IDEA Editorial Note #3 • Response to "Zero Correlation Between Evaluations and Learning "

Ken Ryalls, President • Steve Benton, Senior Research Officer • Dan Li, Research Associate
The IDEA Center

We are writing in response to Colleen Flaherty's recent article published in *Inside Higher Education*, "Zero Correlation Between Evaluations and Learning," (https://www.insidehighered.com/news/2016/09/21/new-study-could-be-another-nail-coffin-validity-student-evaluations-teaching) a conclusion she reached after reading "Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related," published in *Studies in Educational Evaluation*, by Bob Uttl, Carmela A. White, and Daniela Wong Gonzalez (in press).

At the outset, we should state that two of us are experienced higher education faculty members who have sat on both sides of the desk when it comes to evaluation of teaching. As former teachers and administrators, we know the angst associated with student ratings of instruction (SRI) class reports, the disappointment of getting lower scores than we had hoped for, the anger when we get feedback we consider blatantly hostile or unfair, and the frustrations colleagues have expressed to us about their own ratings. And we admittedly represent a position that values proper use of SRI in formative and summative evaluation of teaching effectiveness. Our experiences as educators and researchers convince us that students (a) have something valuable to say that can be used by faculty to improve teaching, (b) generally assign higher ratings to teachers who have higher achievement standards (Benton, Guo, Li, & Gross, 2013), and (c) provide accountability for faculty as they go about their day-to-day teaching activities. Using SRI data wisely in order to improve teaching, and using SRI data carefully as part of a holistic analysis when evaluating faculty remain legitimate avenues. Student voice matters, because students are an important part of the teaching/learning dynamic, and to argue that we should ignore the perspective of the student altogether when analyzing instructor effectiveness in the classroom is absurd.

The blanket assertion of the uselessness of SRI is problematic, because enormous amounts of literature dedicated to SRI are discounted in one fell swoop by an over-reliance on the latest meta-analysis. Psychometric tools are fraught with so many human-error problems as to make comparison of one to another very difficult. So to say "SRI are useless" is about as helpful as saying "SRI are infallible" – neither statement is true, and completely obscures the more helpful approach of asking "Under what conditions, using what instrument, asking what questions, can SRI data be helpful in illuminating instructor performance in the classroom?" Moreover, although the authors are quick to point out the shortcomings of SRI, could not the low correlations also be due to the poor quality and lack of uniformity in the measures of learning used in the studies analyzed?

Analysis of the Critique of Previous Meta-Analytic Studies

Uttl et al. argue that previous meta-analyses of multisection studies (Clayson, 2009; Cohen, 1981; Feldman, 1989) of the relationship between SRI and student learning are flawed for a number of reasons. Clayson (2009) apparently neglected to include key studies that were included in previous meta-analyses. Moreover, Uttl et al. make the case that the moderate positive correlations reported by Cohen (1981) and Feldman (1989) were an artifact of small study size effects. The authors present funnel plots to show the SRI/learning correlations are a function of study size, whereby some small size studies resulted in very high correlations, and some large size studies reported small or no correlations. Finally, the previous studies failed to remove outliers.

Uttl et al., therefore, took on the task of retrieving the original studies and including all multisection studies to date. They then conducted new meta-analyses that took into account the effect of small section studies, examined whether the SRI/learning correlations were smaller in studies that controlled for student prior learning/ability, and removed outliers. Their analyses revealed: a) small study size effects where small studies often reported higher correlations, and b) lower correlations when prior student learning/ability was controlled. They concluded that "the multisection studies do not support the claims that students learn more from more highly rated professors" (p. 18).

However, with meta-analysis the assumptions one makes and the approaches one takes can lead to different results. That's why Cohen, Feldman, and Uttl et al. analyzed the same data but came to different conclusions. As David Freedman (2010, p. 37), the renowned Berkeley statistician, wrote, "meta-analysis would be a wonderful method if the assumptions held. However, the assumptions are so esoteric as to be unfathomable and hence immune from rational consideration." While we are not as cynical as Freedman, one meta-analysis should never be used as the only data point for an argument.

Moreover, the reliability of funnel plots has been challenged (e.g., Lau, Ioannidis, Terrin, Schmid, & Olkin, 2006):

> Strong empirical evidence exists that the appearance of the plot may be affected by the choice of the coding of the outcome (binary versus continuous), the choice of the metric (risk ratio, odds ratio, or logarithms thereof), and the choice of the weight on the vertical axis (inverse variance, inverse standard error, sample size, etc.)… Even in the unlikely event that agreement is reached on what metric and what expression of weight to use on the axes, enormous uncertainty and subjectivity remains in the visual interpretation of the same plot by different researchers (p. 598).

This is not meant to belittle the valuable work Uttl et al. have done, which will be included in our future analyses of student ratings literature. But doubt inherent in both funnel plots and meta-analyses should give pause to anyone who argues as broadly as IHE does about the impact of this one study on the veracity of SRI. Further doubt is added to IHE's hasty conclusion when one considers the other issues that concern us in the Uttl et al. study, as described in the following sections.

Comments on Uttl et al.'s Criticism of the Methodological Flaws in the Previous Meta-analyses

Uttl et al. emphasized the necessity of comprehensiveness and accuracy of information taken from primary studies for valid meta-analyses, and pointed out that previous meta-analyses (Clayson, 2009; Cohen, 1981, 1982, 1983; Dowell & Neal, 1982; Feldman, 1989; McCallum, 1984) are lacking in this aspect. While data accuracy is indeed the foundation for valid research, Uttl et al.'s criticism that search strategies used in previous meta-analyses were unrepeatable is an overgeneralization. Take the example of Cohen (1981), who clearly described the procedures to locate the studies he used, including the three databases he searched, the keywords used, and snowballing bibliographies in search results from the previous steps. Cohen also reported the number of titles as the results of his initial search, the number of articles he deemed to be relevant, which implies he read the full texts, and the final number of studies he included in the analysis. Although Uttl et al. described similar search strategies (without reporting the number of articles in each step), his statement that "the identified articles were hand searched for relevance" (p. 12) is vague, and is itself difficult, if not impossible, to replicate.

Uttl et al. claimed the unrepeatable search strategies account for the discrepancies in the number of studies across the previous meta-analyses. We found Uttl et al.'s statement to be valid to some extent, for instance the discrepancies between Dowell and Neal (1982) and Cohen (1983). However, he neglected to note that the previous meta-analyses used different inclusion criteria for primary studies, tailored to their specific and individual research purposes. After reviewing the method section in those meta-analyses, we considered it legitimate for meta-analyses with a more focused purpose to include only a fraction of primary studies that were contained in more comprehensive meta-analyses. Why would anyone expect, as Uttl et al. do, that Cohen's (1982) review of student ratings validity studies in psychology courses should contain the same number of studies as his work in 1981, which covered multiple disciplines such as psychology, education, and science? The expectation that all meta-analyses should include the same primary studies regardless of their purposes is unrealistic, and does not by itself constitute a valid criticism of prior work.

The accuracy of Uttl et al.'s claim of other methodological flaws in the previous meta-analyses is questionable given some simple fact-checking. For example, Table 1 in Uttl et al. shows five as the numbers of articles and studies reported in Dowell and Neal (1982) for each. A review of Dowell and Neal (1982) and Cohen's subsequent analysis in 1983 reveals six articles and seven multisection courses. Moreover, although Table 1 presents all the seven previous meta-analyses to be "No" in terms of *Individual studies identified*, all of them actually do include a bibliography of studies analyzed.

Comments on Uttl et al.'s Up-to-Date Meta-Analysis

Even though we appreciate Uttl et al.'s endeavor to conduct an up-to-date comprehensive meta-analysis, there are several issues in their work that warrant great caution when interpreting the results.

While Uttl et al. specified six criteria for including primary studies for their meta-analysis, at least two studies in their data fail to meet at least one criterion and thus should be excluded. Prosser and Trigwell (1991) used students' responses to a question concerning learning outcomes as the learning measure and thus violated the fourth criterion by Uttl et al. that "the learning measures had to be objective, assessing the actual learning rather than students' subjective perception of their learning." (p. 12) The 36 two-semester psychology courses in Koon and Murray (1995) were taught across four academic years, with the textbook and final exams changed from year to year. According to the third inclusion criterion by Uttl et al., both the student ratings and the learning measure need to be the same within the multisection course. Therefore, the sample in Koon and Murray (1995) should be broken down into 7, 8, 12, and 9 across the four years, and their correlations should be reported respectively. However, Koon and Murray (1995) did not report such information and should be deemed ineligible for Uttl et al.'s meta-analysis.

Considering more than three decades have passed since Cohen's initial analysis (1981), we expected to see an expanded meta-analysis with more recent primary studies. Indeed, Uttl et al. analyzed 97 multisection courses from 51 articles, making it the largest meta-analysis on the topic to date. Among those, we counted 11 unique citations containing 27 new multisection courses that were not included in the past meta-analyses. Whereas it is exciting to explore the long-standing research question with newer data, we are concerned with the quality of the correlation data extracted from some of the new multisection courses in Uttl et al. After locating and reading through the full texts, we either failed to locate some of the correlations presented in Uttl et al.'s Table 2 or questioned the inclusion of some of the following.

1. Capozza (1973) wrote "the resulting correlation coefficient was .94" (p. 127) and did not specify which correlation coefficient he was reporting. Uttl et al.'s Table 2 shows -.94 for both CIS *r* and CAS *r*. When we contacted Professor Uttl, he justified reporting the negative correlation because Capozza (1973, p. 127) reported "The results indicate that every 10% increase in amount learned reduced the professor's rating by half a point." In truth, this was a one-page summary of a study with inadequate information, and Uttl et al. should have excluded it, as Cohen (1981) did. Interestingly, Uttl et al. criticized previous meta-analyses for including "impossibly high" correlations, but chose to report the -.94 in their study. Incidentally, the authors reported many times that correlations of .8 or higher are "impossibly high." Correlations greater than 1 would be impossibly high, but correlations less than 1, no matter how close they are to 1, would not be impossibly high.

2. We found four discrepancies in the correlations reported from Galbraith and Merrill (2012a) in Uttl et al.'s Table 2. When we contacted Professor Uttl he responded the reference in their article was incorrect. The correlations reported in Table 2 of Uttl et al. actually come from Galbraith, Merrill, and Kline (2012).

3. In Study 1 of McKeachie, Lin, and Mann (1971), the authors reported correlations between a learning measure and student ratings, using data from the same multisection courses but analyzed with different strategies. The first analysis, as shown in Table 1 on p. 438, reported the correlations between the learning measure and student ratings on skill, overload, structure, feedback, interaction, and rapport (*n* = 37). McKeachie et al. later revisited the correlations, with the addition of a new learning measure and with instructor gender taken into account. Table 2 (p. 439) presents the correlations, using "Knowledge" (along with the previous "Thinking" based on the Introductory Psychology Criteria Test) as one of the learning measures and grouped by instructor gender (*n* = 34). In their Studies 2 to 5, McKeachie et al. reported correlations between student ratings and other learning measures, for male and female instructors separately.

   Uttl et al.'s meta-analysis included correlations from McKeachie et al.'s five studies by averaging the correlations across male and female instructors. In contrast, Cohen (1980) only included McKeachie et al.'s Study 1, using correlations from data where male and female instructors were pooled. We include this as another example of how researchers can take different approaches when conducting meta-analyses on the same set of studies.

4. Fenderson, Damjanov, Robeson, and Rubin (1997) only reported a range of correlations along with scatterplots. In reading that article we were unable to determine how Uttl et al. came up with the correlations they reported in Table 2. Professor Uttl reported they used a software package, although he did not recall which one it was, to derive the correlations from the scatterplots. However, the scatterplots in Fenderson et al. do not show exact

data points. So, we are baffled as to how anyone could determine exact correlations without having the actual data to analyze.

5. Drysdale (2010) is an unpublished doctoral dissertation from Utah State University. We wonder why this study was included in Uttl et al.'s meta-analysis.

In addition to these concerns about individual studies included in Uttl et al.'s meta-analysis, we are disappointed the authors did not document their coding reliability procedures and report the inter-rater reliability, as Cohen (1981) did. With three individuals authoring this meta-analysis, we expected to see such supporting details.

Finally, we acknowledge Uttl et al.'s criticism of small size effect is scientifically valid. One must, therefore, be cautious when interpreting meta-analyses based on mostly small samples. As Uttl et al. pointed out, the majority of student ratings/learning correlations reported in Cohen (1981) and Feldman (1989) were statistically nonsignificant, although Uttl et al. failed to disclose an even higher proportion of statistically nonsignificant correlations in their own meta-analysis as shown in Figure 6. Even so, is it methodologically realistic to expect the majority of multisection course studies to have a large sample size? Multisection course studies are inherently limited by the educational settings in higher education (e.g., institution size, disciplines, etc.) and some limitations cannot be overcome by researchers. The largest sample size we found in Uttl et al. was 190. However, the 190 sections were taught across nine years at one of the largest universities in the U.S., and their measure of learning was regression-adjusted grades in the subsequent advanced courses. How many researchers would have the luxury of at least 30 sections as Doyle and Whitely (1974) suggested? In fact, only seven out of the 68 multisection courses in Cohen (1981) and 15 out of the 84 multisection courses in Uttl et al. did (after excluding the seven questionable ones as we described above). Given the limited number of large-sample multisection courses studied to date, a systematic review, taking into account the differences in study features and setting, may be more informative than meta-analyses that synthesize quantities of data without differentiation.

What Can We Learn from the Uttl et al. Study?

Putting aside our critique of the methods and conclusions of the Uttl et al. study, what does it tell us? First, we know that substantial opposition to how SRI are used as a measure of teaching effectiveness can be found in higher education. We agree with the authors that it is poor practice to use SRI primarily or exclusively in evaluations, and that it is ridiculous to expect that all faculty be above average in their ratings. Student ratings of instruction should not be the only information source in decisions about teaching quality. Uttl et al.'s work supports our continued contention that *SRI are a necessary but insufficient measure of teaching effectiveness*.

But, we disagree with the authors that "universities and colleges focused on student learning may need to give minimal or no weight" to student ratings (p. 19). If that is the case, then what measures should be used? Should we use embedded assessments, which are collected during class, as evidence that students have made progress on learning objectives by performance on activities, assignments, projects, papers, and so forth? In spite of their validity, such measures are highly subjective and, in terms of reliability, pale in comparison to that of SRI (Marsh, 2007). What about standardized tests? We've seen what these have led to in elementary and secondary education. How many of us want to teach to a test, such as the Collegiate Learning Assessment (CLA)? According to Arum and Roksa (2011), on the CLA only 45% of students demonstrate any improvement in learning during the first two years of college and only 36% across four years. How would we feel if the CLA were the only measure of student learning outcomes?

A more sensible approach is to include SRI as one of several measures of teaching effectiveness. As a matter of fact, Uttl et al. reported that in a re-analysis of Cohen's data "the SET/learning correlation estimated using only the studies with 30 or more sections is .27" (p. 7). If all studies were included and weighted by sample size the correlation is .39. Correlations of approximately .3 to .4 might seem low at first glance. However, given the restricted range in most student-rating scales, and the less than perfect reliability of classroom exams, the magnitude of the correlations is meaningful. Moreover, because teachers are not the only cause of student learning, and perhaps not the most important one, one would **not** expect students' ratings of instruction to correlate perfectly with how much they learn in a course (Hativa, 2013).

The best practice, then, is to take a comprehensive approach to evaluation that assesses whether faculty peers, students, and the instructor see evidence of positive change in the classroom. These are key data sources to consider whenever we make decisions about how to improve teaching effectiveness. When combined data are in agreement, we have greater reliability (Cashin, 1996). But, each single measure has its shortcomings in validity and/or reliability, which is why multiple measures should always be used.

A second contribution of this study confirms what was already known—some student and course characteristics influence ratings. This is why IDEA statistically controls for students' work habits, desire to take the course, and more recently background preparation. In addition, course difficulty and class size can impact ratings. All of these variables are included as controls in the adjusted scores IDEA provides to level the playing field among teachers whose classes vary in size, student level and motivation, academic domain, and difficulty. Moreover, as Uttl et al. point out, ratings vary by discipline, which is why IDEA provides comparative scores by academic discipline.

A final point we wish to challenge in the Uttl et al. article is their commentary on student- determined academic standards, a commentary which really does not follow from the findings in their study:

> SETs are some sort of measurement instrument device enabling professors to find what students' perceive to be an appropriate workload and an appropriate amount to learn for specific grades, in short, an appropriate academic standard from student's perspectives…professors who are either unable to do it well [i.e., teach to student determined standards] or do not do it because they believe that such student determined academic standards are detrimental to students' themselves and/or to the society at large will get poor SETs (p. 19).

This misconception has been around since the first time SRI were collected, and, unfortunately, gets repeated frequently along with its corollary that "easier" teachers get higher SRI. The assumption that students are out for the easy "A" is insulting to students who are working hard to gain an education. In a study involving over 50,000 classes across eight academic disciplines, Centra (2003) found that the grade students expected to earn was only weakly related to SRI. Others have similarly reported low, positive correlations. However, weak positive correlations between grades and ratings may not necessarily indicate instructors are lowering standards to get higher ratings. It could well indicate that students who learn more earn higher grades and assign higher ratings, which supports the validity of SRI. A second possibility is that student characteristics, such as motivation and interest in the subject matter, could lead to greater learning and, therefore, higher grades and student ratings (McKeachie, 1997).

Evidence shows that the assertion teachers should teach to student-determined standards in order to get high ratings is not only wrongheaded but perhaps actually the inverse of the truth. If teachers really want to improve course ratings, they would do well to practice other more productive behaviors than assigning lenient grades. Challenging students, stimulating their interests (Marsh & Roche, 2000), and making appropriate changes to instruction and the course based on student feedback (Centra, 2003) are more likely than leniency to lead to higher SRI and greater student learning. Moreover, research conducted in nearly 500,000 classes across more than 300 institutions reveals that instructors are more likely to earn high SRI when their students say their teacher challenged them and had high achievement standards (Benton, Guo, et al., 2013).

In conclusion, SRI have value, and they provide data that can assist faculty in getting better at teaching. Student perspective is critical because students are the only ones who have multiple first-hand experiences of what actually occurred in the classroom. Therefore, student ratings of instruction should certainly be considered an important source of data in a comprehensive approach to the evaluation of teaching.

REFERENCES

Arum, R., & Roksa, J. (2011). *Academically Adrift: Limited Learning on College Campuses*. Chicago, IL: University of Chicago Press.

Benton, S. L., Guo, M., Li, D., & Gross, A. (2013). *Student ratings, teacher standards, and critical thinking skills*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Capozza, D. R. (1973). Student evaluations, grades, and learning in economics. *Western Economic Journal, 11*(1), 127.

Cashin, W. E. (1996). *Developing an Effective Faculty Evaluation System. IDEA Paper No. 33*. Manhattan, KS: Kansas State University, Center for Faculty Evaluation and Development.

Centra, J. A. (2003). Will Teachers Receive Higher Student Evaluations by Giving Higher Grades and Less Course Work? *Research in Higher Education, 44*(5), 495-518. doi:10.1023/a:1025492407752

Clayson, D. E. (2009). Student Evaluations of Teaching: Are They Related to What Students Learn?: A Meta-Analysis and Review of the Literature. *Journal of Marketing Education, 31*(1), 16-30. doi:10.1177/0273475308324086

Cohen, P. A. (1981). Student Ratings of Instruction and Student Achievement: A Meta-analysis of Multisection Validity Studies. *Review of Educational Research, 51*(3), 281-309. doi:10.3102/00346543051003281

Cohen, P. A. (1982). Validity of Student Ratings in Psychology Courses: A Research Synthesis. *Teaching of Psychology, 9*(2), 78-82. doi:10.1207/s15328023top0902_3

Cohen, P. A. (1983). Comment on A Selective Review of the Validity of Student Ratings of Teaching. *The Journal of Higher Education, 54*(4), 448-458. doi:10.2307/1981907

Dowell, D. A., & Neal, J. A. (1982). A Selective Review of the Validity of Student Ratings of Teachings. *The Journal of Higher Education, 53*(1), 51-62. doi:10.2307/1981538

Doyle, K. O., & Whitely, S. E. (1974). Student Ratings as Criteria for Effective Teaching. *American Educational Research Journal, 11*(3), 259-274. doi:10.3102/00028312011003259

Drysdale, M. J. (2010). *Psychometric properties of postsecondary students' course evaluations*. (Doctoral dissertation), Utah State University.

Feldman, K. A. (1989). The association between student ratings of specific instructional dimensions and student achievement: Refining and extending the synthesis of data from multisection validity studies. *Research in Higher Education, 30*(6), 583-645. doi:10.1007/BF00992392

Fenderson, B. A., Damjanov, I., Robeson, M. R., & Rubin, E. (1997). Relationship of students' perceptions of faculty to scholastic achievement: Are popular instructors better educators? *Human Pathology, 28*(5), 522-525. doi:10.1016/S0046-8177(97)90072-1

Freedman, D. A. (2010). *Statistical Models and Causal Inference: A Dialogue with the Social Sciences* (D. Collier, J. S. Sekhon, & P. B. Stark Eds.). New York, NY: Cambridge University Press.

Galbraith, C. S., Merrill, G. B., & Kline, D. M. (2012). Are Student Evaluations of Teaching Effectiveness Valid for Measuring Student Learning Outcomes in Business Related Classes? A Neural Network and Bayesian Analyses. *Research in Higher Education, 53*(3), 353-374. doi:10.1007/s11162-011-9229-0

Hativa, N. (2013). *Student ratings of instruction: Recognizing effective teaching*. Oron Publications.

Koon, J., & Murray, H. G. (1995). Using Multiple Outcomes to Validate Student Ratings of Overall Teacher Effectiveness. *The Journal of Higher Education, 66*(1), 61-81. doi:10.2307/2943951

Lau, J., Ioannidis, J. P. A., Terrin, N., Schmid, C. H., & Olkin, I. (2006). The case of the misleading funnel plot. *BMJ, 333*(7568), 597-600.

Marsh, H. W. (2007). Students’ Evaluations of University Teaching: Dimensionality, Reliability, Validity, Potential Biases and Usefulness. In R. P. Perry & J. C. Smart (Eds.), *The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Perspective* (pp. 319-383). Dordrecht: Springer Netherlands.

Marsh, H. W., & Roche, L. A. (2000). Effects of grading leniency and low workload on students' evaluations of teaching: Popular myth, bias, validity, or innocent bystanders? *Journal of Educational Psychology, 92* (1), 202-228. doi:10.1037/0022-0663.92.1.202

McCallum, L. W. (1984). A meta-analysis of course evaluation data and its use in the tenure decision. *Research in Higher Education, 21*(2), 150-158. doi:10.1007/BF00975102

McKeachie, W. J. (1997). Student ratings: The validity of use. *American Psychologist, 52*(11), 1218-1225. doi:10.1037/0003-066X.52.11.1218

McKeachie, W. J., Lin, Y.-G., & Mann, W. (1971). Student Ratings of Teacher Effectiveness: Validity Studies. *American Educational Research Journal, 8* (3), 435-445. doi:10.3102/00028312008003435

Prosser, M., & Trigwell, K. (1991). Student evaluations of teaching and courses: Student learning approaches and outcomes as criteria of validity. *Contemporary Educational Psychology, 16*(3), 293-301. doi:10.1016/0361-476X(91)90029-K

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology, 104*(3), 743-762. doi:10.1037/a0027627

Uttl, B., White, C. A., & Gonzalez, D. W. (in press). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation*. doi:10.1016/j.stueduc.2016.08.007

T: 800.255.2757
T: 785.320.2400

301 South Fourth St., Suite 200
Manhattan, KS 66502-6209
E: info@IDEAedu.org
IDEAedu.org