Coherence in measuring student evaluation of teaching: a new paradigm

Fan Huang\textsuperscript{1}, Emily Pey-Tee Oon\textsuperscript{1} and W P Fisher Jr\textsuperscript{2}

\textsuperscript{1}Faculty of Education, University of Macau, Taipa, Macau SAR, China
\textsuperscript{2}Research Associate, BEAR Center, Graduate School of Education, University of California, Berkeley, Berkeley California 94702 USA

\textsuperscript{1}E-mail: hefu411@126.com

Abstract. Student Evaluations of Teaching (SET) have been adopted worldwide as a standard practice to enhance teaching and learning quality. However, improvement efforts are threatened by various problems involving ordinal scores and interpretive frameworks restricted to individual item sets and uninformed by a well-defined invariant construct. In this study, an approach to framing SET items according to theory rather than personal or subjective decisions is described. A sample of 920 students from the University of Macau participating in this study provided responses to 42 SET items. Applying a scientific model of measurement, the data collected were analysed and interpreted. Justification between theoretical and measurement analysis informed the creation of a theoretical SET construct map. That map and the associated item hierarchy were used as reference sources for configuring SET reports and recommendations. A new coherent paradigm for SET documentation is proposed for further study.

1. Introduction: framing SET items according to theory

Growing concerns with quality assurance in higher education focuses on students as the most reliable and often the only audience of information indispensable to the feedback needed to enhance teaching and learning effectiveness. The use of SET has increased rapidly since the 1960s, and in 1990, more than 90\% of higher education institutions in the United States used the rating data [1]. Nowadays, SET have accordingly been adopted worldwide as a standard practice [2–4].

However, improper constructs used, questionnaire design, data analysis, reporting, information use in policy making decision and lack of a coherent system documenting SET measures lead to relationships between variables be over-interpreted or overlooked, and they could also provide unwarranted, mistaken information concerning course quality and instructors’ effectiveness. The quality improvement efforts may be threatened by the following problems:

(a) On subjective selection of constructs and poorly framed item content that lack of theoretical support
(b) Lack of a coherent system for documenting and applying SET measures.

The problems mentioned above may set obstacles to provide instructors with information that can guide them to make changes in curriculum design or in teaching behaviours, and what’s worse, may lead to erroneous personnel decisions made by administrators, which could influence the instructors’ future careers. Therefore, poor SET practices and misuse of data may be harmful for course quality
2. Research objectives

Accordingly, we aim to explore the following three research objectives in this study:

1) To ascertain SET constructs (Course and Instructor) theoretically based on a body of SET pertinent studies.
2) To map measurement analysis on content against SET theories.
3) To project a coherent system in documenting SET measures capable to compare SET findings across contexts.

3. Theoretical framework

3.1. Justification between theoretical and measurement analysis

Less difficult items are reflective of most basic essential aspects of course quality, and so they carry less weight in the interpretation of the measures. More difficult items are reflective of aspects beyond essential expectations, and so they carry more weight in the measurement of course quality. Moderate course quality expectations are indicated by items with agreeabilities between these two distinctions (Figure 1). For instance, an instructor providing only responses to direct questions exhibits less quality compared to an instructor who also inspires students’ interest in the subject. Items pertaining to the former (responsiveness during class) should be less difficult to agree with and items pertaining to the latter (interest inspiration), more difficult to agree with. In other words, it is theoretically less difficult to agree with responsiveness during class but more difficult to agree with interest inspiration. Instructors that garnered high ratings of agreement on interest inspiration measure higher than instructors whose that garnered most agreement on responsiveness during class.

![Figure 1. Mapping measurement analysis against theoretical justification.](image)

3.2. Scientific measurement theory

Evidence of SET was evaluated by fitting the response data to a scientific measurement model explicitly hypothesizing an invariant relation between student agreeability and SET item difficulty:

$$\ln\left(\frac{P_{ni}}{(P_{nj-i})}\right) = B_n - D_i - K_j$$

which says that the log-odds of any student choosing any given category on any SET item is equal to the difference between the estimate B of student n’s ability and the difficulty estimate D of item I, and
the additional difficulty $K$ of rating category $j$ [5-7]. Assessment data must reasonably fit a model of this kind if we are able to evaluate the teaching effectiveness.

4. Methods

4.1. Sample
Students were invited to voluntarily participate in a research survey. A convenience sampling method was employed, whereby students were chosen from different educational levels at the University of Macau (UM). In pilot study, a total of 241 responses were collected from seven classes. In main study, 679 responses were collected from 28 classes.

4.2. Instrument
In essence, 36 items with six response categories (1 = Very Strongly Disagree, 2 = Strongly Disagree, 3 = Disagree, 4 = Agree, 5 = Strongly Agree, 6 = Very Strongly Agree) were constructed for the pilot study questionnaire, guided by the literature and research. 42 items were included in the main study.

4.3 Expected reliability
Given 36 items and a six-point rating scale, 180 (36×5) distinctions in total were made for 90 each of the two focal constructions (Course and Instructor). The expected uncertainty was equal to 0.2 logits. Measurement reliability is conservatively estimated at 0.96, for a true SD of 1.0 logits [8].

5. Results

5.1. Data quality assessment
Before conducting further investigations, data quality was first assessed. All the data were imported to Winsteps software, version 3.81.0 [9]. The SET data were evaluated for their capacity to support estimation of linear units of measurement before further investigations were undertaken.

5.1.1. Model fit and reliability
In the pilot study, the person and item reliability values were 0.96 and 0.97. Compared to the expected reliability, the results show that person reliabilities in both pilot and main study are no less than expectation. The information-weighted (infit) and outlier sensitive (outfit) mean square fit statistics [7] for the 18 items were between 0.58 and 2.26, with the infit and outfit for item A11 (I spent a lot of time studying for this course) being at 2.29 and 2.64, respectively. The person and item reliability values from main study were 0.97 and 0.99. The infit and outfit statistics [7] for the items were between 0.63 and 2.64, with the infit and outfit for item A11 (time studying) being at 1.79 and 2.09, respectively.

5.1.2. Usefulness of response categories
The shapes of category probability curves are presented in figure 2. Category 2 (strongly disagree) showed no recognizable peak (Figure 1). Depending on the status of the other indicators (number of responses, fit statistics, mean values of the number of students who answered in each category, etc.), future research on these data should collapse Categories 1, 2, and 3, or perhaps only 2 and 3, for a more optimal representation of response category usefulness.
Figure 2. Probability curve of six response categories.

5.1.3. Invariant results between the pilot and the main study The item difficulties estimated from the pilot study correlated 0.96 (disattenuated, 0.98) with the main study (Figure 3), which indicated that the results from two studies remain invariant.

Figure 3. Pilot study item calibrations vs Main study estimates.

5.2. Formation of item hierarchy based on construct map
In order to achieve stability between theory and item calibrations, rounds of iterations are inevitable. Three theoretically intended groups (difficult to agree with, average agreeability and easy to agree with) were developed based on construct map (Figure 4). The construct map and the associated item hierarchy were used as a reference sources for configuring SET reports and recommendations.
6. Discussion: Propose a new coherent paradigm for SET documentation

In this study, a new paradigm was provided to overview teaching effectiveness of every instructor scientifically, which is a shift of the status quo in evaluating teaching effectiveness. In figure 5, a coherent system was developed which can be used to project SET according to horizontal coherence (across faculties), vertical coherence (as individuals, to classes, schools, universities, etc.) and developmental coherence (from different time points).
7. Conclusion
A new paradigm was proposed for coherence in measuring SET, which set objective foundation for administrators and instructors making reasonable policies or adjusting meaningful teaching activities to enhance course quality, teaching and learning effectiveness.

References
[1] Seldin P 1999 Changing practices in faculty evaluation (San Francisco: Jossey Bass)
[2] Denson N, Loveday T and Dalton H 2010 Student evaluation of courses: what predicts satisfaction? High. Educ. Res. Dev. 29 339-56
[3] Kember D, Leung D Y and Kwan K 2002 Does the use of student feedback questionnaires improve the overall quality of teaching? Assess. Eval. High. Educ. 27 411-25
[4] Narasimhan K 2001 Improving the climate of teaching sessions: the use of evaluations by students and instructors Qual. High. Educ. 7 179-90
[5] Andrich D 2010 Sufficiency and conditional estimation of person parameters in the polytomous Rasch model Psychometrika 75 292-308
[6] Rasch G 1960 Probabilistic models for some intelligence and attainment tests (Copenhagen, Denmark: Paedagogiske Institut)
[7] Wright B D and Masters G N 1982 Rating scale analysis: Rasch measurement (Chicago, Illinois: MESA Press)
[8] Linacre J M 1993 Rasch-based generalizability theory: Reliability and Precision (S.E.) Nomogram Rasch Measurement Transactions 7 283
[9] Linacre J M 2014 WINSTEPS (Version 3.81.0) [Computer Software]. Chicago: Winsteps.com.