Differential Item Functioning in Students Rating of Teaching Effectiveness Surveys in Higher Education According to Academic Disciplines: Data from a Saudi University

Mahmoud AlQuraan*
Yarmouk University, Jordan

& Ahmed AL Kuwaiti
Imam Abudalruman Bin Faisal University, KSA

Received: 30/3/2017 Accepted: 4/6/2017

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.

الاداء التفاضلي للفقرات في استبانات تقييم الطلبة للفاعلية التدريسية في التعليم العالي حسب حقل المعرفة الاكاديمي: بيانات من جامعة سعودية

محمود القرعان*
جامعة اليرموك، الأردن

وأحمد الكويتي
جامعة الإمام عبد الرحمن بن فيصل، السعودية

مستخلص: هدفت الدراسة الى تقصي اثر حقل المعرفة (التخصص) للطالب كمصدر للأداء التفاضلي في فقرات تقييم فعالية وجودة التعليم العالي من وجهة نظر الطالب. ولتحقيق هدف الدراسة، تم الاستفادة من البيانات التي قامت جامعة الامام عبد الرحمن بن فيصل بجمعها لأغراض الحصول على الاعتماد الاكاديمي العام. وقد تم تحليل (٣٦٤٥٩) استبانة في ثلاث كليات (التربية، الصحة، الهندسة) باستخدام نموذج راش المعدل. وأظهرت النتائج وجود اربع فقرات ابدت اداءا تفاضليا حسب الكلية، واكد محتوى الفقرات احتمالية ان تكون تلك الفقرات متحيزة لكلية دون اخرى. كما بينت الدراسة ان حذف الفقرات ذات الاداء التفاضلي يسهم وبدلالة احصائية في تحسين صدق البناء للأداة.

الكلمات المفتاحية: الأداء التفاضلي للفقرات، تقويم الطلبة للتدريس، الجودة، التعليم العالي.

*mahmoud.q@yu.edu.jo

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 & Hoshower, 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 multidimentional, 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 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

of the selected colleges are available at more than one campus.

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 intervally 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
Two Way ANOVA Summary Table for Assessing DIF

Item	Class Interval (Ability)			Discipline			Ability Level by Discipline		
	MS	F	Prob	MS	F	Prob	MS	F	Prob
I0001	41.35	48.21	<.001	9.07	10.57	<.001	0.72	0.84	0.651
I0002	69.33	85.15	<.001	6.11	7.51	<.001	1.01	1.25	0.214
I0003	80.21	99.6	<.001	30.52	37.9	<.001	1.08	1.34	0.152
I0004	28.55	31.72	<.001	4.87	5.41	0.004	0.51	0.57	0.922
I0005	83.17	103.46	<.001	33.16	41.25	<.001	-0.61	-0.76	0.999
I0006	43.94	48.27	<.001	29.92	32.87	<.001	0.23	0.25	0.999
I0007	14	14.53	<.001	87.37	90.68	<.001	4.23	4.39	0.0006*
I0008	69.75	86.61	<.001	4.01	4.97	<.001	0.7	0.87	0.618
I0009	81.84	103.54	<.001	25.41	32.15	<.001	1.75	2.22	0.001*
I0010	15.28	16.91	<.001	6.27	6.94	<.001	0.74	0.82	0.675
I0011	8.75	9.03	<.001	26.56	27.38	<.001	2.09	2.16	0.002*
I0012	607.2	448.02	<.001	108.76	80.25	<.001	-2.15	-1.59	0.999
I0013	387.17	307.36	<.001	36.93	29.32	<.001	4.36	3.46	<.0001*
I0014	18.13	18.6	<.001	83.58	85.73	<.001	-0.7	-0.71	0.999

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.

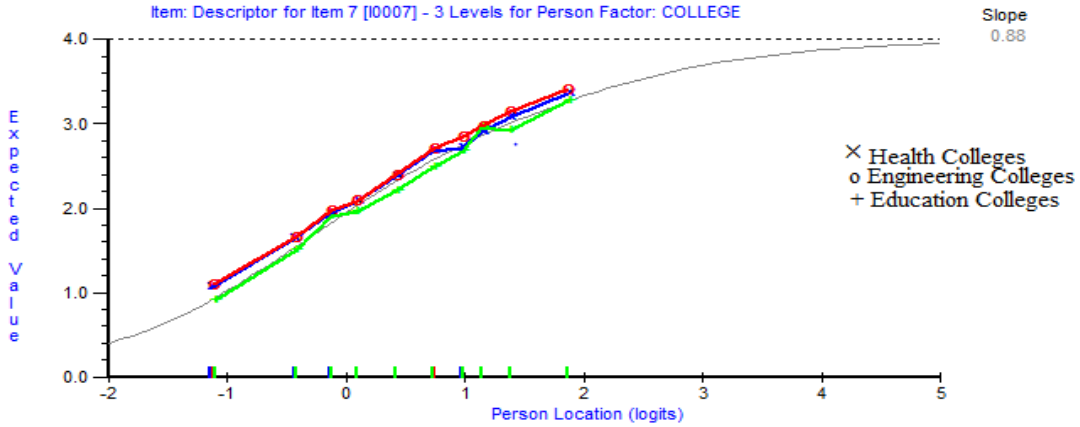

***Figure 1*: Differential Item Functioning graph of the three colleges (disciplines) for item 7**

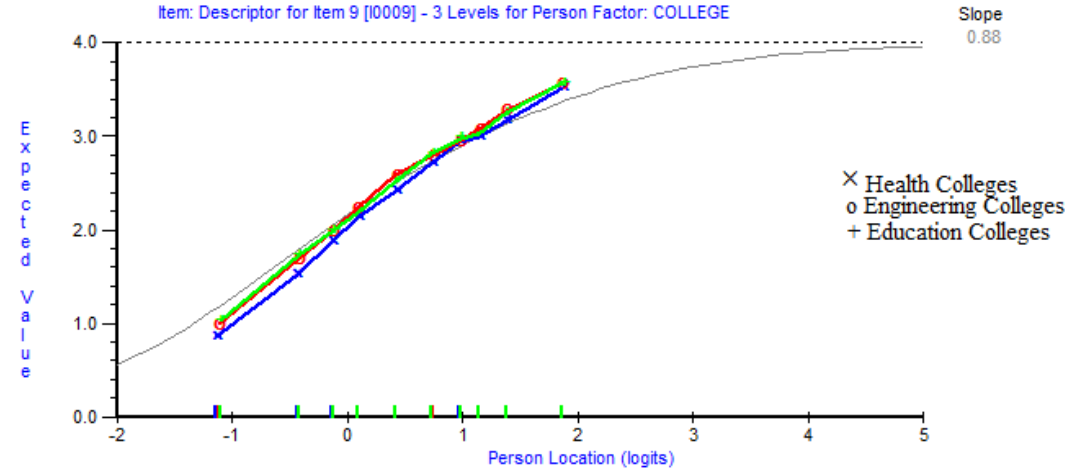

***Figure 2*: Differential Item Functioning graph of the three colleges (disciplines) for item 9.**

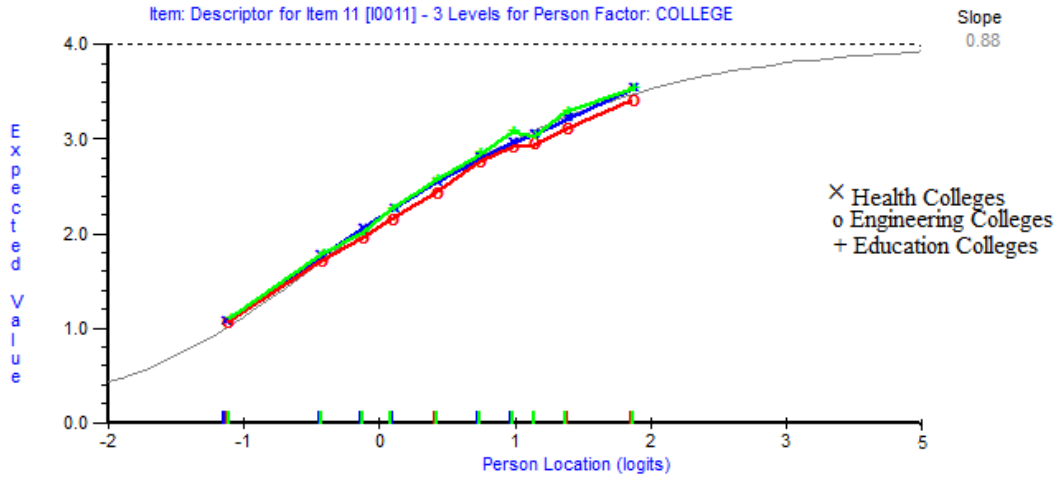

Figure 3: **Differential Item Functioning graph of the three colleges (disciplines) for item 11.**

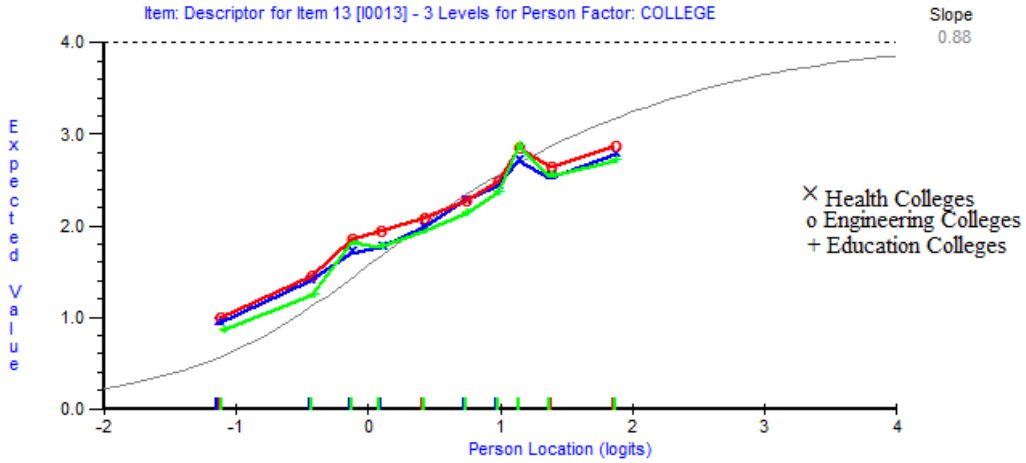

Figure 4: Differential Item Functioning graph of the three colleges (disciplines) for item 13.

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 unidimentionality 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.
Fit Indices for the two models (After and Before Deleting DIF items)

Fit Indices	Befor deleting DIF items		After Deleting DIF items	
	Value	CI: 95%	Value	CI: 95%
χ^2	51153.26	DF=90	22831.74	DF=44
RMSEA	.12	.12-.13	.12	.12-.12
ECVI	1.40	1.38-1.43	.63	.61-.64
FO	1.40	1.38-1.42	.63	.61-.63
NCP	51063.26	50320.94-51810.53	22787.74	22293.62-23288.23
χ^2	51153.26	DF=90	22831.74	DF=44

Table 2 shows the following fit indices for the two models: Chi-Squared (χ^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 χ^2 difference of the models. Table 2 shows that the χ^2 differences equals (28321.52) and DF equals (44) and this difference is statistically significant at α=.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 multidimentionality. 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

Addison, W. E., Best, J., & Warrington, J. D. (2006). Students' Perceptions of Course Difficulty and Their Ratings of the Instructor. *College student journal*, *40*(2), 409-416.

Al Rubaish, A., Wosornu, L., & Dwivedi, S. N. (2011). Using Deductions from Assessment Studies towards Furtherance of the Academic Program: An Empirical Appraisal of Institutional Student Course Evaluation, *iBusiness*, 3(2), 220-228.

Al Rubaish, A., Wosornu, L., & Dwivedi, S. N. (2012). Appraisal of Using Global Student Rating Items in Quality Management of Higher Education in Saudi Arabian University. *iBusiness*, *4*(1), 1-9.

Al-Kuwaiti, A. & Maruthamuthu, T. (2014). Factors influencing student's overall satisfaction in course evaluation surveys: An exploratory study. *International Journal of Education and Research*, *2*(7), 661-674.

Al Kuwaiti, A., AlQuraan, M., & Subbarayalu, A. V. (2016). Understanding the effect of response rate and class size interaction on students evaluation of teaching in a higher education. *Cogent Education*, *3*(1), 1204082.

Anders, S., Pyka, K., Mueller, T., von Steinbuechel, N., & Raupach, T. (2016). Influence of the wording of evaluation items on outcome-based evaluation results for large-group teaching in anatomy, biochemistry and legal medicine. *Annals of Anatomy-Anatomischer Anzeiger*, *208*, 222-227.

Andrich, D. (1988). *Rasch models for measurement*. Thousand Oaks: Sage.

Andrich, D. (2005). *Probabilistic models for some intelligence and attainment tests*. Chicago: The University of Chicago Press.

Archibong, I. A., & Nja, M. E. (2011). Towards improved teaching effectiveness in Nigerian public universities: Instrument design and validation. *Higher Education Studies*, *1*(2), 78-91.

Badri, M. A., Abdulla, M., Kamali, M. A., & Dodeen, H. (2006). Identifying potential biasing variables in student evaluation of teaching in a newly accredited business program in the UAE. *International Journal of Educational Management*, *20*(1), 43-59.

Basow, S. A., & Montgomery, S. (2005). Student ratings and professor self-ratings of college teaching: Effects of gender and divisional affiliation. *Journal of Personnel Evaluation in Education*, *18*(2), 91-106.

Beran, T., & Violato, C. (2005). Ratings of university teacher instruction: How much do student and course characteristics really matter?. *Assessment & Evaluation in Higher Education*, *30*(6), 593-601.

Berk, R. A. (2005). Survey of 12 strategies to measure teaching effectiveness. *International Journal of Teaching and Learning in Higher Education*, *17*(1), 48-62.

Bond, T., & Fox, C. (2001). *Applying the Rrasch model: Fundamental measurement in the human science*. New York: Lawrence Erlbaum Associate.

Brockx, B., Spooren, P., & Mortelmans, D. (2011). Taking the grading leniency 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.

Camilli, G., & Shepard, L. (1994). *Methods for identifying biased items* (Vol. 4). Thousand Oaks: Sage.

Chen, G. H., & Watkins, D. (2010). Stability and correlates of student evaluations of teaching at a Chinese university. *Assessment & Evaluation in Higher Education*, *35*(6), 675-685.

Chen, Y., & Hoshower, L. B. (2003). Student evaluation of teaching effectiveness: An assessment of student perception and motivation. *Assessment & evaluation in higher education*, *28*(1), 71-88.

Clayson, D. E. (2009). Student evaluations of teaching: Are they related to what students learn? A meta-analysis and review of the literature. *Journal of Marketing Education*, *31*(1), 16-30.

Clayson, D. E., & Sheffet, M. J. (2006). Personality and the student evaluation of teaching. *Journal of Marketing Education*, *28*(2), 149-160.

Contreras-McGavin, M., & Kezar, A. J. (2007). Using qualitative methods to assess student learning in higher education. *New Directions for Institutional Research*, *2007*(136), 69-79.

Dommeyer, C. J., Baum, P., Hanna, R. W., & Chapman, K. S. (2004). Gathering faculty teaching evaluations by in-class and online surveys: their effects on response rates and evaluations. *Assessment & Evaluation in Higher Education*, *29*(5), 611-623.

Duncan, S. (2006). Improving the Prediction of Differential Item Functioning: A Comparison of the Use of an Effect Size for Logistic Regression DIF and Mantel-Haenszel DIF Methods. Unpublished Dissertation, Texas A&M University.

Embreston, S. & Reise, S. (2000). *Item response theory for psychologist*. NJ: Lawrence Erlbaum Associates.

Giles, A., Martin, S. C., Bryce, D., & Hendry, G. D. (2004). Students as partners in evaluation: Student and teacher perspectives. *Assessment & Evaluation in Higher Education*, *29*(6), 681-685.

Goos, M., & Salomons, A. (2017). Measuring teaching quality in higher education: assessing selection bias in course evaluations. *Research in Higher Education*, *58* (4), 341-364.

Greimel-Fuhrmann, B., & Geyer, A. (2003). Students' evaluation of teachers and instructional quality--Analysis of relevant factors based on empirical evaluation research. *Assessment & Evaluation in Higher Education*, *28*(3), 229-238.

Griffin, T. J., Hilton III, J., Plummer, K., & Barret, D. (2014). Correlation between grade point averages and student evaluation of teaching scores: taking a closer look. *Assessment & Evaluation in Higher Education*, *39*(3), 339-348.

Gump, S. E. (2007). Student evaluations of teaching effectiveness and the leniency hypothesis: A literature review. *Educational Research Quarterly*, *30*(3), 56-69.

Hambleton, R., & Swaminathan, H., & Rogers, H. (1991). *Fundamentals of item response theory*. London: SAGE.

Hambleton, R., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Dordrecht, Netherlands: Kluwer. Nijhoff Publishing.

Harnish, R. J., & Bridges, K. R. (2011). Effect of syllabus tone: students' perceptions of instructor and course. *Social Psychology of Education*, *14*(3), 319-330.

Kogan, L. R., Schoenfeld-Tacher, R., & Hellyer, P. W. (2010). Student evaluations of teaching: perceptions of faculty based on gender, position, and rank. *Teaching in Higher Education*, *15*(6), 623-636.

Koh, H. C., & Tan, T. M. (1997). Empirical investigation of the factors affecting SET results. *International Journal of Educational Management*, *11*(4), 170-178.

Lattuca, L. R., & Domagal-Goldman, J. M. (2007). Using qualitative methods to assess teaching effectiveness. *New Directions for Institutional Research*, *2007*(136), 81-93.

MacNell, L., Driscoll, A., & Hunt, A. N. (2015). What's in a name: exposing gender bias in student ratings of teaching. *Innovative Higher Education*, *40*(4), 291-303.

Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of educational psychology*, *76*(5), 707-754.

Marsh, H. W. (2007). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. In *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 319-383). Springer Netherlands.

Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist*, *52*(11), 1187.

Martin, J. (1998). Evaluating faculty based on student opinions, problems, implications and recommendations from Deming's theory of management perspective. *Issues in Accounting Education* 13: 1079_94.

McPherson, M.A., and R.T. Jewell. 2007. Leveling the playing field: Should student evaluation scores be adjusted? *Social Science Quarterly* 88: 868_81.

Monahan, P., (2002). The Mantel-Haenszel Procedure For DIF: Alternative Matching Scores to Control Type 1 Error and Improve Distributional Properties. Unpublished Dissertation, The University of Iowa.

Neumann, R. (2001). Disciplinary differences and university teaching. *Studies in Higher Education*, *26*(2), 135-146.

Oon, P. T., Spencer, B., & Kam, C. (2017). Psychometric quality of a student evaluation of teaching survey in higher education. *Assessment & Evaluation in Higher Education*, *42*(5), 788-800.

Ory, J. C., & Ryan, K. (2001). How do student ratings measure up to a new validity framework?. *New directions for institutional research*, *2001*(109), 27-44.

Ostini, R., & Nering, M. (2006). *Polytomous item response theory*. London: SAGE.

Patrick, C. L. (2011). Student evaluations of teaching: effects of the Big Five personality traits, grades and the validity hypothesis. *Assessment & Evaluation in Higher Education*, *36*(2), 239-249.

Penny, A. R. (2003). Changing the agenda for research into students' views about university teaching: Four shortcomings of SRT research. *Teaching in higher education*, *8*(3), 399-411.

Raju, N., & Ellis, E. (2002). Differential Item and Test Functioning. In Drasgow, F and Schmitt, N (Eds). *Measuring and Analyzing Behavior in Organizations; Advance in Measurement and Data Analysis.* (pp. 123-155). San Francisco: Jossey-Bass A Wiley Company.

Rantanen, P. (2013). The number of feedbacks needed for reliable evaluation. A multilevel analysis of the reliability, stability and generalisability of students' evaluation of teaching. *Assessment & Evaluation in Higher Education*, *38*(2), 224-239.

Rienties, B. (2014). Understanding academics' resistance towards (online) student evaluation. *Assessment & Evaluation in Higher Education*, *39*(8), 987-1001.

RUMM Laboratory .(2005). *RUMM 2020 Rasch unidimensional measurement models (Computer Program and manual).*

Santhanam, E., & Hicks, O. (2002). Disciplinary, gender and course year influences on student perceptions of teaching: Explorations and implications. *Teaching in Higher Education*, *7*(1), 17-31.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003).Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of psychological research online*, *8*(2), 23-74.

Shevlin, M., Banyard, P., Davies, M., & Griffiths, M. (2000). The validity of student evaluation of teaching in higher education: love me, love my lectures? *Assessment & Evaluation in Higher Education*, *25*(4), 397-405.

Slocum-Gori, S. L., & Zumbo, B. D. (2011). Assessing the unidimensionality of psychological scales: Using multiple criteria from factor analysis. *Social Indicators Research*, *102*(3), 443-461.

Spencer, P. A., & Flyr, M. L. (1992). The formal evaluation as an impetus to classroom change: Myth or reality? http://eric.ed.gov/?id=ED349053

Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching: The state of the art. *Review of Educational Research*, 83(4), 598-642.

Sproule, R. (2000). Student evaluation of teaching: A methodological critique of conventional practices. *Education Policy Analysis Archives*, *8*(2), 1-23.

Thompson, B. (2006). *Foundation of behavioral statistics*. New York: Guilford Press.

Ting, K. F. (2000). A multilevel perspective on student ratings of instruction: Lessons

from the Chinese experience. *Research in Higher Education*, *41*(5), 637-661.

Uttl, B., White, C. A., & Gonzalez, D. W. (2016). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation*, 54, 22-42. http://dx.doi.org/10.1016/j.stueduc.2016.08.007.

Wachtel, H. K. (1998). Student evaluation of college teaching effectiveness: A brief review. *Assessment & Evaluation in Higher Education*, *23*(2), 191-212.

Weinberg, B. A., Fleisher, B. M., & Hashimoto, M. (2007). *Evaluating methods for evaluating instruction: The case of higher education* (No. w12844). National Bureau of Economic Research.

Wolbring, T., & Riordan, P. (2016). How beauty works. Theoretical mechanisms and two empirical applications on students' evaluation of teaching. *Social science research*, *57*, 253-272.

Yao, Y., and Grady, M.L. 2005. How do faculty make formative use of student evaluation feedback? A multiple case study. *Journal of Personnel Evaluation in Education* 18(2), 107_26.