Judging the Credibility of Quantitative Student Ratings of Instruction Research

IDEA Paper #80 • August 2019

Stephen L. Benton and Dan Li • The IDEA Center

Abstract

Periodically, articles reporting research on student ratings of instruction (SRI), aka student evaluations of teaching, appear in the higher-education press. This literature often summarizes studies that challenge the validity and reliability of SRI. However, before drawing a conclusion about a quantitative study touted in the media, readers should evaluate both the credibility and generalizability of the primary source. In this paper, the authors review one set of criteria that aids in such evaluation—David Krathwohl's (2009) judgments about (a) internal validity or linking power; and (b) external validity or generalizing power. *Internal validity* is the extent to which a study demonstrates that its investigated variables are linked in a relationship. *External validity* is how well a study establishes that its findings are generalizable. Applying such criteria can prevent biased takeaways created from merely reading a news article without assessing the quality of the research paper it summarizes.

Keywords: Internal validity, external validity, student ratings of instruction, student evaluations of teaching

Periodically, articles reporting research on student ratings of instruction (SRI), aka student evaluations of teaching, appear in the higher-education press (e.g., Bunge, 2018; Falkoff, 2018; Flaherty, 2016; Lawrence, 2018). This literature often cites studies that challenge the validity and reliability of SRI. However, before drawing a conclusion from a quantitative study touted in the media, readers should apply established research standards to evaluate both the credibility and generalizability of the primary source. In this paper, we review one set of criteria that aids in such evaluation—David Krathwohl's judgments about internal validity, or linking power; and external validity, or generalizing power (2009). *Internal validity* is the extent to which a researcher demonstrates that the investigated variables are linked in a relationship. Judgments about internal validity include explanation credibility, translation validity, demonstrated result, and the elimination of rival explanations. *External validity* is how well a study establishes that its findings are generalizable. Judgments about external validity focus on explanation generality, translation generality, demonstrated generality, and the elimination of restrictive explanations. We then explain how these judgments build conceptual and empirical support in the chain of reasoning that reinforces a study's credibility and generalization. Applying such criteria can prevent biased takeaways created from merely reading a news article without assessing the quality of the research paper it summarizes.

Krathwohl's Judgments About Internal Validity

Internal validity concerns the linking power of a study—its ability to create a consensus that the investigated variables are interrelated. Krathwohl (2009) uses the term *linking power* because Campbell and Stanley's original definition of *internal validity* (1963) was limited to the control of confounding variables and did not account for judgments about how well a study's explanation and rationale were linked with the credibility of its results. Similar to Campbell and Stanley's original definition, internal validity reduces the uncertainty that a relationship exists between the variables that are being investigated. Readers should therefore look for elements in the research design that either add to or lessen that uncertainty.

Conceptual Evidence

Conceptual evidence, the first element that bolsters internal validity, entails the researcher's justification for the study: connecting it to theory and previous research. When presented well, conceptual evidence lays solid groundwork for the study and helps distinguish chance results from those that more logically relate to it. Readers would know, for example, that if teachers whose astronomical sign is Gemini received the highest SRI scores, this result would most likely have occurred by chance, because it has no conceptual support. As shown in Table 1, conceptual evidence is reinforced by explanation credibility and translation validity.

Table 1
Judgments Supporting Internal Validity and External Validity

Type of evidence	Internal validity	External validity
Conceptual	Explanation credibility	Explanation generality
	Translation validity	Translation generality
Empirical	Demonstrated result	Demonstrated generality
	Rival explanations eliminated	Restrictive explanations eliminated
Final judgment	Credible result	Replicable result

Explanation Credibility

Explanation credibility concerns a judgment about how effectively the author has provided a rationale for the study. This first verdict is critical if one is to be convinced to read on and consider the possibility that the proposed hypothesis could be supported by the data. The research questions and hypotheses should follow logically from previous studies and theory. Are we convinced that the research question was worth pursuing and that there could possibly be a relationship or causal link between the investigated variables? If the explanation of the study's relevance is sound, we are open to exploring whether the data support the proposed relationships.

For example, Sohr-Preston and colleagues (Sohr-Preston, Boswell, McCaleb, & Robertson, 2016) examined potential sources of bias in interpreting and posting comments on RateMyProfessor.com (RPM.com). They began by devoting several paragraphs to the pros and cons of RMP.com. Even though RMP.com suffers from several validity issues (e.g., ratings can be posted by anyone, whether enrolled in a given professor's course or not), the authors made a convincing case that RMP.com was worthy of investigation. Some students rely on it when deciding which classes to take because it is the only such widely available public resource. Moreover, the authors cited studies that indicate that RMP.com influences student expectations and motivation,

actual ratings of instructors, and in-class behaviors, as well as college rankings, faculty promotion, hiring decisions, and faculty affect and self-efficacy. After reading Sohr-Preston et al.'s introduction, most readers would probably be convinced that the study was worth doing.

In contrast, Carrell and West (2010) were less convincing in making the case for investigating the utility of a value-added approach to teacher evaluation. They made unsubstantiated claims that teachers (a) "can influence [student evaluations] in ways that may reduce actual student learning" and (b) "can inflate grades or reduce academic content to elevate student evaluations" (p. 410). No citations were provided to back such allegations. The authors presented them as if they were irrefutable facts. And yet nothing in the article even addressed, let alone supported, those claims. It raises the question, then, whether the point of the investigation was to objectively assess the value-added approach to evaluation or to "take on" SRI. At the outset, the reader might wonder whether the authors would be likely to consider rival explanations for any significant findings.

An assessment of explanation credibility is thus a judgment of how solid the author's rationale for the study and how plausible any proposed relationships really are. If the explanation is convincing, the reader is ready to examine how the study is translated and operationalized in the procedures.

Translation Validity

Translation validity is a judgment of whether the research design and methods are faithful to the study's purpose. Has the investigation been carried out in a manner consistent with the proposed examination of the hypothesized relationships? To assess translation validity, Krathwohl (2009) recommends that readers ask six questions about the research methods: Who? Where? Why (the cause)? What (the effect)? How? And when?

One should first ask, "Who were the study's participants, and were they appropriate for the research question?" Readers might reason that any inquiry into SRI should be conducted on actual college students. The vast majority of investigations probably are. Certainly, for the purposes of external validity, participants should be as similar as possible to the population being targeted for generalization. But with respect to internal validity, participants can be any group to which the research question would apply (Krathwohl, 2009). In investigating ways that teaching evaluations are misinterpreted, Boysen and colleagues (Boysen, Kelly, Raesly, & Casner, 2014), for example, appropriately surveyed university faculty and department heads.

As another example, suppose a researcher wanted to investigate whether gender bias was reflected in first impressions of an instructor. Photos of bogus instructors, differing by gender (the cause), could be shown to a sample of participants, who would then be asked to rate the instructor on various qualities (the effect). Participants would not necessarily have to be actual college students. They would just need to be individuals who potentially could be susceptible to gender bias.

Also relevant is the question of "where" the study was conducted. Most SRI research is probably done on data collected from students completing instruments in either an actual classroom or a Web-based environment. Again, for the purposes of external validity, this makes good sense. However, even though the classroom preserves a study's ecological validity, there are many uncontrolled factors (e.g., distractions, time limits, instructor presence) that might undermine internal validity. The key issue, then, is whether the phenomenon under study could potentially be displayed in the chosen setting (Krathwohl, 2009). Returning to the hypothetical example of gender bias in student first impressions, data could be collected outside the regular classroom, in a setting where the researcher could standardize the procedures for viewing and responding to the photos.

Readers should next ask "why" the proposed cause would be expected to occur. What would supposedly cause the effect to be observed? In the example of the instructor first-impression study, the author would

first need to convince us that the possibility of gender bias exists and, second, that the manipulation of the bogus professor descriptions is credible. The research procedures would also need to be standardized across all participants, thereby making the process almost identical for all. If not, threats to internal validity could creep in, raising doubts about whether something other than gender bias caused any observed effects.

The observations or measures are the "what" of the study, or the means for showing that the effect occurred. Specifically, one should ask whether the instruments or observations were psychometrically sound; that is, backed by reliability and validity evidence. Also, was more than one method of gathering data employed (e.g., quantitative and qualitative)? Was more than one measure or observation used for each method? Although triangulation of data can be an advantage, using a single measure is not necessarily a weakness if it can be theoretically and empirically defended. When it cannot be, however, internal validity is compromised.

For example, in a study intended to investigate instructor-gender bias in SRI, Boring and colleagues (Boring, Ottoboni, & Stark, 2016) used separate SRI instruments to survey students in France and the United States. However, the article's method section was ambiguous about what those instruments actually measured. Regarding the French SRI, the authors reported only that it included both closed- and open-ended questions; they provided no details about the actual items. In addition, evidence of reliability and validity was offered for neither the SRI nor a final exam designed to measure student learning. Therefore, readers could not know what was actually being measured and correlated with gender. To their credit, the authors did provide relatively more information about the instrument in the U.S. sample. Still, they offered no empirical support for the SRI's validity or reliability. To read a more elaborate critique of this article, see Ryalls, Benton, Barr, and Li (2016).

Next, if the hypothesized relationship is supported, one needs to ask "how" it occurred. It helps if the effect appears when the cause is present but disappears when the cause is absent. For example, in the Sohr-Preston et al. study (2016), college students were randomly assigned to read one of four different versions of a bogus professor's online SRI summaries, modeled on RateMyProfessor.com. They were then asked to complete ratings of the professor's dedication, attractiveness, enhancement, fairness, and clarity. The four descriptions of bogus professors contained the same content except for the identity of the professor, who was described as either (a) a woman with a chili pepper, intended to indicate "hotness"; (b) a woman with no chili pepper; (c) a man with a chili pepper; or (d) a man with no chili pepper. Because participants were randomly assigned to the gender and "hotness" conditions prior to completing ratings, precedence of the cause (i.e., "perceived hotness") to the effect (ratings of the professor) was established.

The final question pertaining to translation validity concerns "when" the procedures occurred. Was enough detail provided to remove any ambiguity about that? For example, if an SRI instrument was administered, when did that occur? During class? On students' own time? Before or after the final exam? If comparisons were made between different groups, were measures collected on or around the same day across all conditions?

The importance of answering questions related to translation validity cannot be over-emphasized. Weaknesses in research are usually related to shortcomings in how the study was conducted or how the data were collected and analyzed. If the research question and hypothesis have not been adequately translated and credibly carried out, the authors end up doing a different study than the one that they intended. Moreover, if the gap between intention and translation is wide enough, the empirical evidence collected is irrelevant. For instance, although Sohr-Preston et al. (2016) established a credible rationale for investigating RMP.com postings, they did not actually analyze them. Instead they created proxies of RMP.com, using online scenarios for students to rate bogus professors who varied by gender and attractiveness. Because the study became more

about investigating bias in SRI than about RMP.com, Sohr-Preston et al. fell short of translation validity.

Empirical Evidence

Empirical evidence, the second major component underlying internal validity, is built through proper research design and data analyses that demonstrate the predicted result. The researcher must make a convincing argument that the hypothesized relationship actually occurred under certain conditions. Two factors strengthen empirical evidence: demonstrated result and the elimination of rival explanations.

Demonstrated Result

Three attributes are particularly salient in deciding whether the hypothesized result has been demonstrated: authenticity of the evidence, precedence of cause, and presence of effect (Krathwohl, 2009).

Authenticity of the evidence concerns whether the data collected were what they purported to be. If an SRI was used, what did it actually measure? Is there any evidence that it assessed student perceptions of either teaching effectiveness or course quality? What is known about the individual questions? Has evidence of reliability and validity been presented? Buchert and colleagues (Buchert, Laws, Apperson, & Bregman, 2008), for instance, examined the effects of instructor reputation versus first impressions on a locally developed SRI. Students completed an SRI survey either before the first class, to assess their perceptions of the instructor's reputation, or within the first two weeks of the class, to gauge their first impressions of the instructor. The SRI was administered again to all students at the end of the term. The authors published each of the 18 SRI survey items, enabling readers to make judgments about face validity. However, they did not present information about the instrument's reliability. Therefore, the amount of measurement error found in the comparisons of ratings collected at various times during the course is unknowable.

Precedence of cause means that any manipulation (i.e., the cause) must precede, or be concomitant with, the effect; not the reverse. As described previously, in the first part of the Sohr-Preston et al. study (2016), precedence of the cause (i.e., chili pepper vs. no chili pepper) was established, because students were randomly assigned to conditions where they first read the description of the bogus professor and then completed the ratings. In the second part of the study, the same students were asked to rate their current instructor's teaching performance and report his or her gender and perceived "hotness." Precedence of cause would be more difficult to demonstrate in the natural situation of rating their actual professor, because students may have had various reasons for their ratings based on experience, rather than on instructor gender or his or her perceived "hotness."

Presence of effect is not always easily detectable in the social and behavioral sciences compared to the biological and physical sciences, and measurements are typically less reliable (Tobias, 1976). Assessing student perceptions of teaching quality is not as precise as measuring heart rate. Consequently, a small effect in the hypothesized direction may occur by chance, due to sampling or measurement error. Statistics enable researchers to estimate the probability of a particular score occurring, which aids in decision making. However, one should carefully evaluate how the data were analyzed and interpreted. Was the analysis appropriate for the methods employed and hypotheses being tested? For example, was an appropriate significance level set? In a study of statistical errors in psychology journals, 63% of the articles examined contained at least one incorrect *p* value, and 20% of those would have affected decisions about statistical significance (Veldkamp, Nuijten, Dominquez-Alvarez, van Assen, & Wicherts, 2014). If the results are significant, readers should also assess whether they are meaningful or trivial. Did the author make too much of too little? Or could the results potentially have implications for decision making? Was a measure of effect size reported, such as eta-squared or Cohen's *d* (1988)? Cohen considered effect sizes approximating .20 (1/5 standard deviation) as small, .50 as medium, and .80 as large. If an effect size was not reported, readers may be able to compute one themselves, depending

on the statistics reported in the article.

To summarize, then, a demonstrated result surfaces when (a) the measures or observations are accurate and reliable, (b) there is precedence or concurrence of the cause, and (c) appropriate analyses (quantitative or qualitative) are employed. Even so, the researcher must consider alternative explanations for the results.

Rival Explanations Eliminated

Another aspect of empirical support is the judgment of whether equally plausible explanations exist for the results, other than the one the author puts forth. Counterarguments almost always exist, some more reasonable than others, depending on how well the study was designed. Skilled researchers anticipate rival explanations and design a study in a way that renders them less plausible. They demonstrate that the effect is present only when the proposed cause is present. If appropriate, they randomly assign participants to groups and randomly assign groups to treatments. Moreover, they expose participants to the exact same procedures except for the independent variable (i.e., the cause).

However, sometimes rival explanations turn up. Alvero, Mangiapanello, and Valad (2019) faced one when they investigated the effects of various strategies for increasing student response rates to SRI. They found, as hypothesized, that classes with incentives (i.e., extra credit points) had significantly higher response rates than those without. Nonetheless, Alvero et al. acknowledged the possibility that incentives may have biased the sample by increasing responses mainly from low-performing students (Nulty, 2008). However, they then counterargued with evidence that response rates within incentive groups did not differ by students' course grades. Moreover, class average GPAs did not differ among all four incentive groups. Thus the authors acknowledged, but then competently challenged, a rival explanation.

In contrast, MacNell, Driscoll, and Hunt (2014) left the door open to one. The authors randomly assigned students enrolled in an online introductory anthropology/sociology course to one of four discussion sections, two taught by a male instructor and two by a female instructor. For one of their two sections (apparently not randomly determined), the instructors falsely identified their gender, thereby creating "actual gender" and "perceived gender" conditions. Although no differences were found in end-of-course evaluations between actual-male and actual-female sections, the perceived-male section was rated more highly on fairness, praise, and promptness than was the perceived-female section. This study was championed in an article in *Inside Higher Education* under the headline "Students Give Professors Better Evaluations if They Think They're Male" (Mulhere, 2014).

However, that conclusion may have been premature, because MacNell et al. (2014) reported that "All instructors were aware of the study being conducted and cooperated fully" (p. 296). So, in other words, one can assume that the instructors knew that in one section they were identified as a person of the opposite gender. As Krathwohl (2009) points out, expectancy effects can occur if researchers "inadvertently tip the scales in a variety of ways—verbally (for example with encouragement and clues) and nonverbally . . ." (p. 499). Notably, MacNell et al.'s three dependent measures—instructor praise, fairness, and promptness—could have certainly been affected by expectancy, because the instructors might have intentionally or unintentionally responded differently on the discussion boards across their two sections.

To counter research expectancy effects, and thereby eliminate a rival explanation, MacNell et al. (2014) could have used a double-blind procedure where neither the instructors nor anyone analyzing the data would have known which were the actual-gender and perceived-gender sections, but they chose not to do so. Although the authors reported that the instructors behaved exactly the same way in each section—the one for their actual and their perceived genders—they provided no empirical evidence for that claim, which they could have done through a content analysis of the discussion boards.

Final Judgment: Credible Result

As readers, we must make a final judgment about the credibility of a study based on our previous assessments of internal validity. Are we convinced, based on the conceptual and empirical evidence, that there is a relationship between the variables investigated? Did the research methods follow logically from the research questions? Has a result been convincingly demonstrated? Do the results align with previous research? If not, is it possible that the studies cited in the article may have had flaws? Finally, have credible rival explanations been eliminated? The bottom line is the degree of confidence we have in the author's interpretation of the results.

Krathwohl's Elements of External Validity

Krathwohl's (2009) second criterion, external validity, is the power of a study to reinforce generalization of the findings. Krathwohl defines *generalizing power* in this context as judgments about how well the study connects the plausibility of a generalization claimed with the replicability of its results. In line with Campbell and Stanley's original definition (1963), readers should consider whether a study's findings can be reasonably applied to other settings and populations. As with internal validity, external validity is supported by five judgments, shown in Table 1, which provide both conceptual and empirical evidence. We next consider the decisions readers must make about the credibility of such evidence as it relates to external validity.

Conceptual Evidence

As with internal validity, the first element that bolsters external validity is conceptual evidence, which, as shown in Table 1, is reinforced by explanation generality and translation generality.

Explanation Generality

Explanation generality is a judgment about whether generalization of the explanation for the study, stated or implied, is warranted. Almeida, Silva, and Mohring (2019), for example, looked at the effects of extraneous factors (e.g., student age and expected grade) on SRI in a single business-management course. Their rationale for the study was the premise that student feedback is important for maintaining the competitiveness of a specific business-management course and that feedback should accordingly be free from bias. In discussing their study's limitations, they explicitly stated, "The scope of the study is limited to one class . . . in management of the Business School chosen for the study" (p. 10). Thus, their stated generalization was consistent with what was implied in the rationale and therefore seems warranted.

However, explanation generality attributed to a study is not always so explicitly stated; it is either implied or must be inferred. Such was the case in a study by Young and colleagues (Young, Joines, Standish, & Gallagher, 2018), who investigated whether response rates to SRI would be higher in courses where students were permitted to complete Web-based ratings during class. Young et al. found that faculty who simply allocated time in class for students to complete the ratings saw, on average, a 29% increase in response rates over the previous semester's class. Although the authors did not claim that their findings applied to courses in other institutions, such generalization seems reasonable given the simple and effective nature of the intervention.

If readers are convinced of the credibility of the generalization—claimed, implied, or inferred—they are ready to form judgments about whether generalization is appropriate based on how the study was carried out—i.e., translation generality.

Translation Generality

Inferring generalization is always risky, but less so if the circumstances of the study are representative of the target population. *Translation generality* is a judgment about whether elements of the study's design—participants, setting, treatment, measures, procedures—are representative of the elements about which the researcher intends to generalize. Krathwohl (2009) recommends considering the following design elements.

Research participants and setting. Are the participants in the study and the setting in which it occurred representative of those to which the researcher intends to generalize? The more

information the authors provide, the easier it is to answer that question. As a general rule, authors should "describe the groups as specifically as possible, with particular emphasis on characteristics that may have bearing on the interpretation of results" (APA, 2011, p. 29). For example, in a study of the effects of professor age and gender on SRI, Wilson, Beyer, and Monteiro (2014) recruited undergraduate students enrolled in psychology courses at a southeastern U.S. institution. The authors appropriately reported the number of male and female students and how many were enrolled at the various class levels, as well as statistics on student age. Although Wilson et al.'s sample was one of convenience, it nonetheless comprised college students enrolled in an actual course. Moreover, the students completed online ratings in a manner that was probably similar to that of most settings, thereby supporting ecological validity (i.e., generalizability to real-life situations).

Treatment. For experimental and quasi-experimental designs, external validity is concerned with whether the likely variation in treatment effects occurring in the real world is represented. Wilson et al. (2014), for example, randomly assigned undergraduate students to one of four conditions that were identical save for the pictures of teachers, who varied by gender and age. However, because pictures of professors in the Wilson et al. study varied on only two dimensions each of gender (male, female) and age (young, old), real-life variation in the treatment effect was not represented. To maximize translation generality, the researchers might have instead included a representative variety of gender classifications (including nonbinary) and age ranges.

Observations and measures. Are the selected observations and measures representative of all possible valid observations and measures? If other measures had been used, would the results have been the same, or was there something unique about the ones selected? In truth, it is not uncommon for researchers to opt for a single measure, one uniquely designed for the study. In such cases, translation generalization might arguably be limited to that instrument. In contrast, Berk (2018) identified from the literature multiple sources of evidence that could be used to measure teaching effectiveness, including student end-of-course ratings, student midterm feedback, student exit and alumni ratings, student outcome measures, instructor self-ratings, teaching scholarships and awards, peer classroom observations, peer review of course materials, external expert ratings, video classroom review, teaching or course portfolio review, administrator ratings, and employer ratings.

Time. Some generalizations from research findings decay with the passage of time, especially if the culture changes. For example, because of the changing role of women in the workforce over the past few decades, research on gender and student ratings conducted prior to the 1980s may not necessarily generalize to the present time. As another example, the most pertinent findings from the Sohr-Preston et al. study (2016) were tied to the "hotness" variable, manipulated by the presence or absence of a chili pepper, a symbol connected with current culture. With the passage of time, that icon might not have the same meaning for future generations.

Procedures. Sometimes unique or complicated aspects of a study make it challenging to generalize the results. For example, McDonnell and Dodd (2017) asked undergraduate students to complete course-feedback forms (CFFs) on four different occasions during the semester: the 2nd, 6th, 11th, and 16th weeks. After the completion of the first three CFFs, the instructor went over the results in class, identified strengths and weaknesses based on student feedback, and made three changes in the class, as voted on by students. After the instructor implemented the changes, student responses to the question of "How good is this instructor" were higher on CFF 4 than for any of the previous CFFs, which suggests that the improvements had an impact. However, the multiple design elements in the study could present a challenge to those wishing to apply it to their own situations, thereby muddying generalization.

Empirical Evidence

As with internal validity, empirical evidence reinforces external validity, specifically through "demonstrated generality" and the elimination of restrictive explanations or conditions.

"Demonstrated Generality"

"Demonstrated generality" is present if the generalization "appeared in all the instances of the study in which it would be expected to do so and did not where it shouldn't" (Krathwohl, 2009, p. 180). Krathwohl surrounds *demonstrated generality* with quotation marks because it is logically impossible to demonstrate generality in all instances where it should apply. There will always be exceptions where generality did not occur in a situation in which it should have.

In the Sohr-Preston et al. study (2016), instructor attractiveness influenced SRI in two situations in which it was supposed to; however, the results were contradictory. Whereas presence of the chili pepper (i.e., high attractiveness) caused ratings of the bogus professor to be lower on clarity, actual professors whose students rated them highly on attractiveness received higher ratings on clarity. Thus the effect generalized across both instances, but in opposite ways! The findings were thus equivocal and led to no demonstrated generality regarding instructor clarity.

In contrast, Benton, Duchon, and Pallett (2013) examined the relationship between individual students' ratings of progress on IDEA SRI learning objectives and their performance on five exams in a college course. The instructor identified two objectives as relevant to the course and 10 as of minor or no importance. The authors hypothesized that student self-ratings of progress on course-relevant objectives would correlate positively with performance on course exams and the course total score but that ratings on minor or unimportant objectives would not. Correlations for progress on relevant objectives were indeed positive for four out of five exams and the total score, but those for minor or unimportant objectives were negligible. Thus, generalization appeared in most instances in which it was supposed to but not in those in which it wasn't.

Restrictive Explanations (Conditions) Eliminated

Certain design conditions may place restrictions on generalizing to the intended target population, such as procedures that involve elements unlikely to be carried out in the real world. Unless those restrictions are eliminated, generalizations are best limited to populations operating under those specific design elements. Legg and Wilson (2012), for example, investigated the effect of instructor touch on SRI. The researchers deceived participating students by telling them that the focus of the study was on pulse rate and its relation to measures of student learning. Half the students were then randomly assigned to a condition in which the instructor described how to take a pulse while demonstrating on the student's wrist (instructor-touch condition). The rest of the students listened to the instructor's directions and then practiced on their own wrists (no-touch condition). Following this, students viewed a video lecture delivered by the instructor and then evaluated the quality of the instructor's performance. Those in the instructor-touch condition assigned higher ratings than did students in the no-touch group on excellence of teaching, excellence of the lecture, instructor motivation to do their best work, and positive attitudes toward the instructor.

However, taking a student's pulse may place restrictions on generalization, because it is a design element unlikely to occur in most college classrooms, with the possible exception of health-related fields. Legg and Wilson (2012) addressed this issue by distinguishing between necessary and non-necessary touch. Necessary touch requires tactile contact in order to accomplish a task, such as teaching students how to take their pulse. Non-necessary contact is physical touching that expresses concern and support, such as when an instructor touches a student on the shoulder while helping to solve a math problem. Legg and Wilson thus restricted their study's implications to situations involving necessary touch, thereby correctly limiting generalization to such conditions.

Final Judgment: Replicable Result

The final judgment about replicability depends, in part, on the previous four judgments regarding external validity. First, readers should ask whether the specified or implied generalization seems reasonable. Second, are elements of the study's design representative of those to which the researcher intends to generalize? Third, has generalization been demonstrated in instances where it should occur and instances where it shouldn't? Finally, have restrictive explanations been eliminated? The bottom line is whether the study's findings are replicable. In answering that question, one might imagine conducting the study under varying conditions: with different measures or instruments, participants of different ages, a different proportion of student gender representation, different instructors and settings, different academic disciplines, and so forth.

Conclusion

Taken together, judgments about internal and external validity build conceptual and empirical support in the chain of reasoning that reinforces a study's credibility and generalization. The judgments made about internal validity affect those made about external validity. If the explanation for the study lacks credibility, attempts to generalize are pointless. If the translation of the hypotheses into design elements is flawed, concerns about whether those elements are similar to ones in the target population are irrelevant. If no significant effect has been demonstrated, there is no possibility of generalizing effects to similar conditions. In addition, rival explanations that have not been eliminated undermine the findings. Finally, if there is no credible result, replication is unlikely. Such criteria are useful for evaluating the findings reported in quantitative SRI research. Before reaching a conclusion about a study touted in the higher-education press, it is helpful to consider such criteria when reading a primary source. Doing so will enable one to judge whether claims made about the validity and/or reliability of SRI are well-founded.

Author Biographies

Steve Benton is Senior Research Officer at the IDEA Center where, since 2008, he has led a research team that designs and conducts reliability and validity studies for IDEA products. He received his Ph.D. in psychological and cultural studies at the University of Nebraska–Lincoln in 1983. He is a Fellow in the American Psychological Association and American Educational Research Association, as well as an emeritus professor and former chair of Special Education, Counseling, and Student Affairs at Kansas State University, where he served for 25 years. His current research focuses on best practices in faculty development and evaluation.

Dan Li is a Researcher and Data Analyst at The IDEA Center. She holds a B.A. from Huazhong University of Science and Technology, an M.A. from Marquette University, and a Ph.D. in Media, Technology, and Society from Northwestern University. Her previous research examined the social effects of online technologies, digital inequality, and parental mediation of television viewing. Her current work focuses on student ratings of instruction in higher education.

References

Almeida, I. D., Silva, J. M., & Mohring, M. M. (2019). Student evaluation of teaching effectiveness: Implications for scholars in operations management. *Repositorio ISCTE-IUL*, *2*, 13.

Alvero, A. M., Mangiapanello, K., & Valad, J. (2019). The effects of incentives, instructor motivation and feedback on faculty evaluation response rates in large and small classes. *Assessment and Evaluation in Higher Education, 44,* 501–515. https://doi.org/10.1080/02602938.2018.1521913

APA (2011). *The Publication Manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association. 2010. ISBN 978-1-4338 0562-2.

Benton, S. L., Duchon, D., & Pallett, W. H. (2013). Validity of self-reported student ratings of instruction. *Assessment & Evaluation in Higher Education, 38,* 377–389. https://doi.org/10.1080/02602938.2011.636799

Berk, R. A. (2018). Berk's law: Start spreading the news: Use multiple sources of evidence to evaluate teaching. *Journal of Faculty Development, 32,* 78–81.

Boring, A., Ottoboni, K., & Stark, P. B. (2016). Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research*. Retrieved from https://www.scienceopen.com/document/vid/818d8ec0-5908-47d8-86b4-5dc38f04b23e

Boysen, G. A., Kelly, T. J., Raesly, H. N., & Casner, R. W. (2014). The (mis)interpretation of teaching evaluations by college faculty and administrators. *Assessment and Evaluation in Higher Education, 39,* 641–656. https://doi.org/10.1080/02602938.2013.860950

Buchert, S., Laws, E. L., Apperson, J. M., & Bregman, N. J. (2008). First impressions and professor reputation: Influence on student evaluations of instruction. *Social Psychology of Education, 11,* 397–408. https://doi.org/10.1007/s11218-008-9055-1

Bunge, N. (2018, November 27). Students evaluating teachers doesn't just hurt teachers. It hurts students. *The Chronicle of Higher Education.* Retrieved from https://www.chronicle.com/article/Students-Evaluating-Teachers/245169

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Boston, MA: Houghton Mifflin.

Carrell, S. E., & West, J. E. (2010). Does professor quality matter? Evidence from random assignment of students to professors. *Journal of Political Economy, 118,* 409–432.

T: 785.320.2400
T: 800.255.2757

301 South Fourth St., Ste. 200
Manhattan, KS 66502

Email: info@IDEAedu.org
IDEAedu.org

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

Falkoff, M. (2018, April 25). Why we must stop relying on student ratings of teaching. *The Chronicle of Higher Education.* Retrieved from https://www.chronicle.com/article/Why-We-Must-Stop-Relying-on/243213?cid=rclink

Flaherty, C. (2016, January 11). Bias against female instructors. *Inside Higher Education.* Retrieved from https://www.insidehighered.com/news/2016/01/11/new-analysis-offers-more-evidence-against-student-evaluations-teaching

Krathwohl, D. R. (2009). *Methods of Educational & Social Science Research: An Integrated Approach* (3rd Ed.). New York: Longman.

Lawrence, J. W. (2018). Student evaluations of teaching are not valid. *Academe,* May–June. Retrieved from https://www.aaup.org/article/student-evaluations-teaching-are-not-valid#.XQEdp9NKg6h

Legg, A. M., & Wilson, J. H. (2012). Instructor touch enhanced college students' evaluations. *Social Psychology of Education, 16,* 317–327. https://doi.org/10.1007/s11218-012-9207-1

MacNell, L., Driscoll, A., & Hunt, A. N. (2014). What's in a name: Exposing gender bias in student ratings of teaching. *Innovative Higher Education*. https://doi.org/10.1007/s10755-014-9313-4.

McDonnell, G. P., & Dodd, M. D. (2017). Should students have the power to change course structure? *Teaching of Psychology, 44*, 91–99. https://doi.org/10.1177/0098628317692604

Mulhere, K. (2014, December 10). Study finds gender perception affects evaluations. *Inside Higher Education*. Retrieved from https://www.insidehighered.com/news/2014/12/10/study-finds-gender-perception-affects-evaluations

Nulty, D. D. (2008). The adequacy of response rates to online and paper surveys: What can be done? *Assessment & Evaluation in Higher Education, 33,* 301–314. https://doi.org/10.1080/02602930701293231

Ryalls, K. R., Benton, S. L., Barr, J., & Li, D. (2016). *Response to "Bias against female instructors."* (Editorial Note No. 2). Manhattan, KS: The IDEA Center. Retrieved from http://ideaedu.org/research-and-papers/editorial-notes/response-to-bias-against-female-instructors/

Sohr-Preston, S. L., Boswell, S. S., McCaleb, K., & Robertson, D. (2016). Professor gender, age, and “hotness” in influencing college students’ generation and interpretation of professor ratings. *Higher Learning Research Communications, 6*(3). Retrieve from https://doi.org/10.18870/hlrc.v6i3.328

Tobias, S. (1976). Achievement treatment interactions. *Review of Educational Research, 46*(1), 61–74. https://doi.org/10.3102/00346543046001061

Veldkamp, L. S., Nuijten, M. B., Dominquez-Alvarez, L., van Assen, M. A. L. M., & Wicherts, J. M. (2014). Statistical reporting errors and collaboration on statistical analyses in psychological science. *PLoS ONE, 9*(12), e114876. Retrieved from https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0114876

Wilson, J. H., Beyer, D., & Monteiro, H. (2014). Professor age affects student ratings: Halo effect for younger teachers. *College Teaching, 62,* 20–24. https://doi.org/10.1080/87567555.2013.825574

Young, K., Joines, J., Standish, T., & Gallagher, V. (2018). Student evaluations of teaching: The impact of faculty procedures on response rates. *Assessment & Evaluation in Higher Education, 44*, 37-49. Retrieved from http://doi.org/10.1080/02602938.2018.1467878