Research Paper

Item fairness in SETs in higher education according to academic college: Data from three major universities in Jordan

Accepted 13th September, 2017

ABSTRACT

This study explores students' academic college as a source of item's unfairness in students' rating of teaching quality and effectiveness at higher education institutions. The data utilized in this study were collected from three major universities in Jordan. Each university used different surveys to assess teaching quality and effectiveness and as such, three data sets were analyzed in the current study. The results showed that two out of the three universities used surveys that have unfair items to assess teaching quality and effectiveness.

Key words: Unfairness, SET, students rating of teaching, quality, higher education.

INTRODUCTION

Students evaluation of teaching (SET) has become a universal practice in higher education to measure effectiveness (Goos and Salomons, 2016; Wachtel, 1998; Chen and Hoshower, 2003; Clayson, 2009; Berk, 2005). SETs are often used for critical decisions, such as retention, tenure, and promotion of faculty members (Kogan et al., 2010). The majority of universities use single survey to collect data about students' rating of teaching quality and effectiveness, which is distributed at the end of the semester. Marsh (1984, 2007) and Marsh and Roche (1997) demonstrated that SETs surveys are biased, indicating that SETs might have unfair items according to students' academic colleges. As Camilli (2006) stated, "DIF is synonymous with statistical bias, whereas unfairness can only be established if these measured differences are factors irrelevant to the test construct; there is no direct route from statistical bias to unfairness. To maintain the distinction between statistical bias and unfairness, DIF is used as a kind of screening mechanism for quality control" (p. 234). This means that item fairness can be detected using differential item functioning (DIF) statistical methods (Camilli and Shepard, 2007). Since DIF is a major threat to validity and reliability (Duncan, 2006; Monahan, 2002), SETs' surveys should be given sufficient consideration in order to achieve its intended purpose (Oon et al., 2016). The current study aims at exploring and examining students' academic college as a source of item unfairness (or DIF) in SETs. SETs are affected by students' disciplines (Kember and Leung, 2011) and these differences could be attributed to unfair items, which give advantages or disadvantages to one college over another. Therefore, more attention should be directed towards consequential validity of SET surveys (Ory and Ryan, 2001). This study investigates the presence of unfair items in SET surveys by answering the following research question: "Do SET's surveys contain unfair items (DIF items) according to students' academic college?"

Theoretical background

Many research studies have shown that students are a reliable source to provide information about the extent to which learning experience is productive, informative, satisfying or worthwhile (Archibong, and Nja, 2011). Therefore, SET surveys of teaching, courses, and programs are adopted almost in every university. Thereafter, data are collected, reports are generated across instructors,
departments, and colleges and dealt with as evidence of teaching effectiveness. These are used for professional decisions (Sproule, 2000).

However, there are research studies that have skeptical point of views about SETs (Uttl et al., 2016; Rienties, 2014; Martin, 1998; McPherson and Jewell, 2007; Watchel, 1998, Weinberg et al., 2007; Gump, 2007), and there are others who support and trust such evaluations (Yao and Grady, 2005; Spencer and Flyr, 1992; Contreras-McGavin and Kezar, 2007; Gump, 2007). Despite this controversy, such evaluations are seen by many as a valuable and beneficial tool to improve teaching and student learning outcomes (Lattuca and Domagal-Goldman, 2007; Dommeyer et al., 2004). To maximize the SET benefits, Rantanen (2013) suggests applying SET surveys to suitable courses for each teacher, while Giles et al. (2004) recommend students partnership in designing and implementing evaluations.

Several studies have shown that there are many variables (independent variable) that influence SETs (dependent variable): Grades or expected grades (Griffin et al., 2014; Badri et al., 2006; Brockx et al., 2011), gender (MacNell et al., 2015; Badri et al., 2006), teachers' characteristics (Wolbring and Riordan, 2016; Clayson and Sheffet, 2006; Patrick, 2011; Greimel-Fuhrmann and Geyer, 2003; Shevlin et al., 2000), classroom size and response rate (Al Kuwaiti et al., 2016; Koh and Tan, 1997; Badri et al., 2006), course difficulty (Addison et al., 2006), course level (Santhanam and Hicks, 2002), course type (Beran and Violato, 2005), general versus specific education (Ting, 2000), and course syllabus tone (Harnish and Bridges, 2011).

Also, students' academic discipline is one of the factors that have significant effect on SETs (Neumann, 2001; Chen and Watkins, 2010; Basow and Montgomery, 2005), and wording of SET items could be one of the reasons behind the effect of students' discipline on SETs as shown by Anders et al. (2016). This implies that some items might be worded to be understood in a different way based on students' discipline, or some items content might give advantages or disadvantages to colleges over others. In psychometric terms; students' endorsement of an option (response) on Likert scale item could be influenced by students' academic college rather than what this survey measures. In other words, students' college could be a source of differential item functioning (DIF), or the item might be an unfair one. DIF, or item unfairness, means that students in different colleges (e.g. education vs. arts) respond differently to an item, even though they share the same trait level.

Therefore, in this study, the relationship between students college, or field of study and DIF is a threat to the survey's validity and reliability. The probability of endorsing an option or point in a rating scale item should be determined by the latent trait (e.g. teaching effectiveness) measured by the survey that said item comes from. DIF happens when the probability of selecting an option on the item for two respondents from different groups who have the same trait level is not the same (different colleges or field of study), therefore, the item could be unfair or its function is not the same across these groups (colleges). In other words, Raju and Ellis (2002) indicated that detecting DIF means examining the degree to which two survey takers with identical standing on the latent trait but from different groups (e.g. education and arts) have the same probability of choosing the same option on the item.

There are several methods that could be used to detect DIF. The current study utilized The generalized Mantel-Haenszel (GMH) method for categorical data (Mantel and Haenszel, 1959) based on group differences in the entire response distribution. This method is selected as it is well-established and suitable for polytomous items (4 and 5 point-Likert scale Items). A study on calculation of GMH χ² statistic has been carried out by Zwick et al. (1993).

**Table 1.** Number of items, item type, Cronbach's Alpha, and item-to-total correlation (ITCor.) range for each instrument in each university.

<table>
<thead>
<tr>
<th>University</th>
<th>No. of items</th>
<th>Item's scale</th>
<th>Cronbach α</th>
<th>ITCor. Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>20</td>
<td>5-point Likert</td>
<td>.973</td>
<td>.744-.817</td>
</tr>
<tr>
<td>B</td>
<td>20</td>
<td>4-point Likert</td>
<td>.959</td>
<td>.542-.785</td>
</tr>
<tr>
<td>C</td>
<td>13</td>
<td>5-point Likert</td>
<td>.950</td>
<td>.591-.805</td>
</tr>
</tbody>
</table>

**METHODOLOGY**

**The instruments**

This study used three data sets collected by three major universities in Jordan using three different surveys. Each of these survey is approved by the authorized councils at that university to be used to monitor teaching effectiveness and quality. Each university followed the standardized procedures to build the instrument. Table 1 shows the number of items, item type, Cronbach's Alpha, and item-to-total correlation (ITCor.) range for each
The need for education, science, and arts was agreed upon less by the Reference Group members (Colleges) based on total number of surveys analyzed from each university. The data were obtained from AQACs at three different universities in Jordan at the end of the second semester of the academic year (2015/2016). A total of 20,186 surveys from colleges of education, science and arts were analyzed to achieve the purpose of this study. Table 2 shows the number of surveys analyzed from each university.

### Table 2. The number of surveys analyzed from each university.

<table>
<thead>
<tr>
<th>University</th>
<th>Arts #</th>
<th>Arts %</th>
<th>Education #</th>
<th>Education %</th>
<th>Science #</th>
<th>Science %</th>
<th>Total #</th>
<th>Total %</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2541</td>
<td>43.63</td>
<td>1790</td>
<td>30.73</td>
<td>1493</td>
<td>25.64</td>
<td>5824</td>
<td>100.00</td>
</tr>
<tr>
<td>B</td>
<td>3256</td>
<td>32.56</td>
<td>1918</td>
<td>19.18</td>
<td>4826</td>
<td>48.26</td>
<td>10000*</td>
<td>100.00</td>
</tr>
<tr>
<td>C</td>
<td>1110</td>
<td>25.45</td>
<td>1637</td>
<td>37.53</td>
<td>1615</td>
<td>37.02</td>
<td>4362</td>
<td>100.00</td>
</tr>
</tbody>
</table>

* Selected randomly from 16396 surveys due to Software Limitation.

### Data collection

In each university in Jordan, there is an Accreditation and Quality Assurance Center (AQAC) and one of its responsibilities is to collect data about students evaluation of teaching effectiveness on that university. The data were obtained from AQACs at three different universities at the end of the second semester of the academic year (2015/2016). A total of 20,186 surveys from colleges of education, science and arts were analyzed to achieve the purpose of this study. Table 2 shows the number of surveys analyzed from each university.

### Data analysis

To achieve the purposes of this study, Generalized Mantel-Haenszel (GMH) Method was used through GMHDIF program (Fidalgo, 2010). GMH method is a flexible methodology for assessing the association between categorical variables. Therefore, it is used for the assessment of the differential item functioning of both dichotomous and polytomous items (Fidalgo and Madeira, 2008). The GMH method assesses DIF using a chi-square test ($Q_{MH}$) for multiple groups and performs pairwise comparisons in case the DIF items are detected during analysis; the independent variable is student's college (Education, Science, and Arts) and the dependent variable is SETs items responses across the ability level after matching the groups' members (Colleges) based on total test score as the matching criterion. Separate group (College)-by-score contingency tables are created for each level of the matching criterion (Total score as an ability level) and a chi-square test is used to assess the degree to which the distribution across the score categories (and ability levels) differs for the groups. Significant chi-square ($Q_{MH}$) indicates the presence of DIF (Fidalgo, 2011; McBride, 2008). The matching criterion must be a valid and reliable measure, and should not be contaminated by DIF-items. Moreover, because total test score is used as the criterion variable for grouping examinees or respondents, Holland and Thayer (1988) recommended a potential means of improving that criterion using a two-step form of the MH procedure, in which items identified as displaying DIF on an initial implementation, are removed from the matching criterion for subsequent analysis. The need for purification depends on the number of DIF items detected in the first or initial step and the number of test items (Fidalgo et al., 2000; Fidalgo, 2011).

GMHDIF can only provide significance testing without effect size estimation. Yin (2014) suggested the use of GMHDIF for significance testing and DIFAS 5.0 (Differential Item Functioning Analysis System 5.0) for effect size estimations to complement GMHDIF. In order to support GMHDIF analysis to provide the effect size, the Liu-Agresti Cumulative Common Log-Odds Ratio (L-A LOR) (Liu and Agresti, 1996) was used through DIFAS 5.0 (Penfield and Camilli, 2007). In each DIFAS analysis, only two groups were examined as the DIFAS could not support multi-group analysis. A negative value of L-A LOR indicates that the DIF item is agreed upon less by the Reference Group than as expected and a positive value of L-A LOR indicates the DIF item is agreed upon less by the Focal Group than as expected. While a zero value means that there is no DIF. The classification of effect size used in the current study is presented in Table 3.

### Table 3. DIF criteria for effect size by Penfield and Camilli (2007).

<table>
<thead>
<tr>
<th>Effect size</th>
<th>Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non DIF</td>
<td>Either L-A LOR is not significantly different from zero, or $</td>
</tr>
<tr>
<td>Flagged DIF</td>
<td>If L-A LOR is significantly different from zero; and $</td>
</tr>
<tr>
<td>Large DIF</td>
<td>If L-A LOR is significantly greater than 0.43 and $</td>
</tr>
</tbody>
</table>
### RESULTS

To answer the research question, "Do SET's surveys contain unfair items (DIF items) according to students' academic discipline?" GMH Method was used through GMHDIF program (Fidalgo, 2010) for each university data set. To control some of the problems resulting from stratifying variable contamination as suggested by Penfield and Camilli (2007), the two-stage screening procedure was used to purify the stratifying variable. In the first stage, a standard DIF analysis was run and items are screened for statistical significant of DIF. In the second stage, the DIF analysis was rerun using the items that did not have statistical significant of DIF in the first step. Table 4 shows the number of DIF items in each university SET survey before and after using purification procedures mentioned previously.

Table 4 shows that all the instruments in the three universities have statistical significant DIF items. To complement GMHDIF analyses, DIFAS program was used to estimate the DIF effect size. DIFAS analyses revealed that university A SET survey does not have any flagged DIF items, while a Flagged DIF item was observed in university B SET survey and university C SET survey. Table 5 shows the effect size for the items that flagged DIF according to DIFAS analyses based on the criteria reported in Table 3.

Table 5 shows that the two items from two different universities (B and C) were flagged as DIF items based on the effect size. These items are: "My teacher invests classroom time for teaching and learning purposes" and "My teacher uses different assessment methods to assess students learning." Standardized residuals for the cross tabulation of students' college by their responses to this item was calculated to get clear picture of the students response to these items as shown in Table 6.

Table 6 shows that the content of item "My teacher invests different assessment methods to assess students learning" uses educational terms that might be the reasons behind
CONCLUSIONS AND IMPLICATIONS FOR FUTURE RESEARCH

After more than seven decades of research on SETs in higher education, most researchers believe that it is reliable, valid, and useful. This study provides evidence that disciplines or students' college could be a source of item unfairness. Students who have the same level of ability or ratings of teaching effectiveness but being in a different colleges might understand and perceive some items in CES differently. Therefore, they respond to these items in a different manner. The results of this study showed that two out of the three universities used surveys that have unfair items. This type of error (unfair items) is a major instrument of threat to validity and reliability. The results of this study are in line with those of Wu et al. (2016) who encourages universities to invite scholars and experts to develop teaching effectiveness surveys in order to ensure that these surveys are suitable for use by different colleges. Otherwise, teaching assessment of the colleges can easily lead to a variety of subjects' form that cannot effectively improve teaching quality.

The data analyzed in the current study were obtained from three colleges, and that might be one of the limitations of the current study. Therefore, it is highly recommended to conduct similar studies by analyzing data from different colleges. Also, only the Generalized Mantel-Haenszel method was used in the present study to assess DIF, therefore, it is recommended to investigate the presence of unfair items using different methods.

Higher education institutes should develop free unfair item surveys to assess teaching and teachers. In other words, the items should be perceived and understood in the same way despite the students’ different colleges or disciplines. Another implication of this study is the use of different surveys to assess teaching effectiveness in the same university. Each college can design its SET survey using terminology that is suitable to their students and programs.

REFERENCES


story to the edge. The influence of student, teacher, and course characteristics on student evaluations of teaching in higher education. Educational Assessment, Evaluation and Accountability. 23(4):289-306.


Wolbring T, Riordan P (2016). How beauty works. Theoretical mechanisms and two empirical applications on students' evaluation of...


Cite this article as:

Submit your manuscript at
http://www.academiapublishing.org/ajer