Studies in Educational Evaluation 51 (2016) 77–87

Contents lists available at ScienceDirect

Studies in Educational Evaluation

journal homepage: www.elsevier.com/stueduc

A psychometric approach to the validation of a student evaluation of teaching instrument

Anthony P. Setari, Ph.D.[a], Jungmin Lee, Ph.D.[b], Kelly D. Bradley, Ph.D.[c,*]

[a] University of Kentucky, Dept. of Educational Policy Studies & Evaluation, 144 Taylor Education Bldg, Lexington, KY 40506, United States
[b] University of Kentucky, Dept. of Educational Policy Studies & Evaluation, 144C Taylor Education Bldg, Lexington, KY 40506, United States
[c] University of Kentucky, Dept. of Educational Policy Studies & Evaluation, 144A Taylor Education Bldg, Lexington, KY 40506, United States

ARTICLE INFO

Article history:
Received 9 February 2016
Received in revised form 23 September 2016
Accepted 28 September 2016
Available online xxx

Keywords:
Course evaluation
Evaluation methods
Higher education
Rasch
Teacher evaluation
Validation

ABSTRACT

The purpose of this study was to conduct a validation analysis of an SET and provide a validation framework of SETs that can be included when designing complete evaluations of teaching within higher education institutions. A series of Rasch analyses was conducted on the results of the SET, examining the responses of students within a college and three departments. Results show the majority of items were moderately difficult to endorse in the college and departments, there were issues with DIF, and two items did not consistently fit the model. The study provides an analysis framework that may aid policymakers and institutional administrators in developing higher quality SETs, and demonstrates the need for validating SETs being implemented in higher education settings.

1. Introduction

In response to rising tuition costs and questions over the value of the education students are receiving, policymakers are calling for higher education institutions to demonstrate their relative value and effectiveness (Huisman & Currie, 2010; Leveille, 2006; Sponsler, 2009). This demand has led many higher education institutions to use the data collected during their routine evaluations of instructors, departments, and colleges for accountability reporting purposes. However, there are three crucial issues with the routine evaluations being implemented. First, what components make for an effective higher education institution and which should be considered in an evaluation are debatable, with many evaluations being created and implemented without a framework for interpreting effectiveness (Marsh & Dunkin, 1992). Second, evaluation instruments are often implemented without consideration for reliability and validity (Marsh & Dunkin, 1992; Sproule, 2000; Wright & Jenkins-Guarnieri, 2012). Finally, the interpretation of data coming from the evaluation instruments is often based on an aggregation or mean of respondent scores, with little attention given to responses of individual items and individual student responses (Marsh & Dunkin, 1992; Wright & Jenkins-Guarnieri, 2012). Each of these issues needs to be addressed before better evaluation data can be produced for policymakers to use in their decision-making.

One of the most widely used evaluation instruments in colleges and universities is student course surveys, also known as student evaluations of teaching (SET; Kulik, 2001; Seldin, 1993; Wright & Jenkins-Guarnieri, 2012). The quality of SETs varies, though, with many institutions implementing instruments with serious measurement concerns (Stark & Freishtat, 2014). Item-centered techniques, such as item response theory and the Rasch model, are becoming increasingly popular in the interpretation of SETs (Bradley & Bradley, 2006). These item-centered techniques have the ability to identify issues within a measurement instrument and identify items that might not be measuring constructs accurately, as well as identify differences in how groups are responding to items (Bond & Fox, 2007; Toland, 2013). In addition, these item-centered techniques have the ability to identify how students are responding and provide a context for which items students most commonly and least commonly to endorse. Continued usage of these item-centered techniques is important in the refinement of SET as an evaluation technique.

The purpose of this study was to conduct a validation analysis of an SET implemented at a large public Carnegie Tier-I research university in the United States, and provide a validation framework of SETs that academics and policymakers can include when designing complete evaluations of teaching within higher

* Corresponding author.
E-mail addresses: Setari@uky.edu (A.P. Setari), Lee@uky.edu (J. Lee), Kdbrad2@uky.edu (K.D. Bradley).

http://dx.doi.org/10.1016/j.stueduc.2016.09.006
0191-491X/© 2016 Elsevier Ltd. All rights reserved.

education institutions. In comparison to typical examinations of evaluation instruments that only report general descriptive statistics, this analysis focuses on examining the SET's results at the item level through a Rasch model analysis. There are four research questions that guided this study: (a) how are students perceiving the instructors and courses?; (b) how do students' perceptions of instructors and courses vary?; (c) how well does the instrument measure students' perceptions of instructors and courses?; (d) how can the SET being implemented be altered to be a more precise measure of students' perceptions?

2. Background

SETs are common in higher education institutions (Kulik, 2001; Seldin, 1993; Wright & Jenkins-Guarnieri, 2012). SETs have been implemented in colleges and universities arguably since their creation at the University of Washington in the early twentieth century; and since that time, they have been studied continuously as their structure, usage, and impact has altered over time (Guthrie as cited in Kulik, 2001). SETs had two original purposes: to provide administrators with information about instructors and to provide instructors with feedback from students on ways to improve their instructional practices. Since their creation, the purposes and uses of SETs have been in a continual state of flux, as SETs have been morphed to meet the needs of students, instructors, administrators, and policymakers in their respective time periods (Ory, 2000). For example, in the 1960s, SETs were used as a means to demonstrate accountability to students challenging higher education institutions. Marsh and Dunkin (1992) identified SETs' four main purposes presently: (a) to provide student feedback to teachers; (b) to provide administrators with information on teachers from the viewpoint of students; (c) to provide students with information about instructors and courses; (d) to provide information about and for research purposes. Given the common usage of SETs and their continued applicability in providing data to policy makers and agencies, it is unlikely that SETs will fall out of favor with higher education institutions.

The widespread usage of SETs by higher education institutions has led to the development of a general pattern for creation and implementation (Sproule, 2000). According to Sproule (2000), SETs are generally composed of both close-ended and open-ended items. Close-ended items are often posed as statements in which the student is asked to disagree or agree with to a specific degree. Open-ended items often require students to critique their experience in the course and the instructor. In addition, SETs are usually anonymous and administered in a standardized fashion throughout an institution, typically not by the instructor of a course. Finally, administrators in higher education institutions will compare the SET results of an individual course or instructor to an established average as a means of determining effectiveness. This pattern in creation and implementation has intentionally or not, become the framework for how many institutions use their SETs.

The desire to create effective SET instruments has led to a large number of validity and reliability studies on the subject (Marsh & Dunkin, 1992; Spooren, Brockx, & Mortelmans, 2013; Wright & Jenkins-Guarnieri, 2012). Many of these studies, however, have suffered from a lack of theoretical or model footing (Marsh & Dunkin, 1992). Without a theoretical or model basis guiding these studies, there is not a conceptual indicator to suggest if analysis results about the instrument or the items are appropriate for an SET, making interpretation of these results less meaningful. However, results of these studies are noteworthy when considering SETs. For example, there is evidence to suggest that bias from factors such as ease of class, teacher attractiveness, and teacher charisma (Felton, Koper, Mitchell, & Stinson, 2008; Shevlin, Banyard, Davies, & Griffiths, 2000; Spooren et al., 2013) impact students' responses on SETs. Of note for this study is the finding that what students view as valuable in the evaluation of instructors and courses may vary from what administrators view as valuable (Onwuegbuzie et al., 2007). Interpreting findings such as these using a theoretical or model framework would aid in assuring that SET construction and implementation was being driven by conceptually appropriate ideas.

2.1. Evaluation facets

In the work *The Superior College Teacher from the Students View*, Feldman (1976) put forth a series of facets representing what he interpreted as representing what post-secondary education students considered important. In the article, Feldman synthesized the results of 72 studies examining student surveys of teachers and higher education institutions, and in doing so, identified three facets. The facets the author put forth, in respective order from easiest to most difficult, are presentation, facilitation, and regulation. The presentation facet refers to, "items measuring overall evaluation of the teacher or the course are connected with stimulation of interest, enthusiasm, knowledge of subject matter, preparation and organization of material, clarity, and instructional outcome for the student" (p. 260). The facilitation facet refers to an instructor's work within the classroom, specifically what Feldman cites from Widlak, McDaniel, and Feldhusen (1973), "as the instructor's role as Interactor or Reciprocator" (p. 260), which includes issues such as instructor kindness and willing to help students. The regulation facet was closely related to fairness-of-evaluation in the course, and relates to what Feldman cited from Widlak et al. (1973), "aspects of the instructor's role as Director or Administrator" (p. 262). These three facets have been used in many studies to evaluate teacher performance, and represent a model of student's perception of teacher effectiveness that could potentially be used in other studies (Marsh & Dunkin, 1992).

2.2. The Rasch model

The Rasch model is commonly used in the validation of instruments, as it provides analysis at the item level of an instrument (Bond & Fox, 2007). An important feature of the Rasch model is that it is able to identify item difficulty levels and respondent ability levels. Item difficulty levels are reported on a logit scale where 0.0 represents the point at which a respondent has a 50/50 likelihood of endorsing the item. On the scale, items with a positive logit estimate represent items that are challenging to endorse by respondents, becoming increasingly more challenging as logit estimates increase. Similarly, items with a negative logit estimate are relatively easy to endorse and become increasingly easier to endorse as logit estimates decrease. Respondents' ability levels are also represented with logit estimates. Respondents with high item endorsement ability levels will have positive logit estimates and those with low items endorsement ability levels will have negative logit estimates. In the context of survey research, ability levels refers to possessing higher degrees of a latent trait and being more likely to endorse a high degree of items. Both ability levels and item difficulty levels are key features of the item level analysis.

Furthermore, items are given fit statistics that determine if it an item is functioning appropriately. Fit statistics demonstrate how well the provided data fits the expectations of the Rasch model and effectiveness of the item's measurement (Linacre, 2002). Item infit and outfit estimates are key indicators for determining the quality of the measurement of an item. For surveys, items are typically expected to have a mean-square value between 0.6 and 1.4 (Wright & Linacre, 1994). An item with an infit or outfit estimate outside of this range may have an issue with measurement and be considered