

The following are some very recently published articles that might be of interest:

Dennis E. **Clayson**, "Initial Impressions and the Student Evaluation of Teaching," *Journal of Education for Business*, 88:1 (2013) 26-35. DOI 10.1080/08832323.2011.633580 [Abstract](#)

Do first impressions influence the final evaluations given in a class? The author looked at the initial student perceptions and conditions of a class and compared these with conditions and evaluations 16 weeks later at the end of the term. It was found that the first perceptions of the instructor and the instructor's personality were significantly related to the evaluations made at the end of the semester. Implications for the validity of and utilization of the student evaluation of instruction are discussed.

Linda B. **Nilson**, "Time to Raise Questions about Student Ratings," in: James. E. Groccia and Laura Cruz, eds. *To Improve the Academy: Resources for Faculty, Instructional and Organizational Development* (Jossey-Bass, 2012), Chapter 14.

Portions available at:

<http://books.google.com/books?hl=en&lr=&id=dipWZb3AUJQC&oi=fnd&pg=PA213&dq=linda+nilson&ots=qnyCuVvRz7&sig=nZ7STaU6NTb2Hrgrv-f26RZxapM#v=onepage&q=linda%20nilson&f=false>
<http://www.clemson.edu/OTEI/about/leadership.html>

Relying heavily but not exclusively on Clayson's research, Nilson also finds clear evidence that students' evaluations not only fail to measure learning but are often negatively correlated with it; are more biased than in the past; are often inaccurate about simple factual matters or are dishonest; and are at best measures of student satisfaction with their experience in a course. The latter, she says, is why colleges and universities use them. In correspondence, Nilson has remarked, "I'm frankly surprised that no one has sued a college or university for being denied reappointment, tenure, or promotion based on student ratings misinterpreted and misused in the typical ways that they are. I'm sure the suit would be winnable."

Stephen R. **Porter**, "Do College Student Surveys Have Any Validity?" *The Review of Higher Education* Fall 2011, Volume 35, No. 1, pp. 45-76

"In this article, I argue that the typical college student survey question has minimal validity and that our field requires an ambitious research program to reestablish the foundation of quantitative research on students. Our surveys lack validity because (a) they assume that college students can easily report information about their behaviors and attitudes, when the standard model of human cognition and survey response clearly suggests they cannot, (b) existing research using college students suggests they have problems correctly answering even simple questions about factual information, and (c) much of the evidence that higher education scholars cite as evidence of validity and reliability actually demonstrates the opposite. I choose the National Survey of Student Engagement (NSSE) for my critical examination of college student survey validity"

Stephen R. **Porter**, Corey Rumann, and Jason Pontius, "The Validity of Student Engagement Survey Questions: Can We Accurately Measure Academic Challenge?" *New Directions for Institutional Research*, no. 150, Summer 2011, pp. 87-98 DOI: 10.1002/ir.391

"This chapter examines the validity of several questions about academic challenge taken from the National Survey of Student Engagement. We compare student self-reports about the number of books assigned to the same number derived from course syllabi, finding little relationship between the two measures."

Stephen R. **Porter**, "Self-Reported Learning Gains: A Theory and Test of College Student Survey Response,"

Research in Higher Education (Nov 2012) DOI 10.1007/s11162-012-9277-0

Abstract Recent studies have asserted that self-reported learning gains (SRLG) are valid measures of learning, because gains in specific content areas vary across academic disciplines as theoretically predicted. In contrast, other studies find no relationship between actual and self-reported gains in learning, calling into question the validity of SRLG. I reconcile these two divergent sets of literature by proposing a theory of college student survey response that relies on the belief-sampling model of attitude formation. This theoretical approach demonstrates how students can easily construct answers to SRLG questions that will result in theoretically consistent differences in gains across academic majors, while at the same time lacking the cognitive ability to accurately report their actual learning gains. Four predictions from the theory are tested, using data from the 2006–2009 Wabash National Study. Contrary to previous research, I find little evidence as to the construct and criterion validity of SRLG questions.

Stephen R. **Porter** "Using Student Learning as a Measure of Quality in Higher Education," (2012)

"The purpose of this paper is to review existing measures of student learning, and to explore their strengths and weaknesses as a quality metric for higher education."

http://www.hcmstrategists.com/contextforsuccess/papers/PORTER_PAPER.pdf

The objective of the Context for Success project was to ask scholars of higher education to weigh in on the issues – both theoretical and practical – that need to be considered in designing “input-adjusted metrics” for judging the effectiveness of postsecondary institutions. With the support of the Bill & Melinda Gates Foundation, the consulting firm HCM Strategists invited a number of scholars from around the country to write papers that would discuss the methodological issues in accounting for differences in student populations when evaluating institutional performance. In some cases, these authors were also asked to demonstrate the effects of such adjustments using actual data.

<http://www.hcmstrategists.com/contextforsuccess/papers.html>

[In the latter paper, Porter makes a point seldom remarked upon: almost all of the validity research on the self-reported behaviors, self-reported learning gains and direct measures of student learning that he discusses (e.g., the CLA, relied on in Arum and Roska, *Academically Adrift: Limited Learning on College Campuses*) has "been conducted by researchers heavily involved with the organizations that are designing, marketing and administering these surveys and tests. ... Unlike the field of medicine, which has openly struggled with issues surrounding research funded by drug companies and doctors recommending procedures using medical

devices created by their own companies, the field of postsecondary research has largely ignored this topic."]

Porter is a College of Education "insider" who joined the faculty at NCSU in 2011. (I've never met him.) Some of the work of Nicholas A. **Bowman** is also useful in evaluating student self-reports. <http://www.bgsu.edu/colleges/edhd/directory/hesa/page97703.html>

B.D. McCullough and Darren **Radson**, "Analysing Student Evaluations of Teaching: Comparing Means and Proportions," *Evaluation & Research in Education* Vol. 24, No.3, September 2011, 183-202

Abstract Student Evaluations of Teaching (SETs) play a central role in modern academia. They are used for tenure, promotion, teaching improvement and other important decisions. One would think that the data collected from a SET would be analysed correctly, but such is typically not the case, as can be seen in this study later. Therefore we propose a correct method for analysing SET data. The present paper compares the two methods on a large data set of actual SETs. We show that the traditional method can misrepresent a teacher's performance, and that the traditional method can be extremely sensitive to outliers; neither of these characteristics is desirable. In contrast, the proposed method appears to suffer from neither of these defects.

Some who use survey results in their research assert that responses to items with a 7-point Likert scale are reasonably analyzed as if the scale were an interval scale, but I don't know that this assertion is well justified. I'd think that it would depend rather heavily on what's being evaluated, among other things, and that a contextual, empirical justification ought to be offered. Even the use of "sliders" might not yield behavior best interpreted using interval statistics. Nor do many student evaluation forms present more than a 5-point Likert scale. [I think that these assertions are consistent with what's argued in:

Geoff **Norman**, "Likert scales, levels of measurement and the 'laws' of statistics," *Advances in Health Sciences Education*, Volume 15, Number 5 (December 2010) 625-632. <http://dx.doi.org/10.1007/s10459-010-9222-y>]

The widely cited Marks, Fairris, & Beleche, 2010, has been published as:

Trinidad **Beleche**, David Fairris, Mindy Marks, "Do course evaluations truly reflect student learning? Evidence from an objectively graded post-test," *Economics of Education Review* 31 (2012) 709-719.

<http://dx.doi.org/10.1016/j.econedurev.2012.05.001>

Abstract It is difficult to assess the extent to which course evaluations reflect how much students truly learn from a course because valid measures of learning are rarely available. This paper makes use of a unique setting in which students take a common, high-stakes post-test which is centrally graded and serves as the basis for capturing actual student learning. We match these student-specific measures of learning to student-specific course evaluation scores from electronic records and a rich set of student-level covariates, including a pre-test score and other measures of skills prior to entering the course. While small in magnitude, we find a robust positive, and statistically significant, association between our measure of student learning and course evaluations.

The association is indeed small. Nor do Beleche *et al* speculate on the possible effect of the remedial nature of the courses and the characteristics of the students who are required to take them. Valen Johnson (*Grade Inflation*, 106) reports that the weakest students at Duke were less inclined to attribute their poor performance to instructors than students who were stronger. Duke University admits about 20% of its applicants, UC/Riverside about 85%; a student in the 25th percentile in the freshman class at Duke would have an SAT score (Q+V) over 1300; at UC/Riverside, a student in the 25th percentile in its freshman class would have a score under 1000. It is reasonable to wonder whether the UC/Riverside students studied by Marks et al are more like the weakest Duke students in the way Johnson describes. In 2010, I wrote to Professor Marks about this but have as yet received no reply.

An apt closing:

“The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor.”

Donald T. Campbell, “Assessing the Impact of Planned Social Change” December 1976 Paper #8 Occasional Paper Series (Reprinted with permission of The Public Affairs Center, Dartmouth College), page 49.

[cited in Douglas Harris, *Value-Added Measures in Education: What Every Educator Needs to Know* (Harvard Education Press, 2010), page 246n6.]

Harris' book is about pre-college education. I expect that value-added measures in higher education will become increasingly popular.

I hope that some of the above is useful.

David Austin.
North Carolina State University