Memorandum

To: Executive Committee of the University Academic Senate

From: Brian Lakey, Professor of Psychology

De Re: IDEA student teaching evaluation instrument

Date 10/6/13

This memo provides a *very brief* review of how the evidence for the IDEA student evaluation of teaching instrument compares with other similar instruments. First, I review the evidence typically provided for student evaluations of teaching. Next, I review the evidence for the IDEA scale. I conclude that the evidence for the IDEA scale is weak at best. Finally, I present information regarding my qualifications to have this opinion.

Typical evidence for measures of student evaluations of teaching

There is a large and sophisticated body of research on student evaluations of teaching. To address the extent to which the measures reflect student learning, the standard method is a multi-section validity study. Such a study investigates single courses in which there are multiple sections, each taught by a different professor, but with standardized textbooks, syllabi and exams. Professors are the units of analysis, as each professor has scores that reflect the average across all students in a section. This is typically how scores are reported at GVSU. A good example of a multi-section validity study is Marsh and Overall (1980), in which students were randomly assigned to one of 31 sections of a computer programming course. All sections had a common syllabus, text and exams. Student evaluations forecasted student learning as assessed by exams (r= .38*). That is, professors with higher teaching evaluation scores had students who performed better on the final exam.

There are enough multi-section validity studies to support meta-anal*y*ses in which the results of all high-quality studies are averaged. For example, Feldman (1989) reported that the average correlation between a single-item rating of professors' teaching effectiveness and standardized class exam scores was r .39*. Again, professors who elicited favorable ratings from students had students who performed better on exams. Thus, it is well-established that many different measures of student evaluations of teaching forecast actual learning. Nonetheless, the magnitudes of these effects are moderate. A professor whose teaching evaluations are a standard deviation above the mean of her or his peers, can be expected to produce about 1/3 a standard deviation improvement in students' learning.

Evidence for the validity of the IDEA scales

In reviewing extensive technical reports provided by IDEA representatives, the most striking finding was that IDEA has not conducted any multi-section validity studies of the kind just described. This conclusion was confirmed in an email by Steve Benton, the Senior Research

Officer at the IDEA Center. The best evidence regarding the IDEA scale was provided by Benton et al., (2011). In a single course with a single professor, students who rated their learning as higher on IDEA scales performed better on the exams. Yet, this study was deeply flawed as students were the unit of analysis, rather than professors. Thus, the results might say more about the students than about the professor. For example, it might mean that students who characteristically rate professors favorably also characteristically score well on exams, perhaps reflecting intelligence or interest in knowledge. Studies in which professors are the unit of analysis are not vulnerable to this criticism as student characteristics are averaged out.

According to IDEA documentation (IDEA, 2010), the strongest evidence for the validity of student ratings of learning was shown by the correspondence between faculty members' ratings of the importance of learning objectives, and students' ratings of their own learning on those same objectives. The average correlation between the two was $r = .20$ (Table 15), which indicates a 4% correspondence between what professors say they emphasize and what students say they learn. Keep in mind that students' ratings of learning are imperfect indicators of learning measured by exams. Thus, the link between what professors emphasize and what students learn would be very much smaller.

A distinguishing aspect of the IDEA system is that it emphasizes students' ratings of their learning rather than only professors' teaching. Yet, students' ratings of learning were very highly correlated with students' perceptions of professors' teaching. For example, ratings of learning were correlated at $r \approx .80$ with summary judgments of "excellent course" and "excellent teacher" (IDEA, 2010). Thus, student ratings of learning provide little additional information beyond summary judgments of professors' teaching. In conclusion, although the purveyors of the IDEA system claim that a unique and valuable feature of their measure is that student learning is rated instead of professors' teaching; both types of ratings yield essentially the same information. Thus, the IDEA system appears to provide the same information as other measures, but without similar evidence for its validity. In my opinion, it would be a mistake to adopt a scale with such poor evidence for its validity.

Finally, the reader might be interested in my qualifications to have these opinions. My vita is posted at http://gvsu.edu/psychology/brian-lakey-126.htm and manuscripts include one peer-reviewed publication on student teaching evaluations and a second manuscript under review. I want to emphasize, though, that teaching evaluations are only a side interest about which I do not claim real expertise.

References

Benton, S. L., Webster, R., Gross, A.B., & Pallett, W. H. (2010). Technical Report No. 15. An Analysis of IDEA Student Ratings of Instruction in Traditional Versus Online Courses. Manhattan, KS: The IDEA center.

Benton, S. L., Duchon, D., & Pallett, W. H. (2011). Validity of student self-reported ratings of learning. *Assessment and Evaluation in Higher Education. DOI:10.1080/02602938.2011.636799*

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*, 583-645.

Marsh, H. W. & Overall, J. U. (1980). Validity of students' evaluations of teaching effectiveness: cognitive and affective criteria. *Journal of Educational Psychology*, *72*, 468-475.