath et al.Robert et al.; Ryan and Reid; Shattuck; Shultz et al.; Weaver and Sturtevant
Student subpopulations	7	Eichler and Peeples; Lewis 2011; Lewis 2014; Rath et al.; Robert et al.; Ryan and Reid; Shultz et al.
Normed comparisons	3	Hein; Ryan and Reid; Weaver and Sturtevant
Multiple outcome measures	3	Chase et. al; Hibbard et al.; Mooring et al.

Table S5. Distribution of methods for statistical inference used in the sample of articles. Article labels can be matched to the list of references in Table S2. See the 'Code Descriptions' section for more details about each code.

Code	Subcode	Count	Article Label
Multivariate Statistics	Multiple predictors and outcomes	0	
	Multiple outcomes	0	
	Multiple predictors	10	Crimmins and Midkiff; Cook et al.; Eichler and Peeples; Esterling and Bartels; Hall et al. 2012; Hein; Hibbard et al.; Rath et al.; Robert et al.; Schultz et al.
Univariate Statistics		19	Amaral et al.; Carmel et al.; Chase et al.; Christiansen et al.; Conway; Cook et al.; Hall et al. 2014; Hein; Hibbard et al.; Lewis 2011; Lewis 2014; Malik et al.; Mooring et al.; Muthyala and Wei; Rath et al.; Robert et al.; Ryan and Reid; Shattuck; Weaver and Sturtevant
Descriptive Statistics		22	Amaral et al.; Carmel et al.; Chase et al.; Christiansen et al.; Conway; Cook et al.; Crimmins and Midkiff; Eichler and Peeples; Esterling and Bartels; Hall et al. 2012; Hall et al. 2014; Hein; Hibbard et al.; Lewis 2014; Lewis 2011; Malik et al.; Muthyala and Wei; Rath et al.; Robert et al.; Ryan and Reid; Shattuck; Weaver and Sturtevant

Code Descriptions

Question 1: What types of experimental designs have been used to study the effects of instructional interventions on student performance?

One Group Posttest-Only Design (Design #1)
In this design, a group receives a treatment and afterwards they are assessed on the outcome measure.

One Group Pretest-Posttest Design (Design #2)
In this design, a group is first assessed on a pretreatment assessment, subsequently receive a treatment condition, and afterwards they are assessed on the outcome measure.

Posttest-Only Alternative Treatment (or No Treatment) Control Group Design (Design #3)
In this design, one group receives a treatment while a different group receives an alternative treatment (or the absence of that treatment). Each group is later assessed on the same outcome measure at about the same time and under similar conditions.

Pretest-Posttest Alternative Treatment (or No Treatment) Control Group Design (Design #4)
In this design, two groups are first assessed on the same outcome measure at about the same time and under similar conditions. Then, one group receives a treatment while a different group receives an alternative treatment (or the absence of that treatment). Each group is later assessed on the same outcome measure at about the same time and under similar conditions.

Pretest-Posttest Alternative Treatment (or No Treatment) Control Group Regression Discontinuity Design (Design #5)
Like the previous design, two groups are first assessed on the same outcome measure at about the same time and under similar conditions. Then, participants are distributed across conditions based on a cutoff score or an alternative assignment variable. One group receives a treatment while the other group receives an alternative treatment (or the absence of that treatment). Each group is later assessed on the same outcome measure at about the same time and under similar conditions.

While regression discontinuity (RD) designs have unique properties that afford researchers unbiased causal inferences under certain assumptions, we follow Shadish et al.'s classification of RD designs as quasi-experiments because they lack random assignment.[1] In fact, the only study in our sample with a RD design is technically a "fuzzy" RD design because in some cases the assignment to experimental condition did not adhere fully to the cutoff.[2] But as Shadish et al. note, "a fuzzy cutoff RD design may produce better estimates than many other quasi-experiments if the fuzziness is not too great" (p. 229).[1]

Posttest-Only Alternative Treatment (or No Treatment) Control Group Design with Removed Treatment (Design #6)
This design extends Design #3 by adding another wave of exposure and posttest measurement, but this time both groups receive the alternative treatment or absence of treatment. This design allows for the comparison of long-term effects of a treatment X after the treatment has been removed to a comparison group that received an alternative treatment or no treatment all along.

Pretest-Posttest Alternative Treatment (or No Treatment) Control Group Design with Removed Treatment (Design #7)
This design extends Design #4 by adding another wave of pretest, exposure, and posttest measurement, but this time both groups receive the alternative treatment or absence of treatment. This design allows for the comparison of long-term effects of a treatment X after the

treatment has been removed to a comparison group that received an alternative treatment or no treatment all along. Unlike Design #6, this design allows for within-group comparisons across assessments to understand how the outcome rises and falls with the introduction and removal of treatment.

Posttest-Only Alternative Treatment (or No Treatment) Control Group Design with Repeated Measures (Design #8)
This design extends Design #3 by adding another wave of exposure and posttest measurement. Unlike Design #6, the treated group is repeatedly exposed while the treatment is withheld from the comparison group or they repeatedly receive an alternative treatment. This design affords estimates for the sustained effect of treatment over time in contrast to the sustained effect of an alternative treatment or the absence of the treatment. As Shadish et al. explained, "few threats to validity could explain a close relationship between treatment introductions and removals on the one hand and parallel changes in outcome on the other… Such threats would have to come and go on the same schedule" (p. 113).[1]

Pretest-Posttest Alternative Treatment (or No Treatment) Control Group with Repeated Measures (Design #9)
This design extends Design #4 by adding another wave of pretest, exposure, and posttest measurement. Unlike Design #7, the treated group is repeatedly exposed while the treatment is withheld from the comparison group or they repeatedly receive an alternative treatment. Similar to Design #8, this design affords estimates for the sustained effect of treatment over time in contrast to the sustained effect of an alternative treatment or the absence of the treatment.

Question 2: Which metrics have been used to measure student performance outcomes?

Instructor-authored exams
In general, instructor-authored exams are high-stakes, summative evaluations of students' chemistry content knowledge that were authored by the instructor of a course or authored by a common group of instructors. Both midterm and final exams are included in this category.

Persistence/Retention Rates
Persistence/retention rates were indicated by pass rates, withdrawal rates, or the proportion of students receiving D or F grades or withdrawing from the course (i.e., "DFW" rate).

ACS exams
ACS exams refer to examinations developed and distributed by the American Chemical Society Examinations Institute.

Letter grades
Letter grades refer to course grades on an ordinal scale (e.g., A, B, C, and D). The values rank students according to their performance.

Course grade-point average
Course grade-point average refers to course grades on an interval scale (e.g., 0-4 scale). The values rank students according to their performance.

Question 3: What types of groups have been used to compare performance across experimental conditions?

Concurrent comparison group
In a concurrent comparison, experimental groups receive different treatment conditions at the same time.

Historical comparison group
In a historical comparison, researchers compare outcomes from the treated group with outcomes among participants who received the alternative treatment (or absence of treatment) at a previous time.

Time/successive assessments
This category encompasses studies that compared the same metric over time, such as grades in a course-based intervention and then also grades in subsequent chemistry courses. Alternatively, some studies compared performance on successive assessments within a course, such as midterms, final exams, and course grades.

Student subpopulations
This code refers to comparing outcomes across student subpopulations, such as men and women, racial/ethnic minority and well-represented groups, etc.

ACS norming sample
ACS exam outcomes were compared with the original norming sample. In some studies, contrasts are only made to the nationally normed dataset and no local contrasts were made across experimental groups. In other studies, researchers compare ACS exam outcomes across local groups and to the original norming sample.

Comparing Performance and Affective Outcomes
Performance and affective outcomes were measured and reported, but not necessarily correlated simultaneously with the experimental condition.

Question 4: What quantitative methods have been used to analyze the comparison data?

Descriptive statistics
Researchers used descriptive statistics to summarize performance outcomes and other group comparisons using tables and/or graphs.

Univariate statistics
Researchers working within univariate statistical frameworks constructed falsifiable null and alternative hypotheses about group differences and then tested the tenability of the null hypothesis based on sample data. Common statistical tests include the *t*-test, chi-squared test, analysis of variance, and non-parametric extensions.

Multivariate statistics
Researchers working within multivariate statistical frameworks constructed falsifiable null and alternative hypotheses about group differences beyond the effects of prior abilities, demographic characteristics, or other student-level factors and then tested the tenability of the null hypothesis based on sample data. Common multivariate statistical techniques used in the sample of articles included multiple regression and ANCOVA.

References

(1) Shadish, W. R.; Cook, T. D.; Campbell, D. T. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*; Houghton Mifflin: Boston, MA, 2002.

(2) Shultz, G. V.; Gottfried, A. C.; Winschel, G. A. Impact of General Chemistry on Student Achievement and Progression to Subsequent Chemistry Courses: A Regression Discontinuity Analysis. *J. Chem. Educ.* **2015**, *92*, 1449–1455.