Measuring differences in pencil-and-paper and online course evaluations

A Rasch measurement approach to analyzing differences in pencil-and-paper and online formats for higher education course evaluations

Leslie A. Sweeney¹, University of Kentucky
Kathryn Shirley Akers, University of Kentucky
Kelly D. Bradley, University of Kentucky

¹ Please use Leslie A. Sweeney as the contact author of contact. (145 Taylor Education Building, University of Kentucky, Lexington, KY 40506, laswee3@uky.edu)
Abstract
The purpose of this study was to utilize the Rasch model to identify differences and/or similarities between pencil-and-paper and online student course evaluations at a Midwestern I Research University. Pencil-and-paper course evaluations were collected from 55 advanced undergraduate and graduate students in fall 2008. Online course evaluations were collected from 49 advanced undergraduate and graduate students enrolled in fall 2009. In both cases, the students were enrolled in an introductory statistics course taught at the same day/time by the same instructor. A polytomous Rasch model was utilized to analyze the data with Winsteps. Results demonstrate pencil-and-paper and online course evaluations function analogously for categories and person endorsability, but differently for course and instructor items, as well as learning outcomes.
Measuring differences in pencil-and-paper and online course evaluations

A Rasch measurement approach to analyzing differences in pencil-and-paper and online formats for higher education course evaluations

Since the development and proliferation of the internet during the past two decades, higher education institutions have been considering and/or using web based course evaluation systems in place of traditional paper and pencil collection methods. Such development and usage has spawned numerous studies examining the differences and similarities of online and paper and pencil delivery modes for surveys and evaluations within all domains, including higher education. Myriad studies of student satisfaction, student and faculty perceptions, response rates, response bias, frequency and length of student comments, timeliness, cost effectiveness, and the like have been conducted.

However, most, if not all, have employed various Classical Test Theory (CTT) methods, such as t-tests, chi-square, ANOVA, and MANOVA, but not the Rasch model. A latent-trait IRT model, the Rasch model provides a means to not supply basic aggregate data for course evaluations, such as in CTT methods, but to supply information about how an evaluation’s respondents and items are functioning and, in turn, the reliability and validity of aggregate student satisfaction and teacher effectiveness results. Hence, this study will focus on identifying differences and/or similarities between a web based and traditional paper and pencil course evaluation via the Rasch model to determine if course evaluations may function differently online than on paper.

Literature Review

Higher education institutions have utilized paper based surveys for decades to evaluate teacher effectiveness and student satisfaction with courses. With the proliferation of the internet within the past decade, institutions are entertaining the notion of evaluating courses online.
Hmieleski (2000), as noted in Hoffman (2003), found 98% of the nation’s 200 most wired institutions were still predominantly collecting data via paper and pencil based surveys. However, Hoffman (2003) found 10% of 500 randomly selected colleges and universities were primarily using the internet for collecting student evaluation data. Also, he discovered 56% of institutions were already using the Internet for evaluating online courses, 10% would be implementing an Internet ratings process in 2003, 18% were reviewing this as an option for online courses, and 16% decided not to use the Internet to evaluate online courses (Hoffman, 2003). Thus, it is evident America’s higher education institutions consideration of trading paper and pencil evaluations for online web links is rising.

As with a rise in any new phenomena, differences between the old and the new (web based versus paper and pencil course evaluations) have been evaluated via various scholars from numerous angles. Several scholars have discovered potential response influences for course evaluations in general. For instance, Layne, DeCritofo, & McGinty (1999) discovered academic area can affect evaluation results, while Liegle & McDonald (2005) noted pro-exam and high performing student respondents’ evaluated instructors higher than pre-exam and low performing student respondents.

Student satisfaction with the web based delivery mode of course evaluations has also been continuously examined. For instance, Anderson, Cain, & Bird (2006), Kasiar, Schroeder, & Holsted (2002), Layne et al. (1999), and McGourty, Scoles, & Thorpe (2002) all found students are content and prefer the online course evaluations. McGourty, Scoles, & Thorpe (2002) also discovered students view a web based format as more convenient, since they can fill it out on their own time.
Differences between mean scores of online and paper course evaluations have been examined as well. Researchers have discovered mean scores for both modes of delivery are relatively analogous [Anderson et al. (2005); Avery et al. (2006); Dommeyer, Baum, Hanna, & Chapman (2004)]. However, Kasiar et al. (2002) noted a statistically significant difference between 29% of the evaluation questions, but most questions only had a median/range difference of one point and, in turn, she noted in most cases this did not change the overall meaning of the response. Also, Avery, Bryant, Mathios, Kang, & Bell (2006) noted online scores were slightly higher than paper and pencil results suggesting students provide more favorable ratings to instructors online versus in-class.

Numerous other areas of web based and paper and pencil course evaluations have been explored as well, such as response rates, student perceptions, timeliness, cost, comment frequency and length, and the like. Such studies are fruitful for determining differences between online and paper evaluations, but they do not pertain specifically to our study and/or results. Thus, they have been omitted here for this purpose.

Despite the plethora of research on web based versus traditional course evaluation modes, few, if any, scholars have also examined how effectively course evaluations are functioning. Most scholars have evaluated web based versus paper and pencil course evaluations via utilizing CTT and have employed various CTT methods, such as t-tests, chi-square, factor analysis, ANOVA, MANOVA, and so forth [Anderson et al. (2005); Ardalan, Ardalan, Coppage, & Crouch (2007); Avery et al. (2006); Dommeyer et al. (2004); Donovan, Mader, & Shinsky (2006); Kasiar et al. (2002); Layne et al. (1999); Lilege & McDonald (2005); McGourty, Scoles, & Thorpe (2002);]. Such analyses provide aggregate data concerning responses, but they do not
Measuring differences in pencil-and-paper and online course evaluations

examine how effective a course evaluation is in measuring teacher effectiveness and, in turn, providing reliable and valid results.

However, Bradley & Bradley (2006) have examined the reliability and validity of higher education course evaluations and legitimatized the Rasch model as a means for fully comprehending their results and functioning. A latent-trait IRT model, the Rasch model is designed to not only identify how effectively an instrument is measuring a given construct, but also the differences in how people respond and items function for a given measure. Bradley & Bradley (2006) also demonstrated CTT models only provide surface information regarding aggregate responses, standard deviations, and the like, where as the Rasch model provides more in-depth information regarding individual respondents and items, response categories, and so forth. Such in-depth information is fruitful for administrators in not only understanding the aggregate CTT data, but also in comprehending how effectively their evaluations are measuring students’ satisfaction and perceptions of instructors.

Few, if any, scholars have examined the differences in how course evaluation respondents and questions function for web based and paper and pencil evaluations. More specifically, few if any scholars have utilized the Rasch model to make such determinations. Hence, for this study we will use the Rasch model to try and identify differences and/or similarities between a web based and traditional paper and pencil course evaluation to determine if course evaluations may function differently online than on paper.

Purpose

The purpose of this study was to utilize the polytomous Rasch model to identify potential differences and similarities between pencil-and-paper and online student course evaluations for an introductory statistics course taught at a Midwestern Research I institution by the same
Measuring differences in pencil-and-paper and online course evaluations

instructor. More specifically, the goal was to demonstrate how rating scale categories, people, and survey items function similarly and/or differently between the types of format.

Perspectives/Theoretical Framework

Rasch Model

Following from the mathematical concept of invariance, the Rasch model is utilized for constructing a measure for numerous samples within varying contexts. The Rasch model, introduced by George Rasch (1960), is a one parameter, latent trait model, which allows for the examination of a construct via independent measurement of latent trait reflections in persons and item responses, but simultaneous comparisons of such measures. Rasch provides a comprehensive construct depiction via reporting individual persons and items on an interval scale, where raw scores are converted into logits (log odds). Logits are used to categorize people and items along an equivocal continuum from low to high for comparison of person ability and item difficulty probability estimates (Bond and Fox, 2007). Unlike Classical Test Theory (CTT) where researchers fit a model to the data, Rasch espouses fitting data to the model via providing individual rather than aggregate scores for people and items. Thus, Rasch accounts for anomalies within individual responses, which provides additional information about the construct under investigation (Andrich and Luo, 2003).

The polytomous Rasch model was used for this study because of its ability to measure the properties of rating scale data. With polytomous data specifically, difficulty to endorse is the main characteristic influencing responses (Linacre, 1999). People are placed along an interval continuum based on their willingness to endorse items and items are placed based on their endorsability. Rasch also “...will establish the pattern in the use of the Likert scale categories to
Measuring differences in pencil-and-paper and online course evaluations yield a rating scale structure common to all items on the scale,” (Bond and Fox, 2007, p.103). Ergo, providing detailed information regarding category use, as well as item functioning and person endorsability.

Methods

Data Sources

To measure difference between paper and online course evaluations, course evaluations for a single professor were collected from the two most recent fall semesters. Pencil-and-paper course evaluations were collected from 55 advanced undergraduate and graduate students in fall 2008. Online course evaluations were collected from 49 advanced undergraduate and graduate students in fall 2009. All students enrolled in the same introductory statistics course, offered on the same day and at the same time, by the same instructor, both in the fall term on subsequent years. The course and enrolled students were selected because the course is repeatedly taught by the same instructor, the same way. Therefore, one hypothesized differences would be more attributable to the evaluation administration method versus the instructor.

Instructions regarding the purpose of the survey and participation were read to students in class and e-mailed to students with the web link via the institutional research office, for the pencil-and-paper and online surveys, respectively. Data were then collected and sent to the institutional research office from which an electronic SPSS copy of the data files was obtained by the researcher for this study.

Instrument

The pencil-and-paper and online course evaluations were identical and included four student information questions asking for a student’s classification and reason for selecting, hours spent per week on, and expected grade in the course. Primary sections for the evaluations
Measuring differences in pencil-and-paper and online course evaluations

included course and instructor items, learning outcomes, and summary items, as well as special sections for undergraduate studies courses, seminars, and labs and discussions. The survey instrument included 36 short statements for which respondents would indicate their agreement on a 4-point Likert-type scale where 1 = strongly disagree, 2 = disagree, 3 = agree, and 4 = strongly agree. Respondents were also provided the option to indicate that the item was not applicable. A comment box for any additional respondent feedback was also provided.

Analysis

Data obtained were exported from SPSS into Excel files for clean-up. The data were then run separately via the polytomous Rasch model for analysis to allow for comparison of fit and function of the pencil-and-paper and online evaluation data. After applying the Rasch measurement model, mean square infit statistics, specifically, were utilized for analysis, since such statistics are based on weighted scores and do not allow outliers to throw off responses (Bond and Fox, 2007). Also, observed counts and percentages, average measures, thresholds, and probability curves were utilized to determine if all four categories were functioning satisfactorily for both formats. Item hierarchy maps were also examined to determine how easy or hard items were for people to endorse on the evaluations, as well as to show what items were functioning similarly and differently between the two formats of the evaluation.

Results

Category Structure

As demonstrated in Table 1, the majority of students responded positively in both the paper and pencil evaluation and the online evaluation. For the pencil and paper evaluation, 94% of students responded Agree or Strongly Agree to the evaluation questions. For the online evaluation, 94% also responded Agree or Strongly Agree to the evaluation questions. The
category measure for both evaluations also increases as the categories increase. In other words, as the level of agreement increases, strongly disagree to disagree and so forth, the category measure also increases. Therefore, most students found the statements easy to endorse, showing they mostly agreed or strongly agreed with the evaluation questions for both paper and online evaluation.

Table 1 Category Structure

<table>
<thead>
<tr>
<th>Category label</th>
<th>Paper and Pencil</th>
<th>Online</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Observed Count</td>
<td>Observed Percentage</td>
</tr>
<tr>
<td>1 (Strongly Disagree)</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>2 (Disagree)</td>
<td>50</td>
<td>5</td>
</tr>
<tr>
<td>3 (Agree)</td>
<td>417</td>
<td>40</td>
</tr>
<tr>
<td>4 (Strongly Agree)</td>
<td>573</td>
<td>54</td>
</tr>
</tbody>
</table>

*Item Heirarchy*

Figures 1 and 2 are visual displays of item and person difficulty for each evaluation. Item difficulties can be found on right side of the mapping. In this example item difficulty is a measure of the difficulty of endorsement for an item. An item placed above another item is said to be more difficult to endorse than the lower item. Person difficulty, found on the left side of the mapping, is a measure of the individual’s level of agreement with the items. A person higher on the chart would find an item easier to endorse, or more likely to agree with the item, than a person who was placed lower on the figure. Figure 1 is the variable mapping for the paper and pencil evaluation and Figure 2 is the variable mapping for the online evaluation.
Measuring differences in pencil-and-paper and online course evaluations

Figure 1
Variable mapping of Paper and Pencil Course Evaluation

students - MAP - Evaluations
<more>|<rare>
7   XX +
   |   |
   XXX |
   T|
6   +
   |   |
   |
   |
   |
   |
5   XX +
   |   |
   XX |
   XX S|
   XX |
   XX |
4   X +
   X |
   XX |
   XX |
   XX |
   XXX |
3   XXXXX M+
   XXXX | 19
   X |
   XX |
   XXX | T
2   XX +
   XX |
   | 2 8 15
   XX S|
   XX |S 13
1   X +
   XX |
   XX | 7 11 |
   | 14
   |
   X |
   |
0   X +M
   T| 17 18 |
   | 3 9 |
   | 16 20 |
   | 1 5 6 12
-1   +
   |S 21
   | 4
   | 10
-2   +
   <less>|<frequ>

Figure 2
Variable Mapping of Online Course Evaluation

students - MAP - ratings
<more>|<rare>
7   XXXX T+
   |   |
   |
   |
   |
   |
6   +
   |   |
   XXXX |
   X |
   XX |
   XX |
5   X +
   X |
   XX |
   XX |
   XX |
4   X +
   XXX |
   XXX |
   XXX M|
   XXX |
   XXX |
3   XX +
   X |
   | 2 7
   XX | T
   XX |
   XX |
2   S+
   X |
   | 8 19
   XX |
   |S |
   XXX |
1   +
   X |
   |
   T|
   |
   |
   |
   |
0   +M 3 13 17 20 |
   | 6 18
   | 4 14 16 |
   | 1 5
   |
   |
-1   +
   |S 12
   | 10 21 |
   | 9
-2   +
   <less>|<frequ>
Measuring differences in pencil-and-paper and online course evaluations

Each X represents one student and the mean of these students ability, or level of agreeableness for the students is found on the left side of the figures and is denoted with an “M”. The mean of the item difficulties, or difficulty to endorse, is also denoted with an “M” and is located on the right side of the figures. For both paper and pencil and the online evaluation the mean for the students is much higher than the mean of the items. This agrees with the previous finding that overall, most students agreed with the items.

Despite these similarities, some individual items do appear to be functioning differently for the paper and pencil and the online course evaluation. The mean for students on the paper and pencil evaluation was roughly 3 logits compared to the mean of 3.5 logits for the online evaluation. This indicates items were easier to endorse for online evaluations than paper and pencil evaluations. The item mean was the same for both evaluations.

In addition to the means, differences in the functioning of individual items are also noted. For the paper and pencil evaluation, item 19 was the most difficult to endorse whereas item 10 was the easiest to endorse. For the online evaluation 2 and 7 were the most difficult to endorse, and item 9 was the easiest to endorse. In examining these items more closely, one discovered they were a mix of course items, instructor items, and learning outcomes. More specifically, they were 2, The textbook contributed to my understanding of the subject, 7, Graded tests, assignments, etc. were returned promptly, 9, The instructor presented course material in an effective manner, 10, The instructor had good knowledge of the subject matter, 19, The course stimulated me to read further in the area (See Appendix A). There are seven items between the item mean and person mean for the paper and pencil evaluation, and five items between the means for the online evaluation. Again, this reiterates that students found items on the online evaluation easier to endorse.
Measuring differences in pencil-and-paper and online course evaluations

Discussion

This study demonstrates there are similarities in rating scale categorical functioning, including frequency counts and fit statistics, as noted within the literature. Liegle & McDonald (2005) discovered high performing students’ rate instructors more positively than poor performing students. In examining a course with advanced undergraduate and graduate students, who are more likely to be high performers, since they are enrolled, perhaps this is why most students rated the instructor more favorably. However, the instructor has been continuously teaching this course for many years, which may also account for the high ratings.

Also, online respondents provided slightly more positive responses than paper respondents, which is indicative of the literature, as Avery et al. (2006) found a similar result when comparing online and paper and pencil results. She attributed these results to a highly computer literature population, as well as a population with readily available computer access, which could be the case here. However, Layne et al. (1999) noted academic area affects evaluation results, which is another possibility here, but further tests comparing course evaluations between academic departments would need to be conducted.

Online respondents could also choose when they filled out their course evaluations. McGourty, Scoles, & McGourty, Scoles, & Thorpe (2002) noted pro-exam responses are more positive than pre-exam responses as well. Thus, it is plausible online responses were more positive than paper responses in our study for these two reasons.

Although there are many similarities between online and paper and pencil course evaluations in this study, several evaluation items do appear to function differently across formats of evaluation. Kasiar et al. (2002) noted a similar finding using CTT tests, noting statistically significant differences between 29% of the evaluation questions. Although, she noted
most question differences could be attributed to a median/range difference of only one point. This did not lead to a different meaning of the response for most questions and, thus, she did not discuss the differences between these questions.

However, in our study, there are significant differences between a mix of course and instructor items, as well as learning outcomes. Such differences would usually be attributed to various factors, such as gender, age, etc. in CTT analyses, but the Rasch model identifies these items as functioning differently on the evaluation based on mode of delivery. Further study must and will be conducted to investigate why such differences exist, which could be due to the evaluations coming from fall semesters in two different years.

Educational Significance

Given that many higher education institutions are utilizing or considering online evaluations to replace the more traditional pencil-and-paper course evaluations, studies of this nature are critical to making data-driven decisions and to know the potential consequences of such choices. CTT methods can provide useful aggregate data regarding course evaluation responses via online and paper and pencil formats, as numerous studies have shown differences in regards to student satisfaction, student perceptions, time, cost, and the like. However, this data is based on the assumption that a course evaluation is functioning similarly within both online and paper and pencil formats, as well as only provides a piece of the puzzle, as demonstrated by Bradley & Bradley (2006).

On the other hand, the Rasch model can provide a more detailed and comprehensive display of how course evaluations are functioning in terms of not only respondents and items, but within varying formats. Our study demonstrates assumptions shouldn’t be made about the functioning of course evaluations in varying formats and since few, if any, studies have been
Measuring differences in pencil-and-paper and online course evaluations conducted considering such functioning, more should be considered. Simply stated, it is necessary research to compare and contrast evaluation implementation characteristics, including format. These results should be disseminated to provide educators and administrators with useful information to make more informed decisions regarding evaluation administration methods and the interpretation and comparability of the impending results.

References


Hmieleski, K. “Barriers to Online Evaluation: Surveying the Nation’s Top 200 Most Wired Colleges.” Troy, N.Y.: Interactive and Distance Education Assessment Laboratory, Rensselaer Polytechnic Institute, 2000 (Unpublished Report).


Measuring differences in pencil-and-paper and online course evaluations
