iojes
www.iojes.net

International Online Journal of Educational Sciences

ISSN: 1309-2707

Measurement Invariance of Student Evaluation of Teaching across Groups defined by Course-Related Variables

İlker Kalender[1]

[1]*Graduate School of Education, Ihsan Dogramaci Bilkent University*

ARTICLE INFO

Article History:
Received 13.04.2015
Received in revised form 27.09.2015
Accepted 11.10.2015
Available online 16.10.2015

ABSTRACT

In the present study, comparability of scores from student evaluation of teaching forms was investigated. This is an important issue because scores given by students are used in decision making in higher education institutions. Three course-related variables (grade level, course type, and course credit) were used to define student subgroups. Then, multi-group confirmatory factor analysis was used to assess invariance of factorial structure, factor loadings and factor means across groups. It was found that although a common factorial structure held across groups, fully invariant factor loadings were observed only across instructors who teach different course types. For other groups, only partial invariance of factor loadings was obtained. Analyses also revealed that none of the subgroups had invariant factor means, indicating a possible bias. Results indicate that comparison of instructors based on student ratings may not be valid as it is mostly assumed.

Keywords:
student evaluation of teaching, measurement invariance, comparability

Introduction

Evaluation of instructors based on students' ratings has been a dominant approach in higher education. The most common way to gather feedback is to use standardized student evaluation of teaching (SET) forms which include Likert-type and open-ended items for students to assess different aspects regarding instructor and course. Although there are some other ways to get feedback such as peer observation, self-assessment, etc., SET forms have been standing as the primary source of knowledge for years. Scores given by students are of significant importance to instructors since these scores are used in decision-making, sometimes high-stake, such as promotion, incentives, etc. as well as instructional development (Ehie & Karathanos 1994; Emery, Kramer & Tian, 2003; Kulik, 2001).

Whatever the purpose for using SET forms is, they should be valid instruments. In other words, they should be developed in a way that items are able to measure instructor-related traits, i.e. instructional effectiveness. Otherwise, decisions to be made may be misleading (Solomon et al., 1997). Unfortunately, validity is one of the points on which SET has been receiving severe criticism. Some researchers found that there are some factors affecting students' opinions, stating SET forms measure different aspects other than what they are designed to measure. For example, grading policy of instructor is probable one of the most discussed factors as influential on students' scores. Students who expect higher grades in a class tend to give higher scores for their instructors and those who have lower expectations in grades may use the scores to punish instructors (Greenwald & Gillmore, 1997; Nimmer & Stone, 1991; Sailor, Worthen & Shin, 1997; Scherr & Scherr, 1990; Trick, 1993; Wilson, 1998). But other research showed the contradictory results reporting no relationship between grading leniency and students' ratings (Baird, 1987; Gigliotti & Buchtel, 1990; Greenwald & Gilmore, 1997). Another factor is grade level of the course. Studies by Braskamp and Ory

[1] Corresponding author's address: Graduate School of Education, Ihsan Dogramaci Bilkent University
Telephone: +90 312 290 1095
Fax: +90 312 266 4065
e-mail: kalenderi@bilkent.edu.tr
DOI: http://dx.doi.org/10.15345/iojes.2015.04.006

(1994), Donaldson et al. (1993), and Conran (1991) showed students at upper grade levels give higher scores of instructors. Some researchers explain this relationship by maturity of students (Kalender, 2011) or high motivation leading to high learning at higher grades (Cashin, 1990). Similarly, Aleamoni and Hexner (1980), and Goldberg and Callahan (1991) reported that students give higher scores at graduate level courses as compared to undergraduate courses. Instructors who teach elective and must courses also have differences in SET scores. Students at elective courses give higher scores to their instructors (Kockelman, 2001). This is probably due to the fact that students are more motivated when they select their courses. Credit was also shown to be related with SET scores. Courses with higher credits are more taken seriously by students who put more effort in these classes (Kalender, 2011).

As to effect of these variables, not exhaustively, on validity of the SET forms, Beran and Violato (2005) found that the effects were minimal, concluding that scores given by students measure instructional effectiveness. Similarly, in a recent study, Zhao and Gallant (2012) showed that SET forms have empirical support for construct validity using a confirmatory factor analysis. They showed that a large amount of the variance (49% to 58%) in SET scores is explained by items in the forms. The rest of the variability is mostly explained by some other variables which were stated above. Besides these studies, there are other studies in the literature supporting validity of SET forms (Abrami, 2001; Lemos, Queiros, Teixeira, & Menezes, 2011; Marsh, 2007; Theall & Franklin, 2001).

Despite the evidence for their validity, SET forms are validated at the developmental stage in order to ensure construct validity. This is generally done at the whole population of university students without considering subgroups of the whole body. But, if the results of SETs forms are used for comparisons among different groups, which is not an uncommon practice, then it should be shown that the construct validity of SET forms exists across all groups, or that they should have measurement invariance (Dimitrov, 2010). In other words, items in the forms should be interpreted the same by all students across different groups which may be defined based on instructor- and course-related characteristics. If the individuals in a different group have different understanding of items, then comparability of these groups and decisions based on the comparisons may not be valid due to the reasons such as uncommon factor structure, construct/method/item bias, etc. As stated above, students' ratings may show variation based on several course-related factors (whether they take a course must or elective/credit of courses they take). Thus demonstration of invariance of a form across groups is a prerequisite condition if one wishes to compare groups and make meaningful inferences.

Despite its importance in decision making including comparison of instructors who teach different groups of students, sometimes comparability is naively assumed and it is not tested (Byrne 1989). Indeed, in the SET context, based on the findings reported in the literature, it is quite rational to expect that a common factor structure may not be applicable for whole body of students to which SET forms is intended to be administered. In other words, construct validity may not be tenable for all in the target group. The main reason for this is because what SET forms measure, i.e., instructional effectiveness, is a multifaceted trait. Kalender (2014) found that different groups have varying definitions of instructional effectiveness based on SET forms. A similar study by Young and Shaw (1999) also defined different instructional profiles based on students' responses to SET forms. Additional support came from studies by Trivedi, Pardos and Heffernan (2011) and Marsh and Hocevar (1991) who identified several subgroups of students showing distinct characteristics different than whole body. Furthermore, it was shown that subgroups students might exhibit varying behaviours in SET forms (Kalender, 2014; Smith & Cranton, 1992; Tenenbaum, 1977). It was the existence of such groups with varying responses to SET items that motivated the researcher to study on invariance of SET forms. The relationships between these variables and student ratings might be a source of measurement invariance and cause comparability problems across groups.

Invariance of scores across different groups is commonly tested via confirmatory factor analysis (CFA) (Byrne, 2004; Chen, Sousa, & West, 2005; Mann, Rutstein, & Hancock, 2009; Vandenberg & Lance, 2000; Wu, Li & Zumbo, 2007). CFA-based check of invariance for an instrument involves a set of hierarchical tests. At each level, the measurement model is made stringent by imposing constraints. In the literature, number of studies that focused on construct validity of SET forms under subgroups is limited. To fill this gap, this study attempted to examine measurement invariance of SET forms across different subgroups varying with respect to some selected course-related variables. The results are expected to provide information about comparability issue of instructors who teach classes with different characteristics based on SET scores.

Methodology

Sample

Students from a non-profit university constituted the sample of the study which included a total of 625 courses with 20388 students enrolled. Distribution of grade levels of courses was as follows: 214 freshmen, 154 sophomores, 126 juniors, and 131 seniors. Most of the courses are must courses (n=338). Fifty-one per cent (n=319) of the classes had one section, while 19.0 and 11.4 per cent the classes were two and three sections, respectively (the rest with more than four sections). Credit of the courses in the sample varied between 2 and 5. Mean end-of-semester grade was 2.33 with a standard deviation of 0.57. Unit of analysis was classes in this study to eliminate within-class interactions which is expected since students may be sharing their opinions when filling out the forms after instructor leaves the class.

Subgroups

Subgroups were defined based on three course-related variables on the basis of availability. The first variable is grade level of courses. Course type (must, elective) was selected as another factor. And last factor is credit of the course. As stated in the Introduction section, these variables were mostly shown to have a relation with SET scores, and thus students may be expected to exhibit differentiation with respect to them. Categories of these variables and number of courses in each category were given in Table 1.

Table 1. Number of students in the groups

Grade Level	n (%)	Course Type	n (%)	Credit	n (%)
1st	214 (34.2)	UG must	338 (54.1)	2-3	480 (76.8)
2nd	154 (24.6)	UG elec.	109 (17.4)	4-5	145 (23.2)
3rd	126 (20.2)	GR must	178 (28.5)		
4th	131 (21.0)				

Note. UG must: Undergraduate must, UG elec.: Undergraduate elective, GR must: Graduate must

Instrument

SET form used in the present study includes 10 items. Students are presented Likert-type items with five options ranging from Strongly Agree (5) to Strongly Disagree (1) with Neutral option (3). The items are (i) The instructor clearly stated course objectives and expectations from students (*expectations*), (ii) The instructor stimulated interest in the subject (*interest*), (iii) The instructor was able to promote effective student participation in class (*participation*), (iv) The instructor helped develop analytical, scientific, critical, creative, and independent thinking abilities in students (*thinking*), (v) The instructor interacts with students on a basis of mutual respect (*respect*), (vi) The instructor was on time and has not missed classes (*timing*), (vii) The instructor taught the course in English (*English*), (viii) Rate the instructor's overall teaching effectiveness in this course (*effective*), (ix) I learned a lot in this course (*learnt*), and (x) The exams, assignments, and projects required analytical, scientific, critical, and creative thinking (*assessment*).

Measurement Invariance

To test the measurement invariance, first a common model (baseline model) is hypothesized and tested to see whether a common factorial structure is tenable across model. This is called configural invariance which imposes no constraint on the model across groups other than the relationships between items and the latent variables. At the second level, weak invariance is tested whether factor loadings are fixed across groups. In other words, this level of invariance involves the comparison that the same meaning is given by individuals within subgroups and, when holds, allows within-group comparisons given for observed items. At that level, between-group differences are not allowed to compare since the existence of a common origin is not yet shown. Third level test, strong invariance, is conducted to see whether different groups have a common starting point. At that level, factor intercepts are fixed as well as factor loadings. If strong invariance does hold, possibility of items bias is eliminated and factor means can be compared across groups. If it does not hold, the same observed scores may have different scores in factor scores across groups due to bias, prohibiting group comparisons based on means.

Although there are higher levels of invariance which can be tested such as latent means, variance, etc. (Milfont & Fischer, 2010), as stated by several researchers, demonstration of configural, weak and strong invariance would be enough for most conditions (Little, 1997; Horn & McArdle, 1992; Vanderberg & Lance, 2000). This suggestion is also in parallel with the context of the present study. When considered that analyses on SET scores generally involve either comparison of item scores separately or calculation of factor means, invariance test for invariant factor loadings (weak invariance) and across groups (strong invariance) would be enough to confirm measurement invariance of SET forms.

Analyses

The SET form was originally designed as unidimensional. Thus this model was hypothesized as the common baseline model to be applicable all groups. After that model was assessed to have a good fit across all groups, invariance analysis was conducted across groups separately within variables. First, the baseline model was tested to see if configural invariance holds. Then weak and strong invariance tests were conducted by fixing factor loadings and factor means across subgroups within each factor, respectively. If invariance did not hold at level, modifications in the model were made and the model was tested again to seek partial invariance. Following suggestions by Byrne et al., (1989), maximum 20% of the parameters were freed when seeking partial invariance.

Normality of data was assessed via LISREL (Jöreskog & Sörbom 2004) and results showed that data were non-normal with highly skewed distributions, which was an expected situation as average of SET scores tend to be above midpoint (Chandreskar et al., 2013; Zumrawi, 2014). Thus, assessment of fit of baseline model was made based on scaled Chi-square statistic developed by Satorra and Bentler (S-Bχ2) (Satorra & Bentler 1988; Satorra & Bentler 2001), an index typically applied under non-normality and with maximum likelihood estimation for the invariance checks. Due to sensitivity of S-Bχ2 to sample size, test results were also supported by several fit indices such as Root Mean Square Error of Approximation (RMSEA), Standardized Root Mean Square Residual (SRMR), Comparative Fit Index (CFI), and Non-normative Fit Index (NNFI). Generally accepted values for an adequate fit are as follows: below .06 for RMSEA, below .05 for SRMR and above .90 for CFI and NNFI (Fan & Sivo, 2005; Hu & Bentler, 1999; Marsh et al 2004; Schermelleh-Engel et al 2003; Vandenberg & Lance, 2000).

The same index (S-Bχ2) and the fit indices were also used to evaluate invariance tests. If differences between two S-Bχ2 tests produce a non-significant result, then the new added constraint does not result in a significant change in the measurement model. In other words, invariance at that level holds (Bryant & Satorra 2012; Satorra & Bentler 1988; Satorra & Bentler 2001). When required, several parameters were fixed for ease of computation.

Checking multicollinearity revealed that variance inflation factor was 2.8 (acceptable level is below 10, Norusis, 2004) and all inter-item correlations were below .85 (Kline, 2005).

Results

Investigation of means across categories showed that instructors who teach classes which are at higher grade levels, graduate must and lower credits received higher ratings from students (Table 2). Scores were mostly left-skewed. Standard deviations revealed that they are similar and this reduced, if not eliminated, possibility of different error variances different due to observed variance. High Cronbach's Alpha values indicated that scores across subgroups were reliable.

First a hypothesized baseline model was tested across subgroups within each course-related variable, with Q1 (expectations) was fixed to 1 as reference. Preliminary CFA conducted via LISREL indicated that 10-item unidimensional model did not provide a good fit, as indicated by fit indices (RMSEA=.12 [90% CI=.09:.15], SRMR=.09, CFI=.87, and NNFI=.88). Two items, *timing* (vi) and *English* (vii) were excluded from the model due to low inter-item correlations and factor loadings. Exploratory factor analysis conducted additionality also revealed that the two items were grouped under a separate factor. The new 8-item unidimensional model had a good fit and defined as the baseline model. Estimated fit indices were: RMSEA=.06 [90% CI=.04:.07], SRMR=.03, CFI=.98, and NNFI=.94. Figure 1 present standardized path coefficients of the unidimensional 8-item model tested on whole sample. All items loadings were acceptable. All path coefficients were significant at .05.

Table 2. Means and standard deviations of items across subgroups

	Grade Level							
	1		2		3		4	
	M	SD	M	SD	M	SD	M	SD
expectations	4.56	0.50	4.59	0.43	4.63	0.38	4.69	0.40
interest	4.28	0.66	4.33	0.64	4.42	0.6	4.55	0.50
participation	4.25	0.66	4.31	0.66	4.41	0.61	4.57	0.49
thinking	4.31	0.61	4.36	0.61	4.42	0.57	4.56	0.51
respect	4.71	0.43	4.70	0.42	4.75	0.37	4.86	0.25
effective	4.38	0.61	4.41	0.58	4.49	0.53	4.59	0.48
learnt	4.21	0.67	4.34	0.57	4.39	0.53	4.46	0.57
assessment	4.26	0.58	4.32	0.57	4.38	0.56	4.51	0.51
Cronbach's Alpha	0.97		0.97		0.98		0.97	

	Course Type						Credit			
	UG must		UG elec		Grad must		2/3		4/5	
	M	SD	M	SD	M	SD	M	SD	M	SD
expectations	4.58	0.46	4.6	0.48	4.67	0.36	4.63	0.43	4.53	0.48
interest	4.31	0.64	4.4	0.67	4.49	0.53	4.43	0.58	4.21	0.69
participation	4.28	0.65	4.4	0.68	4.51	0.52	4.43	0.59	4.16	0.71
thinking	4.34	0.60	4.41	0.66	4.51	0.50	4.44	0.57	4.26	0.65
respect	4.72	0.41	4.75	0.41	4.81	0.33	4.77	0.37	4.69	0.42
effective	4.40	0.59	4.45	0.61	4.56	0.47	4.50	0.53	4.31	0.65
learnt	4.26	0.63	4.34	0.65	4.45	0.49	4.39	0.57	4.14	0.68
assessment	4.29	0.58	4.38	0.59	4.46	0.52	4.39	0.56	4.21	0.58
Cronbach's Alpha	0.97		0.98		0.97		0.98		0.98	

Note. UG must: Undergraduate must, UG elec.: Undergraduate elective, GR must: Graduate must

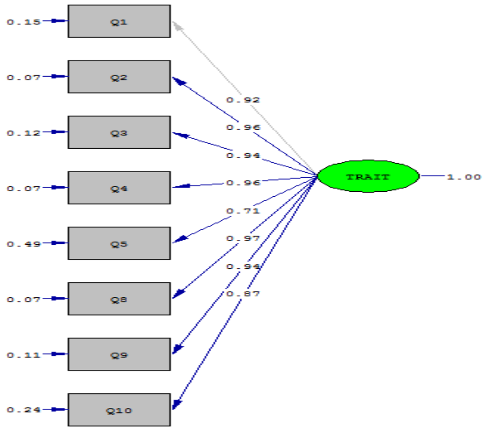

Figure 1. Baseline model *(Q1: expectations, Q2: interest, Q3: participation, Q4: thinking, Q5: respect, Q8: effective, Q9: learnt, and Q10: assessment)*

Table 3 shows the fit indices for the hypothesized baseline model across subgroups. Fit indices indicated that the original model fit the data perfectly.

Table 3. Fit indices for the baseline model in subgroups

	Subgroups	S-Bχ2	RMSEA	SRMR	NNFI	CFI
Grade Level	1st	10.44	.01 (.00;.08)	.01	.99	.99
	2nd	5.53	.00 (.00;.05)	.01	1.00	1.00
	3rd	18.35	.08 (.01;.14)	.01	.97	.99
	4th	19.14	.08 (.02;.14)	.02	.96	.99
Course Type	UG must	6.79	.00 (.00;.04)	.01	.99	.99
	UG elective	20.54	.09 (.03;.16)	.02	.95	.98
	GR must	25.55	.09 (.04;.14)	.02	.97	.99
Credit	2 / 3	25.89	.06 (.03;.09)	.08	.99	.99
	4 / 5	6.21	.00 (.00;.06)	.01	1.00	1.00

Note. UG must: Undergraduate must, UG: undergraduate elective, GR must: Graduate must

One-way ANOVAs for groups based on grade levels and course type, and one-sample t-tests for credit-based groups revealed that all items sere statistically different except the item *expectations* for grade level and type-based groups (Table 4). Statistically different means indicated that groups of students might have been exhibited systematically varying response behaviours, which can be an indicator for the lack of construct validity and an invariance problem.

Table 4. Results of mean difference tests

	Grade Level	Type	Credit
	F	F	t
expectations	2.53	2.41	2.48*
interest	5.56*	5.08*	3.48*
participation	8.01*	7.93*	4.06*
thinking	5.31*	5.07*	3.29*
respect	5.19*	3.32*	2.20*
effective	4.04*	4.75*	3.23*
learnt	5.48*	6.11*	3.88*
assessment	5.69*	5.48*	3.31*

* $p < .05$

Table 5 presents invariance test results for each factor. Configural invariance held for all three variables, indicating one-factor model was acceptable across all subgroups. Check of weak invariance (equality of factor loadings) revealed that the model had invariant factor loadings across undergraduate must and elective, and graduate must courses. However, weak invariance was not observed across courses at different grade levels and with different credits. Although RMSEA, CFI and NNFI indices indicated good fit, S-Bχ2 difference tests yielded statistically significant results, suggesting unequal factor loadings.

Table 5. Results of invariance tests

Invariance Level	Groups	S-Bχ2	df	ΔS-Bχ2	Δdf	p	RMSEA [90% CI]	CFI	NNFI
Configural	Grade Level	101.77	94	-	-	-	.02 [.00;.05]	.98	.98
	Type	64.22	66	-	-	-	.00 [.00;.04]	.98	.98
	Credit	134.94	38	-	-	-	.09 [.07;.11]	.99	.98
Weak	Grade Level	139.30	115	65.93	21	<.001	.04 [.00;.06]	.97	.97
	Type	73.99	80	9.67	14	.79	.00 [.00;.03]	.99	.98
	Credit	160.01	45	23.51	7	<.001	.09 [.08;.12]	.99	.98
Strong	Grade Level	693.18	146	1534.06	44	<.001	.16 [.14;.17]	.93	.92
	Type	522.43	103	4712.89	23	<.001	.14 [.13;.15]	.92	.92
	Credit	918.70	60	1389.99	15	<.001	.21 [.20; .23]	.92	.91

Due to lack of weak invariance, the original model was modified for groups defined by grade level and credit based on the modification indices produced by LISREL. The item with the lowest loading, *respect*, were freed for both of grade level and credit, the models were tested again using the new 7-item model. Across grade levels, weak invariance was satisfied, as indicated by non-significant S-Bχ2 test result and values of the fit indices (S-Bχ2=135.33, df=24, p>.05, RMSEA=.03 [.00;.06], CFI=.97, NNFI=.98). However, modifications were not able to produce any non-significant results for test of strong invariance across grade levels. Similarly, the new model was tested for different credit levels. Results indicated weak invariance was held (S-Bχ2= 19.38, df=11, p>.05, RMSEA=.09 [.08;.11], CFI=.98, NNFI=.98). Like grade level, groups defined by credit of courses did not show strong invariance. Further efforts provided little change in the result of difference test.

Thus, results indicated that, for all three variables, a unidimensional 8-item model was tenable. Factor loadings were the same across groups based only on course type for the originally proposed model. The hypothesized model was also found to be invariant after a minor modification on the model for groups based on grade level and credit. None of the groups had invariant factor means across their subgroups as evıdence by the lack of strong invariance, indicating there might be source of bias across scores for all three course-related variables.

Discussion and Conclusion

The present study focused on comparability of scores from SET forms across student subgroups. Often, scores reported by students are naively assumed to be comparable, which is a false assumption, as indicated by the findings of the present study. To make reliable comparisons, invariance should be held. However, results showed that the SET form did not function similarly across subgroups, indicating construct validity of the SET form did not fully hold across these groups

First of all, the hypothesized 8-item unidimensional model was found to tenable across subgroups in terms of the relationship between items and latent variables. Groups which were defined by course types were fully observed to have invariant factor loadings for originally hypothesized model. However, weak invariance tests revealed factor loadings were not the equal in groups defined based on course grade levels and credit. Thus, item scores and difference scores cannot be compared across these groups using the original model. For groups of grade level and credit, some modifications in the measurement model were made to achieve weak invariance. One of the items (*respect*) in the original model was excluded. After that, weak invariance was observed. In terms of strong invariance, analyses showed that factor means were not invariant in none of the subgroups. Further modifications did not provide any improvement in the invariance. Invariant factor means prohibits comparisons on means, since starting points for the groups were not the same. In other words, scores given for the same opinion produces different factor scores across groups. The lack of strong invariance suggested that the SET form did not have construct validity across different groups due to construct, method, and/or item bias. Although further research is needed, it might be rationale to assume that the main source of unequal factor means is construct bias (there are some traits affecting students' ratings). Rating procedure is simple and wordings are of the items not complex in SET so these two sources of bias can be eliminated.

Reasons for presence of bias, evidenced by variant factor means, can be explained by findings reported in the literature. The lack of strong invariance may be associated with maturity of students. Students who were taught by the same instructors might have different interpretations for instructional practices assessed in SET forms. Higher grade level students may be more experienced on how to assess instructors (Braskamp & Ory, 1994; Donaldson et al., 1993; Conran, 1991; Kalender, 2011). Another reason may be that students at higher grade level courses are motivated since these courses are generally specialized courses. It is known that students' motivation and eager to learn is associated with higher ratings (Cashin, 1990). Similarly, course type is a predictor of student motivation. As reported by Kockelman (2001), when students are allowed to select course, they become more willing to learn, results in higher SET scores. Credit of courses is also a factor associated with the importance given by students for a course. Courses with higher credits are more taken seriously by students and students put more effort in these classes (Kalender, 2011). All these factors may be a cause for developing different rating scale for the instructors. Another explanation for biased scores is that students who rate their instructors with the same ratings may have different levels of opinions. An instructor should show observed actions in the SET forms with different levels to receive the same scores from students (Baas et al., 2011)

One of the points that worth discussion is the model modification. Original model did not have weak invariance, but the modified one was observed to have invariant factor loadings. Even if the partial weak invariance was satisfied for grade level- and credit-based groups, modified model has one item missing (*respect*) and, in turn, may be conceptualizing a different trait. Even if groups defined based on grade level and credits within themselves, instructors rated by groups with different models may not be comparable.

Analyses revealed that comparative analyses which are used in decisions, sometimes high-stakes, regarding instructors should be taken cautiously if instructors teach course with different grade levels, types, and credit, as indicated by that study. Any ranking including instructors with respect to three course-related variables is not a meaningful since instructional effectiveness, or whatever SET form assess, were found to be measured in different scales (Wu, Li, & Zimbo, 2007). However, it should be noted that definitions of groups were arbitrary in the present study. Different formations of groups may yield varying results.

In the present study, group membership was found to be a confounding factor in SET scores. The existence of biased scores should be considered seriously. When this is the case, commonly applied statistical procedures may become unreliable. It is strongly suggested that correction mechanisms should be used if bias is detected. Different means may be incorporated along with SET, such as peer evaluation, self-evaluation, or assessment of the degree to which educational objectives are attained.

Findings obtained in this study confirmed suggestions made by Vandenberg and Lance (2000) and Meredith and Millsap (1992) who stated that invariance of an instrument should be check across groups before making comparisons if one wishes to make meaningful comparisons. The results made research think that there might be other variables such as class size, disciple, etc. on which measurement invariance should be checked. It is recommended that the study should be replicated with other course- and instructor-related variables on which comparisons are made.

References

Abrami, P.C. (2001). Improving judgments about teaching effectiveness using teacher rating forms. In M. Theall, P.C. Abrami, & L. A. Mets (Eds.), The student ratings debate: Are they valid? How can we best use them? [Specialissue]. *New Directions for Institutional Research, 109*, 59-87. doi: 10.1002/ir.4

Aleamoni, L., & Hexner, P. (1980). A review of the research on student evaluations. *Instructional Science, 9,* 67-84. doi: 10.1007/BF00118969

Baas, M., De Dreu, C. K. W., and Nijstad, B. A. (2011). When prevention promotes creativity: the role of mood, regulatory focus, and regulatory closure. *Journal of Personality and Social Psychology, 100*, 794–809. doi: 10.1037/a0022981

Baird, J. S. (1987). Perceived learning in relation to student evaluation of university instruction. *Journal of Educational Psychology*, 79(1), 90-97. doi:10.1177/0273475308324086

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. doi: 10.1080/02602930500260688

Braskamp, L.A., & Ory, J.C. (1994). *Assessing faculty work: Enhancing individual and institutional* performance. San Francisco: Jossey-Bass. doi: 10.1023/A:1007682101295

Bryant, F. B., & Satorra, A. (2012). Principles and practice of scaled difference chi-square testing. *Structural Equation Modeling*, *19*(3), 372-398. doi: 10.1080/10705511.2012.687671

Byrne, B. M. (1989). Multi-group comparisons and the assumption of equivalent construct validity across groups: Methodological and substantive issues. *Multivariate Behavioral Research*, *24*(4), 503–523. doi: 10.1207/s15327906mbr2404_7

Byrne, B. M. (2004). Testing for multigroup invariance using amos graphics: A road less traveled. *Structural Equation Modeling, 11*, 272–300. doi: 10.1207/s15328007sem1102_8

Cashin, W. E. (1990). Students do rate different academic fields differently. *New Directions for Teaching and Learning, 43*, 113–21. doi:10.1002/tl.37219904310

Chandrasekhar, A. J., Durazo-Arvizu, R., Hoyrt, A, & McNulty, J. A. (2013). Do student evaluations influence the teaching skills of clerkship clinical faculty? *Educational Research and Evaluation: An International Journal on Theory and Practice, 19*(7), 628-635. doi:10.1080/13803611.2013.834616

Chen, F. F., Sousa, K. H., & West, S. G. (2005). Testing measurement invariance of second-order factor models. *Structural Equation Modeling, 12*, 471–492. doi:10.1207/s15328007sem1203_7

Conran, P. B. (1991). High school student evaluation of student teachers? How do they compare with professionals. *Illinois School Research and Development, 27*(2), 81-92.

Dimitrov, D.M. (2010). Testing for factorial invariance in the context of construct validation. *Measurement and Evaluation in Counseling and Development, 43*(2), 121-149. doi: 10.1177/0748175610373459

Donaldson, J. F., Flannery, D., & Ross-Gordon, J. (1993). A triangulated study comparing adult college students' perceptions of effective teaching with those of traditional students. *Continuing Higher Education Review, 57*(3), 147–165.

Ehie, I. C. & Karathanos, D. (1994). Business faculty performance evaluation based on the new aacsb accreditation standards. *Journal of Education for Business, 69*(5), 257-26. doi:10.1080/08832323.1994.10117695

Emery, C. R., Kramer T. R., & Tian, R. G. (2003). Return to academic standards: a critique of student evaluations of teaching effectiveness. *Quality Assurance in Education 11*, 37–46. doi:10.1108/09684880310462074

Fan, X., & Sivo, S. (2005). Evaluating the sensitivity and generalizability of SEM fit indices while controlling for severity of model misspecification. *Structural Equation Modeling, 12*(3), 343-367. doi:10.1080/00273170701382864

Gigliotti, R. J., & Buchtel, F. S. (1990). Attributional bias and course evaluations. *Journal of Educational Psychology, 82*, 341-351. doi: 10.1037/0022-0663.82.2.341

Goldberg, G., & Callahan, J. (1991). Objectivity of student evaluations of instructors. *Journal of Education for Business, 66(6),* 377-378. doi: 10.1080/08832323.1991.10117505

Greenwald, A. G., & Gillmore, G. M. (1997) Grading leniency is a removable contaminant of student ratings. *American Psychologist, 52*, 1209-1217. doi: 10.1037/0003-066X.52.11.1209

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to mi in aging research. *Experimental Aging Research, 18*, 117-144. doi: 10.1080/03610739208253916

Hu, L. T., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1-55. doi: 10.1080/10705519909540118

Jöreskog, K. G., & Sörbom, D. (2004). *LISREL 8.7 for Windows* [Computer software]. Skokie, IL: Scientific Software International, Inc.

Kalender, I. (2011). Contaminating factors in university students' evaluation of instructors. *Education and Science, 36*(162), 56-65. doi: 10.1080/10705519909540118

Kalender, I. (2014). Profiling instructional effectiveness to reveal its relationship to learning. *The Asia-Pacific Education Researcher, 23*(3), 717-726. doi: 10.1007/s40299-013-0145-2

Kline, R. B. (2005), *Principles and practice of structural equation modeling* (2nd Edition ed.). New York: The Guilford Press.

Kockelman, K. M. (2001). *Student grades and course evaluations in engineering: What makes a difference?* ASEE Annual Conference Proceedings, 9085–9110.

Kulik, J.A. (2001). Student ratings: Validity, utility, and controversy. In M. Theall, P.C. Abrami, and L.A. Mets (Eds.), The student ratings debate: Are they valid? How can we best use them? [Special issue]. *New Directions for Institutional Research*, *109*, 9-25

Lemos, M. S., Queirós, C., Teixeira, P. M., Menezes, I. (2011). Development and validation of a theoretically based, multidimensional questionnaire of student evaluation of university teaching. *Assessment & Evaluation in Higher Education, 36*(7), 843-864. doi:10.1080/02602938.2010.493969

Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research, 32*, 53-76. doi:10.1207/s15327906mbr3201_3

Mann, H., Rutstein, D. W., & Hancock, G. (2007). The potential for differential findings among invariance testing strategies for multisample measured variable path models. *Educational & Psychological Measurement, 69*(4), 603–612. doi:: 10.1177/0013164408324470

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

Marsh, H. W., & Hocevar, D. (1991). The multidimensionality of students' evaluations of teaching effectiveness: the generality of factor structures across academic discipline, instructor level, and course level. *Teaching and Teacher Education, 7*, 9–18. doi: 10.1016/0742-051X(91)90054-S

Marsh, H. W., Hau, K. T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing hu and bentler's findings. *Structural Equation Modeling, 11*(3), 320-341. doi: 10.1207/s15328007sem1103_2

Meredith, W., & Millsap, R.E. (1992). On the misuse of manifest variables in the detection of measurement bias. *Psychometrika, 57*, 289–311. doi: 10.1007/BF02294510

Milfont, T. L., & Fischer, R. (2010). Testing measurement invariance across groups: Applications in cross-cultural research. *International Journal of Psychological Research, 3*(1), 2011-2084. doi: 10.1177/014920639902500101

Nimmer, J. G., & Stone, E. F. (1991). Effects of grading practices and time of rating on student ratings of faculty performance and student learning, *Research in Higher Education, 32*(2), 195–215. doi: 10.1007/BF00974437

Norusis, M. (2004). *SPSS 13.0 statistical procedures companion*. Upper Saddle-River, NJ: Prentice Hall.

Sailor, P., Worthen, B. & Shin, E. H. (1997). Class level as a possible mediator of the relationship between grades and student ratings of teaching. *Assessment & Evaluation in Higher Education, 22*(3), 261–269. doi: 10.1080/0260293970220301

Satorra, A., & Bentler, P. M. (1988). *Scaling corrections for chi-square statistics in covariance structure analysis*. Proceedings of the American Statistical Association, 308-313.

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika, 66*, 507–514. doi: 10.1007/BF02296192

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

Scherr, F. C., & Scherr, S. S. (1990). Bias in student evaluations of teacher effectiveness, *Journal of Education for Business, 65*(8), 356–358.

Smith, R. A., & Cranton, P. A. (1992). Students perceptions of teaching skills and overall effectiveness across instructional settings. *Research in Higher Education 33*, 747–764. doi: 10.1007/BF00992056

Solomon, D. J., Speer, A J., Rosebraugh, C. J., & DiPette, D. J. (1997). The reliability of medical student ratings of clinical teaching. *Evaluation & Health Professions, 20*(3), 343–52.

Tenenbaum, A. B. (1977). Task-dependent effects of organization and context upon comprehension of prose. *Journal of Educational Psychology, 69*, 528–536. doi: 10.1037/0022-0663.69.5.528

Theall, M., & Franklin, J. (2001). Looking for bias in all the wrong places: A search for truth or a witch hunt in student ratings of instruction. In M. Theall, P. C. Abrame, & L. A. Mets (Eds.), *New directions for institutional research (*pp. 45–56). San Francisco: Jossey-Bass. doi: 10.1002/ir.3

Trick, L. R. (1993). Do grades affect faculty teaching evaluations? *Journal of Optometric Evaluation, 18*(3), 88-92.

Trivedi, S., Pardos, Z. A., & Heffernan N. T. (2011). *Clustering students to generate an ensemble to improve standard test score predictions*. In Proceedings of the 15th International Conference on Artificial Intelligence in Education, Auckland, New Zealand.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the mi literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3,* 4-69. doi: 10.1177/109442810031002

Wilson, R. (1998). New research casts doubt on value of comparing adult college students' perception of effective teaching with those of traditional students. *Chronicle of Higher Education, 44*(19), 12-14.

Wu. A. D., Li, Z., & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment, Research and Evaluation, 12*(3), 1-26.

Young, S. M., & Shaw, D. G. (1999). Profiles of effective college and university teachers. *The Journal of Higher Education, 70*(6), 670–686.

Zhao, J. & Gallant, D. J. (2012). Student evaluation of instruction in higher education: exploring issues of validity and reliability. *Assessment & Evaluation in Higher Education, 37*(2), 227-235. doi: 10.1080/02602938.2010.523819

Zumrawi, A. A., Bates, S. P., Schroeder, M. (2014). What response rates are needed to make reliable inferences from student evaluations of teaching?. *Educational Research and Evaluation, 20*(7-8), 557-563. doi: 10.1080/13803611.2014.997915