Abstract: This study explored academic discipline as a source of differential item functioning (DIF) in students’ rating of teaching quality and effectiveness at higher education institutions. Data utilized in this study was collected by Imam Abudalruman Bin Faisal University - KSA. The total number of surveys analyzed for the purpose of this study is 36459 from three colleges: Education, Health, and Engineering. Using Extended Rasch model (Rating Scale Model), the results show that the instrument contains four DIF items. The content of these four items confirm the possibility of considering discipline as a source of DIF items in students evaluation of teaching in higher education. Moreover, the results of the current study show that removing DIF items from the instrument increases construct validity.

Keywords: Differential item functioning, student’s rating of teaching, quality, higher education.
Higher education institutes have developed relatively complex procedures and instruments for collecting, analyzing, and interpreting data about institutional performance (Penny, 2003). As a result, students evaluation of teaching (SET) has become an increasingly common practice in higher education institutes as a measure of institutional performance and system effectiveness (Goos, & Salomons, 2017; Wachtel, 1998; Chen & Hoshowr, 2003; Clayson, 2009; Berk, 2005). SETs are often used for critical decisions, such as retention, tenure, and promotion of faculty (Kogan, Schoenfeld-Tacher, & Hellyer, 2010). Therefore, reliability and validity of SETs surveys should be given sufficient consideration in order to achieve the intended purpose (Oon, Spencer, & Kam, 2017). The way SETs items are worded will have an effect on the usefulness of the collected information (Marsh, 2007).

As the teaching process is a mix of several components that should be covered by SETs items, there is a possibility that some items in SETs might be worded in a way that suits one college over another; when SETs items ask students to rate pedagogical practices of teachers, college of education students could assign different meanings of these items when compared with their colleagues from another colleges. Marsh (1984, 2007) and Marsh and Roche (1997) showed the SETs surveys are multidimensional, and differential item functioning (DIF) could be the reasons behind SETs multidimensionality (Camilli & Shepard, 2007). Since DIF is a major threat of validity and reliability (Duncan, 2006; Monahan, 2002), the current study aims at exploring and examining students’ academic disciplines as a source of DIF in SETs and the effect of detected DIF items on the validity. Ory and Ryan (2001) recommended that greater attention should be directed toward consequential validity, particularly the matters of how ratings are used on today’s campuses and what happens as a result.

Research indicates that students are a qualified source to report on the extent to which the learning experience was productive, informative, satisfying or worthwhile (Archibong, & Nja, 2011). Therefore, students evaluation of teaching, courses, and programs are used almost in every university, and after the data is collected, reports are generated across instructors, departments, and colleges and viewed as evidence of teaching effectiveness that is then used for professional decisions (Sproule, 2000).

There are research studies that have skeptical point of views about SETs (Uttl, White, & Gonzalez, 2016; Rienties, 2014; Martin, 1998; McPherson & Jewell, 2007; Watchel, 1998, Weinberg, Fleisher, & Hashimoto, 2007; Gump, 2007), and there are many who support and trust such evaluations (Yao & Grady, 2005; Spencer & Flyr, 1992; Contreras-McGavin & 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, & Domagal-Goldman, 2007; Dommeyer, Baum, Hanna, & Chapman, 2004). To maximize the SET benefits, Rantanen (2013) suggests applying SET surveys to suitable courses for each teacher, while Giles and colleagues (2004) recommend student partnership in designing and implementing evaluations.

Reviewing the related literature shows that there are many variables that influence SETs: Grades or expected grades (Griffin, Hilton III, Plummer, & Barret, 2014; Badri et al., 2006; Brockx, Spooren, & Mortelmans, 2011), gender (MacNell, Driscoll, & Hunt, 2015; Badri, Abdulla, Kamali, & Dodeen, 2006), teachers' characteristics (Wolbring & Riordan, 2016; Clayson & Sheffet, 2006; Patrick, 2011; Greimel-Fuhrmann, & Geyer, 2003; Shevlin, Banyard, Davies, & Griffiths, 2000), classroom size and response rate (Al Kuwaiti, Alquraan, & Subbarayalu, 2016; Koh & Tan, 1997; Badri et al., 2006), course difficulty (Addison, Best, & Warrington, 2006), course level (Santhanam & Hicks, 2002), course type (Beran & Violato, 2005), general versus specific education (Ting, 2000) and course syllabus tone (Harnish & Bridges, 2011). Also, students’ academic discipline is one of the factors that has a significant effect on SETs (Neumann, 2001; Chen & Watkins, 2010; Basow & Montgomery, 2005), and the wording of SET items could be one of the causes behind the effect of students’ discipline on SETs as shown by Anders and colleagues (2016). This implies that some items might be worded to be understood in a different way based on students’ discipline. In psychometric
terms; students' endorsement of an option on a Likert scale item could be influenced by students' discipline rather than what the survey measures, which means that students' discipline could be a source of differential item functioning (DIF). DIF means the notion that students in different colleges (e.g., education vs engineering) respond differently to an item, even though they share the same trait level. This study contributes to this effort by examining the possibility of discipline, or field of study, as a source of DIF which is a threat to survey 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. When the probability of selecting an option on the item for two respondents who have the same trait level is not the same and they are from different groups (different disciplines or field of study) then the item could be biased or its function is not the same across these groups, and this item is a DIF item. Raju and Ellis (2002) indicate 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., male and female) have the same probability of choosing the same option on the item.

There are several methods that could be used to detect DIF: Analysis of variance method, transformed item difficulty, item discrimination index, chi square, Mantel Haenzel, and Item response theory methods. One of the strong applications of item response theory is detecting DIF (Hambleton & Swaminathan, 1985). Based on IRT, the item which does not have the same Item Characteristic Curve (ICC) for different groups is considered to be functioning differently between these groups. Different ICCs mean that instrument takers who have the same level of a measured trait do not have the same probability of endorsing the same item (Embrestone & Reise, 2000; Camilli & Shepard, 1994).

**Method**

**Instrument**

In this study, a course Evaluation Survey (CES) was used to collect the data. CES contains 14 five point-Likert Scale items divided into two subscales (instructor and course related items), and it is approved by Imam Abudalruman Bin Faisal University and adopted by Saudi National Commission for Academic Accreditation and Assessment (NCAAA) for accreditation purposes. CES was developed by a panel of experts in related areas, and several studies investigated its psychometric properties and usefulness (Al Rubaish, Wosornu, & Dwivedi, 2012; Al Rubaish, Wosornu, & Dwivedi, 2011; Al-Kuwaiti & Maruthamuthu, 2014). Additionally, Corrected-Item-To-Total Correlation and Cronbach's Alpha were calculated from a random sample (n=50) selected from the current data. The results show that Cronbach's Alpha equals (0.963) and the Corrected-Item-To-Total Correlation ranges from (0.568 to 0.888) which adds evidence for the reliability and the validity of CES.

**Data collection**

The data used in this study is part of data collected by Imam Abudalruman Bin Faisal University during the academic year 2013/2014 for accreditation and monitoring purposes which is going to be submitted to NCAAA. Imam Abudalruman Bin Faisal University through the Deanship of Quality and Academic Accreditation has developed a special application called “UDQUEST” (https://udquest.uod.edu.sa/Login/index.html) that is used electronically to collect data related to many different issues at the university.

One of the surveys available in UDQUEST is the Course Evaluation Survey (CES). This survey is used to evaluate instructors and courses at the university, and is distributed every semester to all students registered in every course offered that semester. The number of electronic surveys analyzed in this study is 36459, and the number of courses evaluated is 866 from 21 colleges and 7 campuses (the same college might be in more than one campus) at the Imam Abudalruman Bin Faisal University. At Imam Abudalruman Bin Faisal University, the colleges are grouped in four different clusters (Health colleges cluster, Engineering colleges cluster, Science and Administration colleges cluster, and Arts and Education colleges cluster). For the purpose of this study, three clusters were selected randomly. Therefore, the data analyzed in the current study are from colleges of: education, engineering, and health. Most
Research questions
1. Does the SET’s survey contain DIF items based on students’ academic discipline?
2. What is the effect of detected DIF items on the SET’s internal structure validity?

Data analysis
For the purpose of this study, the Extended Rasch Model using the RUMM2020 program was used. The model deals with the possibility of different thresholds between the categories (Andrich, 1988, 2005; Ostini & Nering, 2006), and Rasch model helps users to develop scales with strong psychometric properties including greater generalizability (Embreston & Reise, 2000; Hambleton, Swaminathan, & Rogers, 1991). It enables users to create an interval scale of scores for both items and persons (ability) that are sample independent and interrally measured (Bond & Fox, 2001). These scores are reported in units called logits and are typically placed on a vertical ruler called “logistic ruler”. The interpretation of the logits is similar to z-score as person and item logits range from minus three to plus three with a mean of zero and a standard deviation of one. The minus values indicate lower performing persons and easy items (easy to be endorsed item), whereas the plus values indicate higher performing persons and more difficult items. The logistic ruler measures persons’ abilities on one side of the ruler and item difficulties on the other. Because logits can be added, subtracted, multiplied, and divided, comparisons and statistical studies can be made and that makes it useful for showing educational gains, displaying strengths and weaknesses, and comparing groups. One test can be compared with another and people’s ability measures may be also compared with different tests (RUMM Laboratory, 2005).

Results
CES Scale Fitting Results
All CES items were used in the analysis, and the results showed that the Person Separation index (IRT equivalent of Cronbach Alpha) is .873. This means that about 13% of the variability is due to measurement error, or 87.3% of the variance is accounted by the model.

Unidimensionality as an assumption of using Rasch Model was assessed using factor analysis with the ratio of first-to-second eigenvalue greater than 2 (Slocum-Gori & Zumbo, 2011). The results show that there are two factors with eigenvalues greater than 1 and that explain 55% of the variance. The first factor’s eigenvalue 6.02 explains 43% of the variance and the second one 1.68 explains 12% of the variance. The first-to-second ratio is 3.58, which indicates that the unidimensionality assumption is met.

Detecting Differential Item Functioning (DIF)
To answer the first research question “Does the SET survey contain DIF items based on students’ academic discipline?” , the possibility of discipline differences in responses to CES items was tested by DIF analysis with a Bonferroni-adjusted value (p=.0035) using RUMM2020 program. RUMM 2020 uses Two Way ANOVA on the item location to assess DIF for every item where the main effects are the discipline (cluster) and the class interval (trait level). The part of the ANOVA Table of interest is the interaction between Class Interval (trait level) and Group (Discipline). If there is a significant interaction, this means that students from the same trait level and different disciplines endorse the item differently, and that is considered evidence of DIF. The summary of this analysis is shown in Table 1. Because of the multiple comparisons involved (14 items), the alpha level is adjusted using the Bonferroni correction due to the number of items yield an alpha of .05/14=.0035 (Thompson, 2006).
Table 1 shows that CES has 4 items with a significant ability by discipline interaction (DIF) after applying the Bonferroni correction which takes into account the familywise error. These items are: 1- My professor used up-to-date and useful course materials (texts, hand-outs, references, etc.), 2- My professor inspired me to do my best work, 3- My professor gave me the marks for continuous assessment on time, and 4- My professor provided effective IT (Information Technology) to support my learning. Figures 1-4 show Item Characteristics Curves (ICC) for each DIF item.

Figure 1: Differential Item Functioning graph of the three colleges (disciplines) for item 7

Figure 2: Differential Item Functioning graph of the three colleges (disciplines) for item 9.
Differential Item Functioning in Students Rating of Teaching Effectiveness Surveys

Mahmoud AlQuraan & Ahmed AL Kuwaiti

Vol.11 Issue 4, 2017

Figure 3: Differential Item Functioning graph of the three colleges (disciplines) for item 11.

Figure 4: Differential Item Functioning graph of the three colleges (disciplines) for item 13.

Figures 1-4 demonstrate items that have DIF or biased items. Students from different disciplines perceived these items differently despite that students having the same level of rating for courses and teachers. Figures 1 and 4 show that college of education students’ ratings of these items are lower than those of their peers in colleges of engineering and health in almost all the ability levels. This means that these items are behaving in different way than it is expected by the model. These items are: “My professor used up-to-date and useful course materials (texts, hand-outs, references, etc.)” and “My professor gave me the marks for continuous assessment on TIME”. This result indicates that students from different colleges perceive or even understand these items differently. On the other hand, Figure 2 shows that colleges of health students’ ratings of Item 9 “My professor inspired me to do my best work” is lower than colleges of engineering and education students. This indicates that students from different colleges or disciplines perceive and understand this teaching practice differently.

DIF items effects on CES Construct Validity

CES is designed to measure students rating of the course. Therefore, the items are focusing on practices that are an interaction between the course content and teaching practices. The results of these study showed evidence that the unidimensionality of CES has been met. So, single factor confirmatory analysis was used to answer the second research question which is "What is the effect of detected DIF items on SET’s internal structure validity?" Single factor Confirmatory Factor Analysis (CFA) was used twice using LISREL Structural Equation Modeling Software; in the first one all the CES items were included, the second time Detected DIF items were deleted. The results of the two analyses are shown in Table 2.
Table 2 shows the following fit indices for the two models: Chi-Squared ($\chi^2$), Root Mean Square Error of Approximation (RMSEA), Expected Cross-Validation Index (ECVI), Population Discrepancy Function Value (F0), and Estimated Non-centrality Parameter (NCP). Schermelleh-Engel, Moosbrugger and Müller (2003) indicated that two CFA models can be compared by calculating the $\chi^2$ difference of the models. Table 2 shows that the $\chi^2$ differences equals (28321.52) and DF equals (44) and this difference is statistically significant at $\alpha=0.01$. Also, by taking the other fit indices into account, Table 2 suggests that removing DIF items from CES has improved the fit indices of the model. This indicates that eliminating DIF items has improved the construct validity of the instrument.

Discussion

After more than seven decades of research on SETs in higher education, most researchers believe that they are reliable, valid, and useful (Wachtel, 1998). This study provides evidence that disciplines or students' college could be a source of item bias or DIF which indicates that students who have the same level of ability or ratings of teaching effectiveness but one from different disciplines or colleges understand and perceive some items in CES differently and therefore they respond to these items in a different manner. The results of this study show that there are four items that are performed differently by students' discipline after controlling for the level of perceived teaching effectiveness. This type of error (DIF items) is a major threat of instrument validity and reliability (Duncan, 2006; Monahan, 2002). This result supports with the results of Marsh (1984, 2007) and Marsh and Roche (1997) whose studies focused on the importance of wording SETs items to prevent CESs surveys from multidimensionality. Therefore, higher education institutes should use and prepare surveys to assess teaching and teachers that are free of DIF items. In other words the items should be perceived or understood in the same way despite students' college or discipline.

One of the items that showed DIF in CES is "My professor used up-to-date and useful course materials (texts, hand-outs, references, etc.)." The content of this item focuses on using up-to-date references and texts. Since the colleges included in the current study are Engineering, Health, and Education, it is expected that the importance of using up-to-date references and texts to be different according the students' college. The need for up-to-date references and texts for the engineering students is more important than it is for college of education students. This could explain why students from different colleges perceive different meanings of this item's contents, and therefore the results show that it has a DIF. Another item that has a DIF in CES according to the current study is "My professor provided effective IT (Information Technology) to support my learning". Although, using IT to enhance learning is needed for all students despite their college, the volume of this need might be different from one college to another. The real practices in classrooms at the university show more IT involvement by college of engineering teachers compared to their colleagues at college of education. This might be the reason behind the existence of DIF for this item.

Also, the results show that removing the detected DIF items from CES enhances its construct validity. Unfortunately, Oon and colleagues (2017) reported that SETs are rarely assessed psychometrically which might lead to potential consequences by providing inaccurate and invalid information, and therefore SETs' results for courses and teachers cannot be justified. Based on the results of this study, it is recommended to investigate DIF sources in SETs' surveys and make sure that the surveys used by higher education institutes are free of DIF items. Fairness of the SETs' surveys is questionable when these surveys contain DIF items.

Conclusions and Implications for further Research

The current study has shown evidence that some Likert-type items in the SET survey function differently across students' college,
and the CFA has shown that removing the detected DIF items from SET survey enhances its internal structure validity. Since the current study is based on one data set (N=36459) and one university’s experience, it is recommended to conduct more research using different data sets from different universities. Also, it is recommended to examine other possible sources of DIF in SET surveys according to other variables such as students’ level, gender, and GPA.

References


