ETS SIR II

Listening. Learning. Leading.®

Testing the Invariance of Interrater Reliability Between Paper-Based and Online Modalities of the *SIR II*™ Student Instructional Report

David Klieger, John Centra, John Young,
Steven Holtzman, and Lauren J. Kotloff

May 2014

Testing the Invariance of Interrater Reliability Between Paper-Based and Online Modalities of the *SIR II*™ Student Instructional Report

David Klieger, John Centra, John Young, Steven Holtzman, and Lauren J. Kotloff
Educational Testing Service, Princeton, New Jersey

May 2014

Action Editor: Marna Golub-Smith

Reviewers: Donald Powers and Brent Bridgeman

Abstract

Moving from paper-based to online administration of an assessment presents many advantages, including cost savings, speed, accuracy, and environmental conservation. However, a question arises as to whether changing the modality of administration affects reliability and thus validity, how scores or ratings should be interpreted. We investigated whether the interrater reliability (within-class variance) for the *SIR II*™ Student Instructional Report differed between the paper-based and online versions. Our results indicated that they did not. The findings provide additional evidence that moving to an online version of the instrument does not change how one should interpret SIR II ratings.

Key words: course evaluation, teaching evaluation, teacher evaluation, college instruction, university instruction, class instruction, course instruction, instruction, interrater reliability

Acknowledgments

We wish to thank Gene Bouie, ETS's Director of College Programs, for providing the funding that made this study possible. We also wish to thank Kean University and St. Ambrose University for providing the data for this study.

The formal use of student evaluation of their teachers (SET) at the college level began in the 1920s at a handful of colleges, including Harvard University, the University of Washington, the University of Wisconsin, and Purdue University (Anderson, Cain, & Bird, 2005; Calkins & Micari, 2010; Guthrie, 1954; Haskell, 1997). With an increased demand for accountability in higher education and a view of the student as a consumer, the use of SETs became widespread in American colleges and universities in the 1960s (Centra, 1993). Since then, SETs have become a routine component of a university's quality assessment process (Anderson et al., 2005). The findings of numerous studies conducted over several decades generally support the reliability and validity of SETs (Centra, 1993; Cohen, 1981; Wachtel, 1998). More recently, SETs have taken on increased importance in decisions involving faculty tenure, reappointment, promotion, and merit pay as well as whether to continue offering particular courses (Adams & Umback, 2011; Anderson et al., 2005; Calkins & Micari, 2010; Haskell, 1997; Layne, DeCristoforo, & McGinty, 1999).

The advent of affordable personal computers in the 1970s (and their increased availability in the 1980s) and the introduction of the modern Internet in the 1990s have provided increased opportunities to deliver SETs, collect and analyze responses, and deliver feedback by means other than traditional paper-based surveys. Since the early 1990s, numerous studies have tested the comparability of traditional paper-based SETs that students complete in class and online SETs that students access and complete outside of class time. In addition to examining differences in the number, length, and usefulness of qualitative feedback and comments, these studies primarily examine whether the two modalities differ in terms of response rate and scores for quantitative ratings.

Studies consistently report that online SETs yield significantly lower response rates than paper-based SETs (see Ardalan, Ardalan, Coppage, & Crouch, 2007; Avery, Bryant, Mathios, Kang, & Bell, 2006; Dommeyer, Baum, Hanna, & Chapman, 2004; Fike, Doyle, & Connelly, 2010; Layne et al., 1999; Sax, Gilmartin, & Bryant, 2003). The overwhelming majority of the studies have found that quantitative scores from the two modalities (typically a comparison of the means of ratings between the two modalities) are highly comparable (Dommeyer, 2004; Donovan, Mader, & Shinsky, 2006; Heath, Lawyer, & Rasmussen, 2007). A few studies found significant differences in comparisons between the two modalities, but these differences were so small as to be considered by the researchers to have little practical meaning (Avery et al., 2006;

Fike et al., 2010; Kasiar, Schroeder, & Holstad, 2002). A study by Ardalan et al. (2007) produced mixed results.

Research on differences in quantitative ratings between paper-based and online SET modalities has addressed potential confounds, with mixed findings. Kordts-Freudinger and Geithner (2013) investigated whether differences in mean scores between paper and online SETs were the result of different evaluation settings (i.e., paper surveys are completed in class while online surveys are completed out of class) rather than evaluation format (i.e., paper-based vs. online). They concluded that evaluation setting rather than format influences evaluation results. Using regression analysis to control for variables that have been shown to affect student evaluations of teachers (class size, teacher experience, student's expected grades, and level of class [freshman/sophomore vs. junior/senior]), Nowell, Gale, and Handley (2010) found significant differences between in-class versus online evaluations, with average student ratings on online surveys significantly lower than average student ratings provided in class. One of the challenges in these studies is the inability to control for many of the factors that have a statistically and practically significant influence on ratings of instructors' performance. Therefore, it is not surprising that much of the prior research on differences in quantitative ratings between paper-based and online SET modalities is inconclusive.

In terms of quantitative ratings, different modalities can potentially differ in several ways. We seek to expand understanding of how quantitative ratings on SETs can differ, particularly in critical ways that have not been previously explored. Specifically, we have been unable to locate any prior research regarding the invariance of interrater reliability of SETs across modalities. Interrater reliability, defined here as the consistency or exchangeability across third-party evaluators of their ratings of an individual, is an essential component of the validity of those ratings: In general, reliability imposes an upper limit on validity (Nunnally, 1978, p. 192). If one wishes to relate scores or ratings to an external criterion (i.e., calculate some form of criterion-related validity), then low reliability is highly problematic. At least since the beginning of the past century, the psychometric study of interrater reliability has been of great interest to researchers (see Pearson, Lee, Warren, & Fry, 1901). One hundred years ago, the field of psychological measurement adopted the modern approach of interrater reliability analyses as involving a ratio of variances (see Harris, 1913). In this study, we use this way of examining interrater reliability to investigate an important aspect of whether the online and paper-based

versions of the *SIR II*™ Student Instructional Report (SIR II) are comparable. Without such comparability, questions would arise about whether the online SIR II is related to external criteria (e.g., a measure of student learning) differently than is the paper-based SIR II. Furthermore, if the interrater reliability of the online SIR II is lower than that of the paper-based SIR II, then use of the online SIR II would be more difficult to justify. One might expect differences in interrater reliability between the paper-based and online versions of SIR II because of construct-irrelevant variance (i.e., for reasons unrelated to the construct of college-level instruction that both formats of SIR II are designed to measure). These reasons may include greater variability among students in motivation to respond to the online version in comparison to the paper-based version, substantial variability in students' familiarity or comfort with computers, and formatting of the online version of SIR II that promotes differences in the ways that students respond to items.

Demonstrating the comparability of the online and paper-based versions of SIR II would support the adoption of online evaluation by those who are reluctant to do so. It would also reassure current online users regarding their continued use of online evaluation. Cost savings, speed, accuracy, and environmental conservation are compelling reasons to adopt or retain the online SIR II in lieu of the paper-based alternative. Paper-based responses are converted into computer-based information anyway, so raters who directly enter their responses into a computer make the evaluation process more efficient. Online SIR II reduces, if not eliminates, the use of paper. Increased efficiency and reduced use of resources translates into cost savings for both the end-user and the assessment administrator and scorer. As an enhancement to the paper-based version of SIR II, the online modality provides a direct link to the appropriate sections of a compendium of actionable recommendations for improving college instruction (see Educational Testing Service, 2014). Based on feedback from educators across the United States, these recommendations are rooted in best practices (see Educational Testing Service, 2014). Notwithstanding the many benefits of an online version, lack of comparability of the online and paper-based modalities would indicate that the SIR II modalities are measuring something different or, as mentioned previously, possibly that the online version is measuring the same thing but with more error (i.e., that it is measuring more construct-irrelevant variance). This finding might deter stakeholders from moving away from the longer established paper-based method.

Hypothesis

We posed the following hypothesis: The interrater reliability for each dimension and the overall evaluation of the online version of SIR II does not differ from the corresponding dimension and overall evaluation of the paper-based version.

Methods

SIR II is a standardized SET that surveys college students for their evaluation of instruction in a college class that they are taking. SIR II (and its predecessor, SIR) have undergone development over a period of more than 40 years and have been used by more than 500 institutions (Centra, 1993). In conjunction with the aforementioned compendium, the mission underlying the SIR II instrument is to provide student feedback that can help faculty and administrators at both 2-year and 4-year colleges and universities improve teaching effectiveness and learning quality. Moreover, SIR II makes available to interested parties comparative survey data from almost eight million SIR II surveys from institutions across the United States. These data include mean ratings for 107,071 classes from 2-year institutions (more than 2.4 million students) and 117,132 classes from 4-year institutions (more than five million students; Educational Testing Service, 2014).

SIR II contains 45 standard questions in total. In addition to an overall evaluation item, SIR II consists of the following dimensions: course organization and planning (five items); communication (five items); faculty/student interaction (five items); assignments, exams, and grading (six items); course outcomes (five items); and student effort and involvement (three items). Items are Likert type, with the following 5-scale points for some of the dimensions: very effective (5), effective (4), moderately effective (3), somewhat ineffective (2), and ineffective (1). Other dimensions employ the following 5-point scale instead: **much more** than most courses (5), **more than** most courses (4), about the **same** as others (3), **less** than most courses (2), and **much less** than most courses (1). For both response scales, respondents may alternatively choose not applicable (0) for each item. Each response is directed at how the item contributed to the student's learning. Furthermore, there are three items about course difficulty, workload, and pace as well as five items about respondents' backgrounds and grade expectations (reason for taking the course, class level, gender, English proficiency, and expected grade in the course). In addition to using the standard questions and response options, faculty and administrators may supplement SIR II with additional questions that have customized response choices. (This study

analyzes only the standard questions and response options.) SIR II is available in paper-based and online formats. (A third format [e-SIR] is available for distance-learning programs.) The paper-based version of SIR II is administered in class only, where as the online version can be administered either in class or remotely. More information about SIR II can be found at http://www.ets.org/sir_ii/about (Educational Testing Service, 2014).

Our subjects consisted of two classes for each of five instructors. One instructor taught at a private, Catholic liberal arts university in the Midwestern United States. Four instructors taught at a large, public university located in the Northeastern United States. In Table 1, we report sample sizes for each SIR II modality in our analyses. Each instructor taught one class that used all paper-based SIR II and another class that used all online SIR II. There is no indication that an instructor's classes differed in ability level or other aspect that would confound interpretation of results reported below.

Table 1

Sample Sizes of Classes for Which Ratings Were Analyzed

Instructor	Course	Online	Paper	Total
1	Operations management	11	25	36
2	Corporate finance	12	23	35
3	Accounting	23	14	37
4	Marketing	15	26	41
5	Occupational therapy	15	15	30
	Total:	76	103	179

Note. Paper = paper-based. Instructors 1-4 taught at the same university. Instructor 5 taught at a different institution

Analysis and Results

For each SIR II dimension, our analysis compares the residual variance of the online SIR with that of the paper-based SIR. The residual variance is variability in ratings not attributable to the instructor (i.e., it is not variance within a class). This residual variance is one way to measure interrater reliability, where a lower residual variance would indicate higher interrater reliability. In particular, this analysis compares to an *F* distribution the ratio of the larger residual variance to the smaller residual variance. According to the *F* distribution, none of the *F* statistics were statistically significant (see Table 2). All *p* values were 0.125 or larger. These results indicate that there is no evidence of a statistically significant difference in the interrater reliability (residual variance) between the online and paper-based versions of SIR II.

Table 2

Comparisons of the Variances of Paper-Based Versus Online Modalities: Testing Ratios of Mean Square Errors (MSEs)

SIR II dimension	Online residual MSE	Online *df*	Paper residual MSE	Paper *df*	*F* statistic (ratio of MSEs)	*p* value
Course organization and planning	.471	71	.400	98	1.178	.225
Communication	.484	71	.428	98	1.130	.286
Faculty/student interaction	.723	71	.598	98	1.210	.190
Assignments, exams, and grading	.538	71	.419	98	1.284	.125
Course outcomes	.558	71	.561	98	1.007	.492
Student effort and involvement	.511	71	.532	98	1.040	.434
Overall evaluation	.709	71	.562	98	1.262	.143

Note. MSEs were obtained via one-way ANOVAs. The ratio of MSEs for the *F* statistic is always the larger residual MSE divided by the smaller residual MSE. *df* = degrees of freedom; paper = paper-based.

Discussion, Limitations, and Conclusion

Our results indicate that the consistency of ratings is invariant between the online and paper-based versions of SIR II. This finding is encouraging, because while reliability does not assure validity, if the online SIR II were less reliable than the paper-based SIR II it would be more likely that the online SIR II is less valid than the paper-based SIR II. Users and potential users of SIR II want to know that their interpretation of ratings from the online version of SIR II would be comparable to (including no less reliable than) those from the paper-based version. Users of SIR II may seek to relate those ratings to important criteria such as measures of student learning outcomes, future teaching performance, and so on.

Our research does have limitations. The number of instructors and universities in this study is rather small. It is possible that analyses based on a larger number of instructors and universities would yield different results. Furthermore, each class used a single SIR II modality (paper-based only or online only). It is possible that different classes for the same instructor differed in ways that affected the analyses. However, we did not possess data for any class that used both modalities. In addition, we do not know if the online and paper-based modalities of SIR II will significantly differ in ways not addressed in this paper. For example, we do not know whether means of ratings will differ or not between the two versions. That determination requires additional investigation. Although the results of this current study are encouraging, follow-up research is warranted.

References

Adams, M. J. D., & Umback, P. D. (2011). Nonresponse and online student evaluations of teaching: Understanding the influence of salience, fatigue and academic environments. *Research in Higher Education, 53*, 576–-591. doi:10.1007/s11162-001-9240-5

Anderson, H. M., Cain, J. C., & Bird, E. (2005). Online student course evaluations: Review of literature and a pilot study. *American Journal of Pharmaceutical Education*, *69*(1), 34–43.

Ardalan, A., Ardalan, R., Coppage, S., & Crouch, W. (2007). A comparison of student feedback obtained through paper-based and web-based surveys of faculty teaching. *British Journal of Educational Technology*, *38*(6), 1085–1101. doi:10.1111/j.1467-8535.2007.00694.x

Avery, R. J., Bryant, W. K., Mathios, A., Kang, H., & Bell, D. (2006). Electronic course evaluations: Does an online delivery system influence student evaluations? *Journal of Economic Education, 37*(1), 21–38. doi:10.3200/JECE.37.1.21-37

Calkins, S., & Micari, M. (2010, Fall). Less-than-perfect judges: Evaluating student evaluations. *Thought and Action, 2010,* 7–22.

Centra, J. A. (1993). *Reflective faculty evaluation*. San Francisco, CA: Jossey-Bass.

Cohen, P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multi-section validity studies. *Review of Educational Research*, *51*(*3*), 281–309.

Dommeyer, C. J., Baum, P., Hanna, R. W., & Chapman, K. S. (2004). Gathering faculty teaching evaluations by in-class and online surveys: Their effects on response rates and evaluations. *Assessment & Evaluation in Higher Education*, *29*(5), 611–623. doi:10.1080/02602930410001689171

Donovan, J., Mader, C. E., & Shinsky, J. (2006). Constructive student feedback: Online vs. traditional course evaluations. *Journal of Interactive Online Learning, 5*(3), 283–296.

Educational Testing Service. (2014, January 8). *About the SIR II Student Instructional Report*. Retrieved from http://www.ets.org/sir_ii/about

Fike, D. S., Doyle, D. J., & Connelly, R. J. (2010). Online vs. paper evaluations of faculty: When less is just as good. *The Journal of Effective Teaching*, *10*(2), 42–54.

Guthrie, E. R. (1954). *The evaluation of teaching: A progress report*. Seattle: University of Washington.

Harris, J. A. (1913). On the calculation of intra-class and inter-class coefficients of correlation from class moments when the number of possible combinations is large. *Biometrika, 9*(3/4), 446–472.

Haskell, R. E. (1997). Academic freedom, tenure, and student evaluation of faculty: Galloping polls in the 21st century. *Education Policy Analysis Archives*, *5*(6), 1–32.

Heath, N. M., Lawyer, S. R., & Rasmussen, E. B. (2007). Web-based versus paper-and-pencil course evaluations. *Teaching of Psychology, 34*(4), 259–261. doi:10.1080/00986280701700433

Kasiar, J. B., Schroeder, S. L., & Holstad, S. G. (2002). Comparison of traditional and web-based course evaluation processes in a required, team-taught pharmacotherapy course. *American Journal of Pharmaceutical Education, 66,* 268–270.

Kordts-Freudinger, R., & Geithner, E. (2013). When mode does not matter: Evaluation in class versus out of class. *Educational Research and Evaluation, 19*(7), 605–614. doi:10.1080/13803611.2013.834613

Layne, B. H., DeCristoforo, J. R., & McGinty, D. (1999). Electronic versus traditional student ratings of instruction. *Research in Higher Education, 40*(2), 221–232.

Nowell, C., Gale, L. R., & Handley, B. (2010). Assessing faculty performance using student evaluations of teaching in an uncontrolled setting. *Assessment & Evaluation in Higher Education, 35*(4), 463–475. doi:10.1080/02602930902862875

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.

Pearson, K., Lee, A., Warren, E., & Fry, A. (1901). Mathematical contributions to the theory of evolution. IX. On the principle of homotyposis and its relation to heredity, to the variability of the individual, and to that of the race. Part I. Homotypos in the vegetable kingdom. *Philosophical Transactions of the Royal Society London, 197*, 285–379.

Sax, L .J., Gilmartin, S. K., & Bryant, A. N. (2003). Assessing response rates and nonresponse bias in web and paper surveys. *Research in Higher Education, 44*(4), 409–431.

Wachtel, H. K. (1998). Student evaluation of college teaching effectiveness: A brief review. *Assessment and Evaluation in Higher Education, 23*(2), 191–212.