Abstract

In the present study comparability or ranking of instructors based on student ratings were investigated under the effect of absenteeism. To this end, invariance of scale properties of student ratings was examined via multi-group confirmatory analysis. Using randomly selected 2098 classes, equality of factorial structure, factor loadings, intercepts and residuals were tested. Results indicated that absent and regularly attending groups have developed the same conceptual meaning for the term instructional effectiveness. Also, ratings in the both groups of classes had a common unit, which makes within-class comparisons of instructors separately for attending and absent groups possible. However instructors who teach classes with absent students systematically receive lower ratings, indicating a bias between the two groups. Student ratings were adjusted against absenteeism to lessen the effect of bias. Results showed significant differences in the rankings of top-rated instructors both before and after the adjustment. Biased ratings pose a serious threat in comparability between instructors who teach absent and attending classes. Thus decisions involving instructors should be supported by other assessment mechanisms.

Keywords: Student rating of instruction, measurement invariance, absenteeism, bias

Introduction

Student ratings of instruction are deemed important as an indicator of instructional performance in higher education institutions. Although they have been used for almost a hundred years by instructors to improve teaching (Remmers 1928), they recently became a critical component in providing administrators with a metric of instructional effectiveness that is used in some critical decisions about instructors' careers such as promotion, extending contract, and granting tenure (Macfadyen et al. 2015). High-stake use of ratings by administrators typically involves either comparing or ranking instructors. However, to be a fair indicator of what they are intended to measure, these ratings should be free of construct-irrelevant factors favouring against one or more student groups. Although there is a large literature on student ratings and confounding factors, number of studies on bias in student ratings due to absenteeism is limited.

Absenteeism in Higher Education

Absenteeism can be defined as a failure to regularly attend timetabled sessions such as seminars, lectures, practical or laboratory classes (Barlow and Fleischer 2011). Students believe that they can make up for what they miss later by several ways such as findings class notes. However, Colby (2004) reported a significant positive relationship between attendance and marks by defining some rules to relate the rate of absenteeism to the probability of failing in a course. These results were recently replicated in a recent study by Newman-Ford, Fitzgibbon, Lloyd and Thomas (2008) and earlier by other studies (Launius, 1997; Martinez, 2001).

Despite of its vital role in student achievement, absenteeism is common in higher education institutions. Romer (1993) reported that one third of university students are regularly absent. Several other studies reported different rates of absenteeism: 18.5% (Marburger, 2001), 25% (Friedman, Rodriguez. & McComb, 2001), between 59% and 70% (Moore, Armstrong and Pearson 2003), and 30% and 50% (Kousalya, Ravindranath, & VizayaKumar, 2006). Absent students miss their opportunity to acquire knowledge, experiences or skills that can be gained only in class. Among these opportunities are face-to-face contact with instructor and other students, working opportunity in groups, chance to evaluate others' opinions.

Some students have valid excuses for their involuntarily absenteeism such as chronic illness, family problems, or having to do paid work (Nicholl & Timmins, 2005), while some other students report trivial ones for their voluntarily absenteeism such as failing to get up earlier or preferring to do other things at the course time (Trivado-Ivern et al., 2013).

However, the academic reasons for absenteeism outweigh non-academic ones. Instructor's effectiveness is a critical predictor of absenteeism. Students report low quality instruction as the principal reason for their absenteeism (Lopez-Bonilla & Lopez-Bonilla, 2015). Another study by Rodríguez et al. (2003) showed that academic organisation, the teaching methodology, the value or usefulness of the lesson, or the teachers' attitudes are related to absenteeism. Students have higher rates of absenteeism in such classes (Babad, Icekson, & Yelinek, 2008). Triado-Ívern et al. (2013) and Rodriguez et al. (2003) found teacher's approaches, scarcely appealing, academic organization, teaching methodology, and teacher attitudes among the factors associated with absenteeism. Massingham and Herrington (2006) reported that motivational, constructivist and authentic teaching processes are directly related to low absenteeism. Hunter and Tetley (1999) interviewed students and reported that students prefer classes in which instructors make topics interesting, present topics hard to make up, make students feel that topics are important. Unexciting and unchallenging instructors and failure to connect what is presented in class to real life are among the other factors associated with student absenteeism (Hughes 2005; Nicholl & Timmins, 2005). Wolbring (2012) identified several factors associated with absenteeism such as course topic, climate among course participants, course- and workload, and timing of the course.

Whether it is related to students' preferences, instructors' teaching style, or any other variables, university students skip classes as a systematic behaviour (Lopez-Bonilla & Lopez-Bonilla, 2015).

Fairness of Student Ratings under the Effect of Absenteeism

When students rate their instructors, they are assumed to have information sufficient information/observation for a sound judgement about their instructors (Wolbring, 2012). This assumption implies that students attend most of, if not all, the classes regularly. Absenteeism gives rise to a question as to the fairness of student ratings. If the non-random absenteeism has an effect on student ratings, their validity becomes questionable.

The relationship between absenteeism and students ratings has been shown in the literature. Wolbring (2012) found that student ratings are biased by absenteeism. Similarly, Martin et al. (2013) found that students absent more frequently rates gave significantly lower ratings, indicating that absenteeism is a predictor of low ratings. Berger and Schleußner (2003) and Babad, Icekson, and Yelinek (2008) pointed out the same relationship. However, these differences may not be regarding as genuine differences between attending and absent students unless scale properties of student ratings (factor loadings, intercepts, etc.) are shown to be invariant between these two groups.

Thus, an empirical examination of equality of scale properties of student ratings under the effect of absenteeism is needed to obtain a better understanding of comparability of instructors using student ratings. However no examination of scale properties of student ratings has been made in the literature. Therefore the question regarding invariance of the ratings between groups with different absenteeism levels still remain unanswered. Its investigation may provide significant information as to fairness of student ratings between groups under the effect of systematic absenteeism. Only after invariance of student ratings is shown, valid comparison can be made across instructors who teach classes with different levels of absenteeism (Dimitrov, 2010; Horn & McArdle, 1992). If invariance is absent, comparing instructors who teach classes with different absenteeism levels based on the ratings may be misleading and result in incorrect decisions regarding promotion, assignment of courses to instructors, etc. (Ehie & Karathanos, 1994). Student ratings which do not have invariance may also be misleading for students and instructors themselves. Instructors may want to compare their ratings across classes with different levels of absenteeism. If ratings become non-invariant due to bias ratings from such classes cannot be compared. Similarly, students use ratings to select courses. Under the effect of absenteeism, information to be used students become unreliable (Schmelkin-Pedhazur et al., 1997).

Invariance analysis provides information about differences in scale properties such as conceptualization of instructional effectiveness and existence of a common unit of measurement and point of origin common to all groups. Such investigation requires testing differences in scale properties between measurement models with sequentially increased constraints (Dimitrov, 2010). Generally it starts with a hypothesized measurement or baseline model to define the relationships between observed and latent variables. Starting with this baseline model, several additional constraints added to the model at each level of invariance and equality of factorial structure (configural invariance), factor loadings (weak invariance), intercepts (strong invariance), and error variances (strict invariance) are hierarchically tested across different groups.

In the specific context of student ratings, configural invariance means that different groups conceptualize instructional effectiveness in the same way (Cheung & Rensvold, 2002). When weak invariance is observed, the importance attributed by absent and attending students to the rating items is the same. Weak invariance indicates that the change in the student ratings for a unit of change in the instructor's performance ins the same in every group. In other words, increasing exposure of an observed instructional practice creates the same amount of change in student ratings across groups. Strong invariance indicates equality between intercepts of item across groups. When it holds, an observed score corresponds to the same factor score in each group. In other words, group membership is not a determinant of a student's ratings. Equality of error variances is considered as an indication of different reliabilities between groups.

Weak invariance allows only comparison of within-group differences. However, at that level, groups are not shown to have a common origin therefore between-class differences cannot be compared. Only when strong invariance holds, scores from different absent and attending students can be compared without having any bias since they have a common metric (Schmitt & Kuljanin, 2008). On the other hand, unequal factor intercepts indicate that observed scores do not represent students' true scores due to lack of a common origin. Strict invariance does not directly limit comparison opportunities. However it should be noted that random error is different across groups.

Effect of Bias on Ranking of Instructors

Another issue that the present study deals with is practical significance of biased ratingsmay have on comparison or rankings of instructors. Despite the evidence of biased ratings due to absenteeism, this bias may not still have practical consequences from a pragmatic point of view after they are controlled using statistical mechanisms. As long as the ranks stay the same after controlling for absenteeism, student ratings can be considered not to be seriously affected by the absenteeism. But, if significant changes are still observed in the ranks, then comparisons and analyses on student ratings render interpretations and inferences suspect. At that point, direction of the relationship between absenteeism and student ratings is also of importance. The position of the

present study is that, regardless of the reasons, absent students are not equipped to evaluate an instructor fairly, and, thus, their ratings should be given less weight.

The research questions of this study can be expressed as follows:

1. Do student ratings have invariant scale properties between absent and attending student groups?
2. Does the bias, if present, affect ranking of instructor?

Method

Sample

The data was drawn from a private, non-profit, English-medium Turkish university, located in the central region of the country. There are approximately 13000 students enrolled at the university. Number of students who filled out the student ratings forms was 31193 in 2098 classes. Response rate was 81.20%, which indicated a fair representation of all students in the university.

Considering that any student is quite likely to attend more than one evaluation sessions or a group of students in class may provide a common pattern of ratings, classes were taken as unit of analysis so to avoid problems associated with dependence. To this end, means of each item in the student ratings form were calculated for each class. Some classes (n=25) were excluded from the study since they had too few ratings defined by Rantanen (2013)[1]. The mean of missing ratings in the remaining varied between 0.40% and 2.86%. These classes were kept in the further analyses since such low missing ratings do not result in considerable decrease in reliability. Mean (standard deviation) of number of students in a class is 29.74 (SD=17.62). Mean number of forms filled out was 18.20 (SD=9.72) per class. Missing response rate per item was very low (less than 1% per item). Thus no procedure was applied to impute missing values.

A randomly selected sample of 2098 classes was drawn for the present study. Grade level distribution was as follows: 710 freshmen, 511 sophomores, 417 juniors and 435 seniors. Most of courses are must courses (n=1121). Fifty percent (n=1036) of the classes had one section. Twenty-one and 12.4 percent the classes had two and three sections, respectively (the rest with more than four credits).

Instrument

The Likert-type items in the student ratings form were developed by the administration of the university to represent different dimensions of instructional quality. All of the items in the student ratings form were included in the present study: (i) The instructor clearly states course objectives and expectations from students (*expectations*), (ii) The instructor stimulates interest in the subject (*interest*), (iii) The instructor stimulates in-class student participation effectively (*participation*), (iv) The instructor develops students' analytical, creative, critical, and independent thinking abilities (*thinking*), (v) The instructor interacts with students on a basis of mutual respect (*respect*), (vi) Rate the instructor's overall teaching effectiveness in this course (*overall*), (vii) I learned a lot in this course (*learned*), and (viii) the exams, assignments, and projects required analytical, scientific, critical, and creative thinking (assessment) (1: Strongly disagree, 2: Disagree, 3: Neutral, 4: Agree, and 5: Strongly agree).

Procedure

The student rating forms are given at the end of the each semester before the final exams. A 20-minute period in a class is given to student to fill the forms (including an open-ended section as well) in the absence of instructors. A volunteer student in each class takes responsibility for instructional evaluation, monitors the process and returns the forms to the administration of department. The forms are scanned by the computer centre of the university and instructors are allowed to see only the averages of the responses for each of their classes.

[1] In Rantanen's study, required number of ratings was defined as a function of class size. Please see Rantanen (2013) for further details.

Absenteeism

In the present study, students' self-rated attendance rates were used. Students reported this information by answering a yes/no question "*Did you attend all classes of this course?*" on the student rating form. Since the unit of analysis is classes, absenteeism was defined as the rate of students in a course who do not attend all classes during the semester. In the university from which data was drawn, attendance policy is applied. Students are not allowed to take the final exam if their absenteeism levels are above a level set by university administration.

Using a self-reported variable might be conceivably argued, since students might provide incorrect information due to fear of negative consequences or to have social desirability as a result of their absenteeism. The author of this study acknowledges that using official records of student absenteeism might be a better approach. However these data was not available. Use of a self-reported variable was justified with the support of the literature. As Johns and Miraglia (2015) reported, self-reported absenteeism can be used as a valid indicator in research studies. Number of studies regarding validity of self-reported absenteeism is very limited. On the other hand, validity of self-reported variables was shown in several studies. For example, Benton et al. (2013) stated that self-reported learning is strongly related to actual learning. In another study, Cassady (2001) showed self-reported grades are highly reliable indicators of actual grades. What is common in these studies is that authors had cautious that self-reported variables should not be used for absolute decisions. Given that students in the sample were informed that the absenteeism question was asked only for research purposes, self-absenteeism could be considered a valid variable for the present study. Furthermore full anonymity and confidentiality of student ratings might help eliminate this problem since students are guaranteed that that their reported will not be available to instructor or any other related parties.

The purpose of the present study is to examine the effect of absenteeism on scale properties of student ratings. To this end, classes were divided into two groups based on their absenteeism rates. First group (*attending)* included classes with 0% absenteeism rate (n=749), whereas the second group (*absent)* comprised of the remaining (n=1324). Mean absenteeism rate in the second group was 0.23 (SD=0.11) per class with a minimum 0.2 and maximum of 0.67, respectively. Average absenteeism rate was 0.36 in the range of 0 and 0.67 in the sample.

Preliminary Data Analysis

Means and standard deviations of each of eight items were given in Table 1. All means given by absent group were higher than attending one. Independent samples t-test was conducted to see if there was a significant mean difference on the ratings. Results revealed a significant overall mean difference (M_{absent}=4.50 and $M_{attending}$=4.29, t(1372.60)=8.64, p<.001). However, mean differences between these two groups can be not attributed to real differences unless measurement invariance is established.

Table 1. Means, Standard Deviations, Cronbach's Alphas for the Groups

Items	Absent		Attended		Whole	
	M	SD	M	SD	M	SD
expectations	4.61	0.45	4.43	0.54	4.54	0.49
interest	4.44	0.59	4.21	0.67	4.36	0.63
participation	4.47	0.57	4.18	0.69	4.36	0.63
thinking	4.41	0.56	4.20	0.61	4.33	0.59
respect	4.76	0.37	4.68	0.42	4.73	0.39
overall	4.53	0.53	4.34	0.62	4.46	0.57
learned	4.38	0.56	4.13	0.64	4.29	0.60
assessment	4.37	0.55	4.14	0.59	4.29	0.57

Cronbach's Alpha	0.90	0.89	0.93

Prior to the analyses, univariate and multivariate normality was assessed on both whole, absent and attending groups using the normality test in Lisrel. Results showed that data were not multivariate normally distributed in none of the groups (Whole group: χ^2=3402.60, p<.001; attending group: χ^2=16919.70, p<.001; absent group: χ^2=8772.56, p<.001). Similarly, skewness and kurtosis values indicated no univariate normality.

Since the data is not normal the principal index to assess goodness of fit was Satorra-Bentler $\chi 2$ (S-B$\chi 2$), which is the scaled version of $\chi 2$ index for non-normal conditions (Bryant and Satorra 2012). Assessment of goodness of fit was also supported by some other fit indices. These indices were Root Mean Square Error of Approximation (RMSEA), Standardized Root Mean Square Residual (S-RMR), Comparative Fit Index (CFI), and Non-Normed Fit Index (NNFI). Values which are generally accepted for a reasonable fit are as follows: below 0.08 for RMSEA and S-RMR and above 0.90 for CFI and NNFI (Hu & Bentler, 1999). Parameter estimates were made by using maximum likelihood estimation with asymptotic covariance and mean matrices due to its robustness against non-normality (Olsson, Troye, & Howell, 2000).

Before starting invariance analysis, omnibus test of equality of covariance structure was conducted in order to check the equality of observed indicator variance/covariance matrix across absent and attending groups (Vandenberg & Lance, 2000). Also unidimensionality of the 8 items were checked using confirmatory factor analysis.

Invariance Analysis

Invariance analysis was conducted using Multi-group Confirmatory Factor Analysis approach with the sequential constraint imposition (Byrne, 2004). In this approach, several constraints are hierarchically added to the baseline model at each step to assess equality of configural, weak, strong, and strict invariance. In the literature, invariance tests are suggested to be made in this particular order (Dimitrov, 2010; Schmitt & Kuljanin, 2008). After each constraint is added, goodness of fit of the new model is tested against the older one using a significance test. The first invariance level, configural invariance, is tested without imposing any constraints on the baseline model to check whether the proposed model is viable for all groups. For testing weak invariance, baseline model is further constrained by fixing factor loadings between different groups. One of the factor loading is fixed to 1 for estimating other parameters. When testing strong invariance, factor intercepts are also constrained as well as factor loadings. Similarly, one of the intercepts is fixed to zero for computational purposes. Strict invariance is tested by fixing error variances across groups. This level of invariance is investigated since it provides evidence for equality of error variance. To determine the reference item to be fixed, contribution of each item to the model was tested separately.

Assessment of invariance between two consecutive invariance levels was made based on the scaled chi-square difference test (ΔS-B$\chi 2$=ΔS-B$\chi 2_{constrained}$–ΔS-B$\chi 2_{unconstrained}$) (Satorra & Bentler, 1994) which was employed due to non-normal data set. ΔS-B$\chi 2$ test checks whether the new constraint results in a poorer goodness of fit. If a significant reduction is observed in goodness of fit after imposing a new constraint, it is concluded that invariance level that is defined with the new constraints does not hold. Non-significant values for the test S-B$\chi 2$ provide proof of invariance. All invariance analyses were conducted using mean and covariance matrices. If invariance is not observed at any level, then partial invariance was sought by freeing some fixed parameters in the model. Levine et al. (2003) suggested that a maximum 20% of the parameters should be freed. When relaxing constraints, Byrne's suggestion (2010) was followed and all factor loadings were tested separately to determine the non-invariant items.

Correction for Bias

To adjust student ratings against absenteeism, ratings given for the item (*I learned a lot in this course*) was selected and multiplied by extent of attendance (1-absenteeim rate) in a class, a common simple approach of weighting used to make ratings less biased (Wolbring 2012). As stated in the Introduction section of this paper, students with higher absenteeism were considered not to be equipped to judge instructional effectiveness due to low opportunity of observation during the semester. By using this approach, ratings were given less weight as compared to classes with high/full absenteeism. If no absenteeism is observed in a class (all student fully attending), then weighting would not change the ratings. On the other hand, ratings of other classes would be

lowered with respect to the degree of absenteeism. After this adjustment instructors were ranked and their changes in the ranking were compared with the previous values.

Results

Omnibus test of equality of covariance matrices showed that two groups are not equivalent in terms of covariance structure, Box's M=231.835, F(36, 839.04)=504.318, p<.001, which provided support for further invariance analysis. Before conducting invariance analysis, a baseline model was hypothesized to 8 items grouped under a common latent construct. A confirmatory factor analysis showed unidimensionality of the data for whole, absent and attending groups. Results of the goodness-of-fit indices are as follows:

1. Whole group: S-Bχ2(20)= 308.94, p<.001, RMSEA=.08 [90%CI=.07;0.09], S-RMR=.02, CFI=.98, NNFI=.97.
2. Absent group: S-Bχ2(20)=219.012 p<.001, RMSEA= .08 [90%CI=.07;0.09], S-RMR=.02, CFI=.97, NNFI=.96.
3. Attending group: S-Bχ2(20)=110.956, p<.001, RMSEA=.07 [90%CI=.06;0.09], S-RMR=.02, CFI=.98, NNFI=.97

Due to the dependency on sample size S-Bχ2 tests were significant but all other fit indices provided good values. Also all of the factor loadings were above .60 (p<.05) and reliabilities are .70 (Nunnally, 1978), as shown in Table 2, which provided supporting evidence for unidimensionality. Thus the 8-item model constituted a single-factor baseline for invariance analyses between whole, absent and attending groups. Factors loadings were similar across two subgroups. In other words, the relationships between observed items and the latent variable that are assessed by 8 items were similar across groups.

Table 2. Standardized (Unstandardized) Factor Loadings for the Baseline Model in Whole, Absent and Attendant Groups

Items	Whole	Absent	Attendant
expectations	.93 (1.00)	.91 (1.00)	.94 (1.00)
interest	.96 (1.33)	.96 (1.38)	.96 (1.27)
participation	.93 (1.28)	.92 (1.27)	.93 (1.26)
thinking	.95 (1.23)	.94 (1.28)	.96 (1.16)
respect	.71 (0.62)	.72 (0.66)	.70 (0.58)
overall	.91 (1.14)	.92 (1.28)	.89 (1.18)
learned	.92 (1.23)	.90 (1.24)	.93 (1.18)
exams	.89 (1.12)	.87 (1.16)	.90 (1.06)

As shown in the Table 3, the baseline model was tested without any constraints across two groups to examine the configural invariance. The item *expectations* were used as reference item for fixing item loading (and intercept) since it was the most invariant (based on mean socres) item between two groups of classes. S-Bχ^2 test was significant as expected, but all the other fit indices indicated that configural invariance held, suggesting equivalence of the baseline model with no insignificant parameters. Then weak invariance tests by imposing a constraint to fix all factor loadings across the groups. The model with fixed factor loadings had a significant result. Freeing the item *participation* resulted in a non-significant change between weak and configural invariance models. Full or partial strong invariance was not observed indicating the lack of invariant item intercepts between two groups. No further attempt was made to investigate strict invariance since previous invariance level did not hold.

Table 3. Results of Invariance Analysis

	S-Bχ2	df	Model comparison	ΔS-Bχ2	Δdf	RMSEA	S-RMR	CFI	NNFI
1.Configural	350.46*	48	-	-	-	.07 [.06;.09]	.03	.97	.96
2.Weak	384.96*	55	2-1	34.50*	7	.07 [.06;.08]	.03	.97	.97
3.Weak-P[1]	377.35*	54	3-1	26.89	6	.07 [.07;.08]	.04	.97	.97
4.Strong	441.31*	63	4-3	63.96*	9	.07 [.07;.09]	.03	.97	.97

[1] The item *participation* was freed. * p<.001.

Since the item intercepts were found to be different, their differences were calculated. Table 4 presents item intercepts for absent and attending groups. Since the item *participation* was freed when weak invariance was investigated, its intercepts were not calculated. All intercepts are smaller in absent group, except the item *participation*. Lower intercepts in absent group mean that this group had higher factor scores (instructional effectiveness) for the same rating. Or the same instructional quality is rated by absent students with lower scores. The non-zero differences indicated a bias on ratings between two groups.

Table 4. Differences between Item Intercepts

Items	Difference*
expectations	0.00
interest	0.51
thinking	0.55
respect	0.35
overall	0.02
learned	0.24
exams	0.42

* $\text{Intercept}_{\text{attended}}$-$\text{Intercept}_{\text{absent}}$.

Having demonstrated a potential bias on student ratings due to absenteeism, a weighting approach was used to investigate the effect of the bias on raking of instructors. Figure 1 shows the histogram of percent changes in the ranks after the adjustment. The rate of instructors whose percentiles were not changed was 12.4% (n=257). Median of the percentage percentile changes after adjustment was .20%.

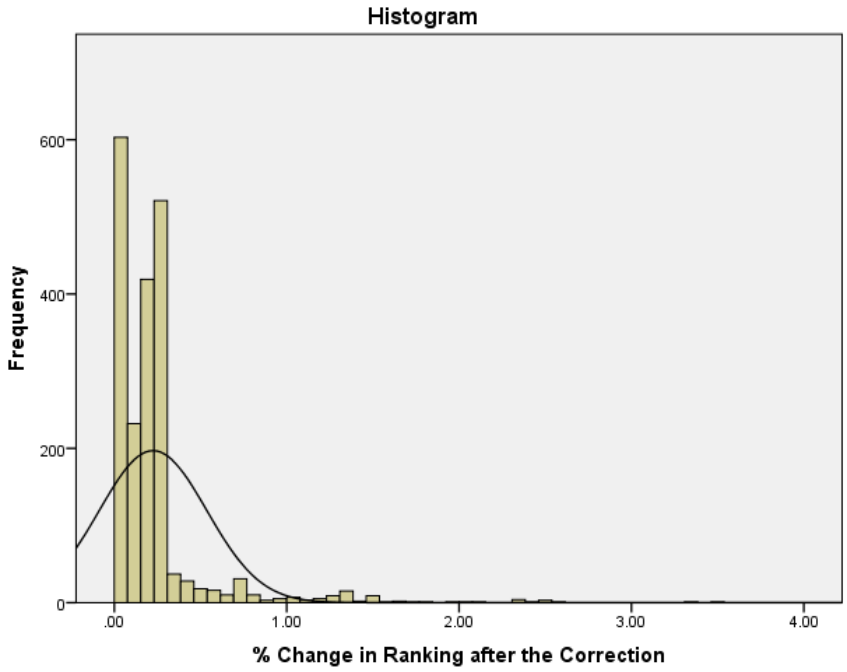

Figure 1. Distribution of Ranking Differences after the Adjustment (%)

Rank of the twenty-five percent of instructors changed by a maximum of 0.05%, while the mean change was 0.20% for 50% of the instructors. Table 5 shows frequencies of instructors who stayed in their original percentiles after the adjustment. As can be seen in the table, only half of the top 1st instructors with the highest student ratings stayed in the same percentile after the adjustment. Similarly, instructors who were at the 5th and 10th percentiles kept their original positions.

Table 5. Percent of Instructors who stayed at their Original Percentiles after the Adjustment

Original Percentile	Percent of Instructors
1st	50.0
5th	63.5
10th	73.0
20th	18.8
30th	2.5
40th	4.7
50th	8.3
60th	18.5
70th	19.3
80th	29.6
90th	71.0

Discussion

Results showed configural and partial weak invariance were observed between the two groups. Equality of factorial structure (configural invariance) indicated that both absenteeism groups have the same conceptualization for instructional performance. Also ratings have a common unit across attending and absent classes, as evidenced by invariant factor loadings except the item *participation* (weak invariance). The same amount of change in instructional performance results in the same amount change in student ratings in both groups of classes. This allows within-class comparisons of instructional performance. Thus instructors can use student ratings for feedback for their class without making comparisons among classes. As an exception, classes with absent students seemed to have a different level of importance for instructors' ability to create in-class participation than attending students do. However, the failure to observe invariant factor intercepts indicated some source of variance in student ratings that is irrelevant to instructional performance due to absenteeism of students. In other words, means given in Table 1 do not indicate students' true ratings, rather they are under the effect of bias.

The non-invariant factor intercepts between regularly attending students and those who are not pose a serious threat given students ratings are used in high-stake decisions. Between-class comparisons are not allowed due to the lack of a common point of origin. Closer examination of the intercepts revealed, unlike the instructors of classes which included regularly attending students, the instructors who teach classes with absent students seem to receive lower ratings for the same instructional performance, regardless of the reason for voluntarily or involuntarily absenteeism. On the other hand, instructors receive higher ratings from attending students' classes.

Lower ratings from absent classes can be explained from different perspectives. First of all, it should be noted that this study did not discriminate between reasons of absenteeism. Still, it is possible to speculate on the relationship between absenteeism and ratings provided by students. Both of grading leniency and validity hypotheses seem to be plausible in explaining the effect of absenteeism on invariance of student ratings.

Absent students may be giving lower ratings since they receive/expect higher grades, if their reason for absenteeism is voluntarily or involuntarily. This explanation supports the grading leniency hypothesis, which states that instructors can either "buy" high ratings from students by giving them high grades (Langbein, 2008; McPherson & Jewell, 2007) or students with low expected/received grades may discredit instructors with giving them lower ratings, although these students are responsible for their low performance (Greenwald & Gillmore, 1997).

Another explanation for lower ratings for the same instructional quality by classes with absent students can also be the validity hypothesis, which states that high ratings are a result of effective instruction (Marsh & Roche, 2000; Spooren et al., 2013). In other words, when an instructor delivers a good instruction, she or he can expect higher ratings. If absent students start missing classes due to instructional performance, they could be expected to give lower ratings due to their negative opinions about instructional performance in those classes. On the other hand, classes with regularly attending students expect more from instructors. For such classes, seeing quality instructional material, assessment items, clear organization, presentation skills, etc. are closely associated with high instructional performance, and in turn, high ratings. This explanation is in agreement with what was reported by Lopez-Bonilla and Lopez-Bonilla (2015). They found that teaching style is the one of the strongest predictor of absenteeism. On the other hand, Moore, Armstrong, and Pearson (2008) stated that absenteeism may be proximity for student conscientiousness and diligence. Such motivated students are expected to attend classes more frequently. However, Durden and Ellis (2003) stated that motivation is separate factor from absenteeism. Similarly, Romer (1993) and Stanca (2004) investigate the relationship between attendance and attainment and found a significant relationship even after controlling for student motivation, ability, effort and motivation.

Whatever the reason is, instructors of classes with absent students receive systematically lower ratings than classes with regularly attending students. In this respect, this study reported supporting findings to the literature. Martin et al. (2013) found that absenteeism is associated with low student ratings, reporting that absent students gave significantly lower ratings than regular students. Similar findings were also reported by Berger and Schleußner (2003) and Babad, Icekson, and Yelinek (2008). These studies implicitly assumed that student ratings were invariant scale properties.

Students who do not value believe in the importance of attending in the courses in order to be successful become absent more frequently. These students may be informed about the evaluation process and motivated to be actively participate into rating sessions (Stalmeijer et al. 2016). Although it might be beneficial to help student

understand importance of attendance in achievement, this might not be enough. Instructors should also try to create classroom environments that can attract students. Moore et al. (2008) stated that increasingly though, the perceived value, impact and quality of the lecture experience, however that lecture is delivered, are likely to be central and important issues that influence the decision by students to attend and engage. However, for involuntarily absenteeism, institutional mechanisms could be employed such as remedial courses, etc. or students may be encouraged to take class another semester.

One might argue that items showing bias should be removed from the list of items to eliminate the effect of absenteeism. However in this study, all items were found to be biased. Thus correction methods should be considered. After statistical correction, only half of the instructors in the top 1st percentile kept their positions in the same percentile. However remaining instructors showed significantly larger deviations from their original ranking. When the top 10th percentile was considered, most of the instructors stayed in the same percentile. In general, instructors in the middle percentiles were affected mostly by adjustment. These findings imply that high-stake decisions about the instructors at the top percentile may be erroneous.

Limited comparison opportunity between instructors who teach students who fully attend classes and those who do not was also in parallel with findings reported by Berger and Schleußner (2003), Babad, Icekson and Yelinek (2008), and Wolbring (2012). Collecting student ratings at earlier stages of the courses may be helpful in lessening the bias due to learning levels. There are some older studies reporting relationships between the ratings collected at the mid-semester and end of the semester (Costin, 1968; Kohlan, 1973). Or implementing student ratings more than once in a semester may be a solution.

Given the relationship between achievement and absenteeism, one might also argue that strict attendance policies by higher education institutions may be considered instead of statistical control of absenteeism. However, results reported in the literature as to relationship between attendance and achievement showed that these policies are not able to eliminate the effect of absenteeism (Ladwig & Luke, 2013). Thus it seems inevitable to control for the influence of absenteeism on student ratings.

It is acknowledged that this study has some limitations. Use of binary split for absenteeism is a problem. However, data availability limited the researcher. Moreoever, use of self-reported absenteeism might not provide a complete picture of absenteeism. In this study, students were informed about anonymity of student ratings, but they might still have feared negative consequences. a study to include (i) the relationship between achievement and absenteeism and (ii) reasons of absenteeism might help depict a more complete picture.

In summary, absenteeism creates non-invariant scale properties for student ratings. The results indicated that naively assuming that student ratings are comparable among group may have serious implications and comparison of instructors who teach classes with varying levels of absenteeism may not be fair. Thus student ratings should not be considered sole indicator of instructional performance due to its absenteeism-based bias. Ratings may not reflect true performance of instructors. Use of student ratings should be supported by other means of assessing instructional performance such as committees of experts, peer evaluation, self-evaluation, students' grades, etc.

References

Babad, E., Icekson, T., & Yelinek, Y. (2008). Antecedents and correlates of course cancellation in a university 'drop and add' period. *Research in Higher Education*, *49*, 293–319.

Barlow, J., & Fleischer, S. (2011). Student absenteeism: Whose responsibility? *Innovations in Education and Teaching International*, *48*, 227–237.

Benton, S. L., Duchon, D., & Pallett, W. H. (2013). Validity of student self-reported ratings of learning. *Assessment & Evaluation in Higher Education*, *38*(4), 377-388.

Berger, U., & Schleußner, C. (2003). Are ratings of lectures confounded with students' frequency of participation? *German Journal of Educational Psychology*, *17*, 125–31.

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

Byrne, B. M. (2004). Testing for multigroup invariance using AMOS graphics: A road less traveled. *Structural Equation Modeling*, *11*(2), 272–300.

Cassady, Jerrell C. (2001). Self-reported gpa and sat: a methodological note. *Practical Assessment, Research & Evaluation*, *7*(12). Retrieved July 19, 2016 from http://PAREonline.net/getvn.asp?v=7&n=12.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*, 233-255.

Colby, J. (2004). *Attendance and Attainment*. Paper presented at the 5th Annual Conference of the Information an d Computer Sciences – Learning and Teaching Support Network (ICS-LTSN), University of Ulster.

Costin, F. (1968). A graduate course in the teaching of psychology: Description and evaluation. *Journal of Teacher Education*, *19*, 425–32.

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.

Durden, G. C., & Ellis, L. V. (2003). Is class attendance a proxy variable for student motivation in economics classes? An Empirical Analysis. *International Social Science Review*, *78*, 22–34.

Ehie, I. C., & Karathanos, D. (1994). Business faculty performance evaluation based on the new aacsb accreditation standards. *Journal of Education for Business*, *69*(5), 257-262.

Friedman, P., Rodriguez, F., & McComb, J. (2001). Why students do and do not attend class. *College Teaching*, *49*, 124–133.

Greenwald, A., & Gillmore, G. (1997) Grading leniency is a removable contaminant of student ratings. *American Psychologist*, *52*(11), 1209-1217.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, *18*, 117–144.

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

Hughes, S. J. (2005). Student attendance during college-based lectures: a pilot study. *Nursing Standard*, *19*(47), 41–49.

Hunter, S., & Tetley, J. (1999). *Lectures. Why don't students attend? Why do students attend?* Proceedings of HERDSA Annual International Conference held in Melbourne 12–15 July 1999, Higher Education Research and Development Society of Australia, Milperra, NSW.

Johns, G., & Miraglia, M. (2015). The reliability, validity, and accuracy of self-reported absenteeism from work: a meta-analysis. *Journal of Occupational Health Psychology*, *20*(1), 1-14.

Kohlan, R. G. (1973). A comparison of faculty evaluations early and late in the course. *Journal of Higher Education, 44*, 587–95.

Kousalya, P., V., Ravindranath, K., & VizayaKumar. (2006). Student absenteeism in engineering colleges: Evaluation of alternatives using ahp". *Journal of Applied Mathematics and Decision Sciences*, 1-26.

Ladwig, J., & Luke, A. (2013). Does improved attendance lead to improved achievement? An empirical study of Indigenous education in Australia. *Australian Educational Researcher*, *41*(2), 171-194.

Langbein, L. (2008). Management by results: Student evaluation of faculty teaching and the mismeasurement of performance. *Economics of Education Review*, *27*(4), 417–428.

Launius, M. H. (1997). College student attendance: attitudes and academic performance.*College Student Journal*, *31*, 86–92.

Levine D. W., Kaplan, R. M., Kripke, D. F., Bowen, D. J., Naughton M. J., & Shumaker, S. A. (2003). Factor structure and measurement invariance of the women's health initiative insomnia rating scale. *Psychological Assessment*, *15*(2), 123-136.

López-Bonilla, J. M., & López-Bonilla, L. M. (2015). The multidimensional structure of university absenteeism: An exploratory study. *Innovations in Education and Teaching International*, 52(2), 185-195.

Macfadyen, L. P., Dawson, S., Prest, S., & Gašević, D. (2015). Whose feedback? A multilevel analysis of student completion of end-of-term teaching evaluations. *Assessment & Evaluation in Higher Education*, *41*(6), 821-839.

Marsh, H. W., & Roche, L. A. (2000). Effects of grading leniency and low workload on students' evaluations of teaching: popular myth, bias, validity, or innocent bystanders. *Journal of Educational Psychology*, *92*(1), 202–228.

Martin S. I., Way, D. P., Verbeck, N., Nagel, R., Davis, J. A., & Vandre D. D. (2013). The impact of lecture attendance and other variables on how medical students evaluate faculty in a preclinical program. *Academic Medicine*, *88*(7), 972-977.

Massingham, P. & Herrington, T. (2006). Does attendance matter? An examination of student attitudes, participation, performance and attendance. *Journal of University Teaching and Learning Practice*, *3* (2), 82-103.

Marburger, D. (2001) Absenteeism and undergraduate exam performance. *Journal of Economic Education, 32*(2), 99-109.

Martinez, P. (2001). *Improving student retention and achievement. What do we know and what do we need to find out?* London: Learning and Skills Development Agency.

McPherson, M. A., & Jewell, R. T. (2007). Leveling the playing field: Should student evaluation scores be adjusted? *Social Science Quarterly*, *88*(3): 868–881.

Moore, S., Armstrong, C., & Pearson, J. (2008). Lecture absenteeism among students in higher education: a valuable route to understanding student motivation. *Journal of Higher Education Policy and Management*, *30*(1), 15-24.

Newman-Ford, L., Fitzgibbon, K., Lloyd, S., & Thomas, S. (2008). A large-scale investigation into the relationship between attendance and attainment: A study using an innovative, electronic attendance monitoring system. *Studies in Higher Education*, 3(6), 699– 717.

Nicholl, H. & Timmins, F. (2005). Programme-related stressors among part-time undergraduate nursing students. *Journal of Advanced Nursing*, *50*(1), 93–100.

Nunnally, J. C. (1978). *Psychometric theory (2nd ed)*. New York: McGraw-Hill.

Olsson, U. H., Foss, T., Troye, S. V., & Howell, R. D. (2000). The performance of ML, GLS, and WLS estimation in structural equation modeling under conditions of misspecification and nonnormality. *Structural Equation Modeling*, *7*(4), 557–595.

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.

Remmers, H. H. (1928). The relationship between students' marks and student attitudes towards instructors. *School and Society*, *28*, 759–760.

Rodríguez, R., Hernández, J., Alonso, A., & Diez-Itza, E. (2003). El absentismo en la Universidad: Resultados de una encuesta sobre motivos que señalan los estudiantes para no asistir a clase [Absenteeism in university: The results of a survey on students' reasons for non-attendance]. *Aula Abierta, 82*,117–145.

Romer, D. (1993). Do students go to class? Should they? *Journal of Economic Perspectives*, *7*, 167–174.

Satorra, A., & Bentler, P. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. Von Eye & C. C. Clogg (Eds.), *Latent variable analysis. Applications for developmental research* (pp. 399–419). Thousand Oaks, CA: Sage.

Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human Resource Management Review*, *18*, 210-222.

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*, 598–642.

Stalmeijer, R., Whittingham, J., de Grave, W., & Dolmans, D. (2016). Strengthening internal quality assurance processes: Facilitating student evaluation committees to contribute. *Assessment & Evaluation in Higher Education*, *41*(1), 53–66.

Stanca, L. (2004). The effects of attendance on academic performance: panel data evidence for introductory microeconomics. *The Journal of Economic Education*, *37*(3), 251-266.

Triadó-Ivern, X., Aparicio-Chueca, P., Guàrdia-Olmos, J., Peró-Cebollero, M., & Jaría- Chacón, N. (2013). Empirical approach to the analysis of university student absenteeism: proposal of a questionnaire for students to evaluate the possible causes, *Quality & Quantity*, *47*(4), 2281-2288.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4-69.

Wolbring, T. (2012). Class attendance and students' evaluations of teaching: Do no-shows bias course ratings and rankings? *Evaluation Review, 36*(1), 72–96.

Zabaleta, F. (2007). The use and misuse of student evaluations of teaching. *Teaching in Higher Education*, *12*(1), 55–76.