ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2017; 8(1),15-33

Exploring Variability Sources in Student Evaluation of Teaching via Many-Facet Rasch Model*

Ders Değerlendirme Anketinde Varyans Kaynaklarının Çok Yüzeyli Rasch Modeliyle Değerlendirilmesi

Bengü BÖRKAN **

Abstract
Evaluating quality of teaching is important in nearly every higher education institute. The most common way of assessing teaching effectiveness takes place through students. Student Evaluation of Teaching (SET) is used to gather information about students' experiences with a course and instructor's performance at some point of semester. SET can be considered as a type of rater mediated performance assessment where students are the raters and instructors are the examinees. When performance assessment becomes a rater mediated assessment process, extra measures need to be taken into consideration in order to create more reliable and fair assessment practices. The study has two main purposes; (a) to examine the extent to which the facets (instructor, student, and rating items) contribute to instructors' score variance and (b) to examine the students' judging behavior in order to detect any potential source of bias in student evaluation of teaching by using the Many-Facet Rasch Model. The data set includes one thousand 235 students' responses from 254 courses. The results show that a) students greatly differ in the severity while rating instructors, b) students were fairly consistent in their ratings, c) students as a group and individual level are tend to display halo effect in their ratings, d) students are clustered at the highest two categories of the scale and e) the variation in item measures is fairly low. The findings have practical implications for the SET practices by improving the psychometric quality of measurement.

Keywords: Student evaluation of teaching, Many Facet Rasch Model, psychometric analysis

Öz
Öğretim niteliğini değerlendirmek neredeyse her yükseköğretim kurumlarında önemlidir. Öğretimin etkinliğini değerlendirmenin en yaygın yöntemi öğrenciler üzerinden gerçekleştirilmektedir. Ders değerlendirme anketi (DDA) yoluyla dönemin herhangi bir zamanında öğrencilerin ders ve öğretim elemanı hakkındaki görüş ve tecrübelerine ilişkin bilgi toplanır. DDA, puanlayıcı aracılı bir tür performans değerlendirmesi olarak kabul edilebilir. Bu kez öğrenciler puanlayıcı, öğretim elemanları ise değerlendirilendir. Performans değerlendirme, bir puanlayıcı aracılı değerlendirme süreci olduğunda, ekstra önlemler, daha güvenilir ve adil bir değerlendirme uygulamaları oluşturmak için dikkate alınması gerekir. Bu çalışmanın amacı, a) öğretim elemanlarının puanlarındaki farklılığa/varyansa, puanlama sürecindeki yüzeylerin (öğretim elemanı, öğrenci ve değerlendirme maddeleri) ne derece katkı sağladığını ve b) öğrencilerin yaptiklari puanlamalarda yanlılığa yol açacak potansiyel kaynakları çok yüzeyli Rasch modeli yardımıyla incelemektir. Çalışmada kullanılan veri seti 254 dersten 1.235 öğrencinin değerlendirmelerini kapsamaktadır. Sonuçlara göre a) öğrenciler, öğretim elemanlarını değerlendirirken farklı katılık/cömertlik derecesi göstermektedirler, b) çoğu öğrenci kendi içinde oldukça tutarlı, c) grup olarak, değerlendirmelerde halo etkisi oldugu görülmektedir, d) öğrenciler beşli ölçeğin üst puanlarında kümelenmişlerdir, e) madde zorluk değerlenlerindeki varyasyon çok düşüktür. Bu bulguların, DDA'nin psikometrik özelliklerinin daha nitelikli hala getirilmesi yönünde sonuçları vardır ve bunlar makalede tartışılmıştır.

Anahtar Kelime: Ders değerlendirme anketi, Çok Yüzeyli Rasch Modeli, psikometrik analiz

* Preliminary results of this study was presented at the European Conference on Educational Research 2015
** Yard. Doç. Dr., Boğaziçi University, Istanbul-Turkey, e-posta: bengu.borkan@boun.edu.tr

DOI: 10.21031/epod.298462
Geliş Tarihi: 29.09.2016
Kabul Tarihi: 28.02.2017

INTRODUCTION

Student Evaluation of Teaching (SET) which is used to gather information regarding students' experiences with a course and instructor's performance at some point of semester seems to be to the most common ways of gathering data for supposedly both formative and summative evaluation purposes (Gravestock, & Gregor-Greenleaf, 2008; Penny, 2003; Seldin, 1993; Zabaleta, 2007). By means of SET results, administrators in higher education institutes aim to (a) improve teaching quality, (b) provide input for appraisal exercises (e.g., tenure/promotion decisions), and (c) provide evidence for institutional accountability (Seldin, 1993; Spooren, Brock & Mortelmans, 2013).

SET can be considered a type of performance assessment. In performance assessment, a person (examinee) displays performance and/or construct a product and, quality of this performance or product is evaluated by at least one evaluator/rater. When performance assessment becomes a rater mediated assessment process, extra measures need to be taken into consideration in order to create more reliable and fair assessment practices. One of the most common threat in rater mediated assessment is called 'rater variability'. This term generally describes the variability that is linked to rater characteristics (i.e. lenient, severe, gender), not to the performance of person being evaluated (Eckes, 2009). In other words, rater variability threatens the validity and fairness of performance assessment when measurement is involved construct irrelevant variance in examinee scores (Lane & Stone, 2006; Messick, 1998). Eckes states that "This long, and possibly fragile, interpretation–evaluation–scoring chain highlights the need for carefully investigation of the psychometric quality of rater-mediated assessments. One of the major difficulties facing the researcher, and the practitioner alike, is the occurrence of rater variability." (p.4).

As literature point out, both theoretical and psychometric issues remain unresolved for SET questionnaires (Gravestock & Gregor-Greenleaf, 2008). Studies have been accumulated around two main concerns which are in fact relevant to each other. The first concerns focus on the question whether the score obtained from students' evaluations are valid and actually measure what we intent to measure so called teaching effectiveness. The second concern focus on potential bias sources which treats the reliability and validity of our measures (Gursoy & Umbreit 2005) such as gender of the instructor, expected grade. The purpose of this study is to (a) examine the extent to which the facets (instructor, student, and rating items) contribute to instructors' score variance and (b) examine the students' judging behavior using the Many-Facet Rasch Model in order to detect any potential source of bias in student evaluation of teaching.

Evaluating Quality of Teaching in Higher Education

Because of the great extent use of SET, an enormous literature has been collected since early 1920 in which the first SET was administered at the University of Washington (Seldin, 1993; Zabaleta, 2007). Since then, some issues such as validity of SET has remained, and other issues like the use of SET results to improve teaching, has recently come to researchers' attention. Majority of the research has been conducted in North American, Australian and UK teaching context (Gravestock, & Gregor-Greenleaf, 2008; Zabaleta, 2007). Majority of those studies have generally positive position for the use of SET (such as Abrami, 2001; Beran, Violato, & Kline, 2007; Gravestock, & Gregor-Greenleaf, 2008; Marsh, 1987); on the other hand, some have displayed a skeptical attitude toward the use of SET since SET could produce bias results due to teacher and course characteristics which are believed to be irrelevant with the quality of teaching (Dodeen, 2013; Koh & Tan, 1997; Williams & Ceci, 1997). Literature displays ambivalent research findings regarding the validity of this method and the use of its results. While some studies claim that SET provides valid data in general (e.g. Marsh, 1984, Nelson & Lynch, 1984, Zangenehzadeh, 1988) and no bias in particular, other studies reported bias in the data and concluded deficiency in validity of SET scores (e.g. Centra, 1993, Haladayna & Hess, 1994, Marsh, 1987, Marsh & Roche, 2000). Therefore, it is suggested that the results of SET should not be used alone for high stake decision such as retention, promotion or tenure (Penny, 2003).

Economic and political changes in the World have been pushing higher education institutes to exhibit their performances equally well in both research and teaching arena. Moore and Kuol (2005) argued that SET provides us quantitative data that we can use for comparison is imprecise ways of evaluating and comparing teaching effectiveness. Therefore, educational institutions should be well informed about how to present, interpret and use these sorts of data (Gravestock, & Gregor-Greenleaf, 2008). Although it is widely accepted that SET should not be the only toll to evaluate ones teaching quality, SET result will continue to be used for longer time as performance indicator of teaching effectiveness (Penny, 2003).

We have big assumption based on the idea that data obtained SET questionnaire is a good measure of teacher effectiveness which leads to students' learning and better education. Well informed decisions should be based on a vigorous SET system including valid and reliable data collection instrument and dependable data collection procedure. Many critical studies highlighted the weakness of student evaluations of teaching and questioned its validity. Instructor's sex or personality could be a determining factor in students' rating. Being a female instructor can be disadvantageous (Basow & Martin, 2012) or perceived attractiveness/expressiveness of the instructor is a shaping factor in students' judgment on the quality of teaching; (Cashin, 1995). Marsh (1982) reported that SET appears to be subject to substantial halo effects, which means students answers the different dimension of the instrument in a similar way. Barnes and colleagues (2008) and Wachtel (1998) reported students' ratings on different dimension were significantly correlated with students' expected course grade. Students tend to rate the instructor or instruction more highly in smaller courses (Hoyt, Che, Pallett, Gross, 1999) and age is negatively correlated with evaluation scores (Zabaleta, 2007). Grading leniency/severity can bias student ratings (Basow & Martin, 2002).

Student Evaluations of Teaching Questionnaire

Effective teaching is considered as a construct that we use to explain desired instructor/teachers behavior in educational process. None of the construct can be directly observed and measured and, therefore we need an operational definition for them. In other words, we need a list of related behaviors that are associated with our construct. Unfortunately, no consensus on the definition of what effective teachings is in this sense. Ory and Ryan (2001) argue that there is no "universal set of characteristics of effective teachers and courses that should be used as a target…"(p.32). Therefore, various measurement tools are available; almost every institution developed their own questionnaire by considering institutional needs, climates and priorities. Keeley (2012) called this questionnaires "home grown", I called "tailor made".

Existing questionnaires have different content/items in various lengths. SET consists of a questionnaire which usually includes mixture of open-ended (qualitative data) and closed-ended items (quantitative data) with a rating scale. Items are usually related with different dimensions of teacher effectiveness such as planning, organization, grading, interaction, instruction, learning, fairness of grading. By means of including all different dimension of teaching better content and construct validity could be achieved. However, this makes the questionnaire longer.

A number of researchers conducted a study for the purpose of identifying the dimensions, sub-construct or factors of the construct that is usually named as students' perceptions of teaching effectiveness. Here some examples for a well-developed, psychometrically evaluated instrument. Barnes and colleagues (2008) developed a questionnaire with 14 items. They identified two distinct factor; teaching excellence and teaching readiness. In another study, Mortelmans and Spooren (2009) developed a questionnaire including 37 items and 12 dimensions of effective teaching; build-up of subject matter, Linking-up with foreknowledge\ content validity of examination, presentation skills, value of subject matter, course difficulty, harmony organization course learning\ course materials, clarity of course objectives, help of teacher during learning process, formative evaluation and authenticity of the examination. Marks (2000) explore five underlying constructs for his student evaluation questionnaire: organization, expected/fairness of grading, workload/difficulty, perceived and instructor liking/concern, learning. Marsh (1982) reported nine dimensions of teaching

effectiveness with 35 items; learning/value, organization, enthusiasm, breadth of coverage, group interaction, individual rapport, examinations and grading, assignments, and workload. The questionnaire has very high internal consistency coefficient and it produced stable results over time. Ginns, Prosser and Barrie (2007) developed their own 23 item five factor SET questionnaire. They named factors as generic skills scale, appropriate assessment, good teaching, appropriate workload and clear goals and standards. They report high internal consistency and inter-rater agreement.

In some higher education institutions SET may play an important role and it effects teaching climate of institution in particular. Negative attitude of instructor towards SET were mention in the literature (e.g. Spooren et al, 2013). Given that instructors are the primary users of this system, their trust is very curial to fully utilize the potential of SET. Primary reason of instructor not to trust is the belief of potential bias in student rating. Yet, negative findings regarding the validity mentioned above elevate their concern. Consequently, SET needs to come under the spotlight in order to develop instructor trust and to increase the practical usefulness of SET.

Rater-Mediated Performance Assessment

The score of examinee on a performance task depends on not only examinee ability but also other various facets related to the nature of assessment. Three most commonly seen aspects (facets) are the ability of the examinee and the difficulty of the performance task (Mulqueen , Baker & Dismukes, 2002) and the rater effect. Rater effects in an evaluation process appears in different forms such as halo effect, rater severity/leniency or central tendency (Hoyt, 2000; Myford & Wolfe, 2003). Such rater effects introduce a method variance for scores. Previous research in different settings show that significant rater effects exist in rater mediated performance assessment (Eckes, 2005). For example, The meta-analysis study shows that 37% of examinee performance can be explained with rater effect and rater-ratee interaction (Hoyt & Kerns, 1999). Different procedures can be used to control reliability of scoring in evaluation processed where there are multiple evaluators.

A Many-Facet Rasch Model approach

Like G Theory, a Many-Facet Rasch Model (MFRM) approach allows researchers or practitioners to analyses potential sources of errors in rating processes. A MFRM developed by Linacre (1989) is based on the basic Rasch model (Rasch, 1960/1980). The Rasch model, a one-parameter latent trait model, provides item free estimates of each person ability and person free estimates of item difficulty and places both estimates on an equal-interval log-linear scale (Wright & Stone, 1979). In other words, estimates of person measures are independent of the difficulty of item or task in measurement processes, and estimates of item or tasks measures are independent of the specific group of people ability (Sudweeks, et al, 2005).

Since a MFRM is a member of the Rasch family, it possesses all of the characteristics of a basic Rasch model and more. MFRM allows assessment of various variability sources in the rating score, for instance examinee ability, task difficulty, rater severity and interaction of these facets. In short, a MFRM has the following benefits;

a) If the data fits the model, each facet are estimated independently (Linacre & Wright, 2002); in other words, the measures obtained by the model are sample, item, and condition-free. Therefore, function of facet can be evaluated separately (Myford & Wolfe, 2003).
b) Since person estimates, item estimates and all other facets' estimates are located on the same logit scale, comparisons between facets are possible.
c) Individual level effect (besides group level effect given in previous Bullet b) within in each facet are examined more closely; for instance, which raters rate more severely or which raters disagree other raters.
d) The MFRM provides us goodness-of-fit statistics showing degree of data fit to the model, and they help us to interpret the fit of each single element in each facet (Sudweeks et al, 2005).

e) The MFRM provides bias analysis, that is, the analysis of the interactions between elements of different facets (see Linacre, 2009a, for details). For instance, researchers can examine raters' severity depends upon the characteristics of ratee or the condition of the ratings (Myford & Wolfe, 2003).

Many Facet Rasch Model

Many facet Rasch model extends the Rasch model into more complex situation including more than two facets (i.e. examinee and item) of interest of interest (Linacre, 1989). This model is particularly useful for analysis of subjectively rated performance assessment and/or various tasks in different difficulty level:

$$\ln\left(\frac{P_{nij}}{P_{ni(k-1)}}\right) = \beta_n - \delta_i - \gamma_j - \tau_k \qquad \text{(MFRM, 6)}$$

where

P_{njimk}, = probability of person *n* receiving a rating of *k* on criterion *i* from rater *j*,

$P_{njim(k-1)}$ = probability of person *n* receiving a rating of *k*-1 on criterion i from rater *j*,

β_n = ability of person *n*,

δ_i = difficulty of criterion *i*,

γ_j = severity of rater *j*,

τ_k = difficulty of receiving a rating of *k* relative to a rating of *k*-1.

MFRM is an additive model and can be expanded to as many facets as we like. Besides persons and item facets, other facets that are susceptible to contributing construct irrelevant variances in measurement, such as raters, occasion, and task facets, can be added to the model. As in Rasch model, this model calibrates each facet on a common logit linear scale after raw scores are corrected for inconsistencies among raters' severity, differences in the relative task difficulty (Lunz, Wright & Linacre, 1990). Along the logic scale, the higher the number is, the more lenient the rater is; the more negative the number is, the more severe the rater is. Moreover, MFRM allows us to detect unusual interactions called as bias between raters and tasks/items, or raters and particular examinee (Linacre, 1994).

METHODOLOGY

Data Source

This study will utilize student evaluation of teaching data collected in the undergraduate courses of a mid-size university in a big city. This public university is located on the north western part of Turkey and serving around 11 thousand students in 32 undergraduate programs and four thousand graduate students in 56 master and 32 doctoral programs.

The university has a 150 year-long historical period. From the beginning, significance of teaching excellence has traditionally been emphasized. The university first started to administer paper based student evaluation of teaching questionnaire at the end of every semester. In 2008, instructors of some courses became a volunteer for web based version of SET questionnaire. For those courses, student filled out online questionnaires. Until 2010 Fall semester, the mixed method administration for student evaluation of teaching questionnaire had been continued; online and paper-based. Since 2010 Fall, instructors of all graduate and undergraduate courses in the university have been evaluated by students online.

SET questionnaire has three parts (see the appendix for the content of the item). The first part includes five items about a course and 10 items about instructor effectiveness. Each item has a five point rating scale (5: Excellent, 1: Poor). The second part includes several items about courses

related information such as students' attendance to the course, expected grade from the course, whether the course is required or elective in students' program. The last part includes a textbox where students may write any comments about the course and the instructor.

Sample

In the current study the analysis requires connectedness in data; when every judge rates every person in a study, data is complete. However, if people were rated by different judges and judges cannot be linked through people, we would have subsets in the data and then, connectedness becomes a problem. Therefore, only one faculty out of five was purposefully chosen to guarantee the connectedness; Since Faculty of Science and Art offers courses to all students at the university, selected students in the data set have high chance to provide representative student sample for the population.

In 2015 Fall Academic semester, response rates varied in undergraduate courses. I only included courses if more than six students' evaluated the instructor of the course in order to secure the connectedness among courses. After data cleaning, 254 courses and 1235 students were left in the data.

MFRM Analysis

Before Rasch analysis was conducted, the assumption of unidimensionality was checked. First, factorial structure of the scale was examined by using both exploratory with IBM Statistical Package for Social Sciences 21 and confirmatory factor analysis with MPlus.

Rasch analysis was completed using Facets v. 3.71.4 (Linacre, 1987-2014). I adopted Rasch Rating Scale Models (Andrich, 1978); a three facet and four facet Rasch models. Students, instructor and item are the common facets in both models. A course type and expected grade is an additional facet in the four facet model. Three mathematical models are given in the Appendix. Facets reported that subset connectedness was obtained in the data.

Listed below are some important indexes and evaluation criteria for the analyses of this research.

a) The Infit and Outfit mean square (MnSq) statistics: The Infit and Outfit MnSq statistics reflect the discrepancy between observed and model-driven expected responses and flag unexpectedness in the data (Linacre and Wright, 2002). The value of these statistics range from zero to infinity. In case of perfect correspondence these values become one. A value greater than one indicates that variance is higher than expected. Regarding rater fit statistics, high variance means that a rater rate inconsistently and unpredictably. A value below one signals the existence of lower variance in the data than that predicted by the model. In the case of rater facet, these statistics can be interpreted as too predictable rater behavior. The rater either rates too consistent or do not distinguish between different performances. Linacre and Wright suggest that the Infit and Outfit MnSq statistics values between 0.5 and 1.5.
b) The Separation Ratio (G): G represents a measure of the spread of the estimates relative to their measurement error. It ranges from one to infinity. G = 2, for instance, means that the dispersion in the measures of the elements in the facet is two times greater than the imprecision in their estimates (Wright, 1996). While high G value is desired for item and person facet, low G value is desired for rater facet.
c) The reliability of Separation Index (R): R shows how reproducibly different the measures are. It ranges between zero and one. If R is close to one, there is a high probability that the elements of the facet with high measure estimates actually have higher measures than those with low measure estimates (Linacre, 2009). Similar to G value, while high G value is desired for item and person facet, low G value is desired for rater facet.

f) The Fixed (all-same) chi-square statistics: Hypothesis test is conducted to determine whether or not the estimations of each elements of a facet have the same estimates after accounting for measurement error.
g) Bias analysis (interaction): The interaction between facets will be evaluated by using z-score. An absolute value of z-score greater than 2.0 is considered as an indicator of statistically significant interaction between facets.

RESULTS

Assumption of Unidimensionality

Internal structure of the scale was investigated with exploratory factor analysis (EFA) first and then, confirmatory factor analysis (CFA). While EFA conducted on SPSS yielded one factor structure, CFA conducted on Mplus confirmed one factor model. Moreover, item fit values (examine in details later) show that the data fit to the Rasch model is acceptable and therefore assumption of unidimensionality was secured. Out of 63,810 data points, 614 (0.94%) have a standardized residuals bigger or smaller than three, 2.870 (4.49%) have standardized residuals bigger or smaller than two. These numbers shows that data model fit is acceptable.

The Rating Scale

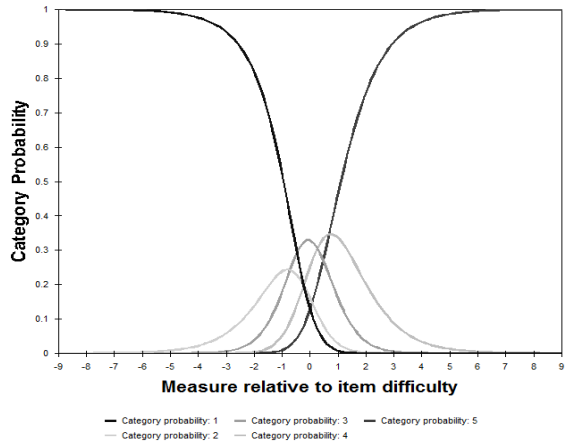

Figure 1. Probability Curves of Five Categories in the Scale

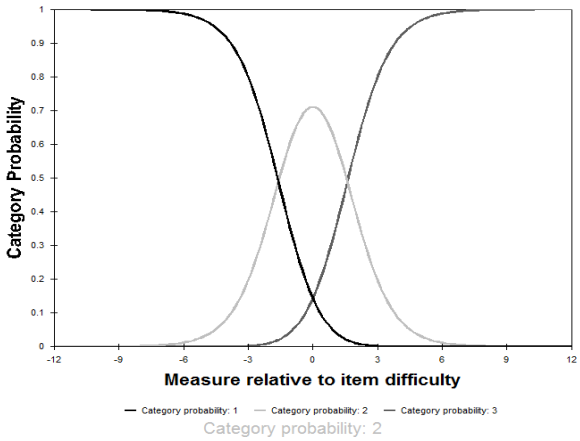

Figure 2. Probability Curves of Three Categories in the Scale

The response scale for items has five categories. This scale is evaluated by how well every point category in the scale conforms to expectations; Figure 1 shows probability curves of categories in the scale. Overlapping categories indicates that the distinction between rating categories students is not clear to students. The measures for thresholds for five categories scale given in Figure 3 show that disordered thresholds exist. This shows that the five category scale does not function as we wish. It is seen that students tend to choice either the first or the last category in the scale.

A five category scale

DATA				QUALITY CONTROL			RASCH-ANDRICH		EXPECTATION		MOST	RASCH-	Cat	Obsd-Expd	Response
Category Counts			Cum.	Avge	Exp.	OUTFIT	Thresholds		Measure at		PROBABLE	THURSTONE	PEAK	Diagnostic	Category
Score	Total	Used	%	%	Meas	Meas	MnSq	Measure	S.E.	Category	-0.5	from	Thresholds	Prob	Residual	Name
1	5496	5091	8%	8%	-.33	-.36	1.3			(-1.74)		low	low	100%	-1.1	lowest
2	4410	4410	7%	16%	.05	.01	1.5	-.03	.02	-.69	-1.22		-.80	23%	-.6	
3	9332	9332	15%	31%	.35	.37	1.0	-.57	.01	-.03	-.34	-.30	-.33	29%	-.5	middle
4	12592	12592	21%	52%	.74	.81	.9	.28	.01	.65	.28	.28	.22	29%		
5	31980	28905	48%	100%	1.59	1.57	1.0	.32	.01	(1.85)	1.26	.32	.86	100%	2.0	highest
										(Mean)		(Modal)	(Median)			

A three category scale

DATA				QUALITY CONTROL			RASCH-ANDRICH		EXPECTATION		MOST	RASCH-	Cat	Obsd-Expd	Response
Category Counts			Cum.	Avge	Exp.	OUTFIT	Thresholds		Measure at		PROBABLE	THURSTONE	PEAK	Diagnostic	Category
Score	Total	Used	%	%	Meas	Meas	MnSq	Measure	S.E.	Category	-0.5	from	Thresholds	Prob	Residual	Name
1	5496	5091	8%	8%	-.61	-.69	1.2			(-2.69)		low	low	100%	-.6	lowest
2	26334	26334	44%	52%	.77	.80	1.2	-1.60	.02	.00	-1.68	-1.60	-1.63	71%	-.9	middle
3	31980	28905	48%	100%	2.72	2.71	1.0	1.60	.01	(2.71)	1.69	1.60	1.62	100%	1.5	highest
										(Mean)		(Modal)	(Median)			

Figure 3. Category Statistics

It appears that collapsing categories 1 and 2, and 4 and 5 optimized the use of the rating scale best because it improved separation and reliability measures and provided better data to model fit than the five category scale (Figure 2). Collapsing middle three categories as an alternative solution did not work. The statistical indicators also showed that combining middle three categories provides poor statistics. Since it is evident that collapsing categories 1 and 2, and 4 and 5 provided the most meaningful information, the data were analyzed on the trichotomous scale.

Facets in the Rasch Model

The extent of the facet contribution to the instructors' score variance was examined. The results are presented for rater facet, instructor facet and item facet respectively as follows.

Rater severity and fit

Figure 4 is a visual representation of Facet analysis results. The first column is the common logit scale, the next three columns present the measures for raters (students) measures and the last column is the scale used in the rating. The second column allows the severity of the raters (students). Their distribution of measures ranged from -7.55 to 9.10 logits severity with a mean of 2.02 and a standard deviation of 2.01. Standard error is .48 with a SD of .47. The results show that majority of logits measures are above zero. This means that majority of students rate their instructor leniently. Moreover, most raters clustered closely around the mean, they are within ±1 logit value. Although the interquartile range is relatively restricted and variability is small, the separation index and reliability was high; 3.93 and .94 respectively. Moreover, the χ^2 of 21348.3 ($p<.000$) was statistically significant and, therefore, the null hypothesis that all raters have the same severity logit estimates must be rejected. In contrast to classical concept of reliability and separation, we do not want high separation index or reliability because we want raters to be equally severe. The separation ratio, 3.93, is an indicator of unwanted variance or construct irrelevant variance and shows that it is 3.93 times greater than the estimation error.

Table 1. Statistics

Statistics	*Examinees*[a] *(Instructor)*	*Raters (Students)*	*Items*
M (measure)	.00[b]	2.02	.00[b]
SD (measure)	1.44	2.01	.25
M (SE)	.17	.48	.03
RMSE	.19	.37	.03
Separation (strata) index (H)	7.69	3.93	7.97
Separation reliability (R)	.98	.94	.98

[a]Examinees with non-extreme scores only
[b]The mean of the measures is constrained in a given facet to be zero.

Outfit and InFit MnSq statistics indicate around 10% of raters had misfit (Table 1). This means that these students rate their instructor inconsistently or consistent with their peers who rate the same instructor. Similarly, around 10% of raters have fit statistics indicating that their ratings are too predictable or provide redundant information. Around 50% of students (611) have InFit value lower than 1.00.

Table 2. Fit Statistics

	Raters (Students)				Instructors				Items			
MnSq	InFit		OutFit		InFit		OutFit		InFit		OutFit	
>1.50	127	(10%)	158	(12.8%)	11	(4.3%)	29	(11.3%)	---		1	(6.7%)
1.5-0.5	980	(80%)	922	(74.7%)	241	(94.6%)	219	(85.9%)	15	(100%)	14	(93.3%)
0.50<	128	(10%)	155	(12.5%)	3	(1.1%)	7	(2.7%)	---		---	

Instructors and fit

The second column in Figure 4 shows teaching effectiveness measure variation among instructors. The instructors are ordered with the most effective at the top and the least effective at the bottom. Measures ranged from -3.58 to 6.45 logits and its distribution is fairly normal around mean of 0.0 with a standard deviation of 1.44; the mean of standard errors of the measures was .17 with a standard deviation of .08. Although the differences in severity are small, the reliability of separation index (7.69) was very high. Instructor separation value is 7.69 that mean this population is separable into 7-8 levels of effectiveness and shows that central tendency effect (Myford & Wolfe, 2004) was not an issue. High separation value provides us high person reliability which is .98. In fact, this coefficient could be a little bit overestimated. For instructor facet, overfit is more of a concern for reliability measure than person estimates. "Overfit tends to stretch the measures along the latent dimension, to reduce their standard errors, and thus, to increase their reliability (or precision); yet, these measures will still be sufficiently accurate for most practical purposes." (p. 102, Eckes, 2015). Fortunately, only small percent (1,1 and 2,7%) of MnSq. values are overfitted.

Items

Table 3 shows statistics and estimates of 15 items in the scale. Observed average of each item is given in the second column of the table on a three category rating scale. It ranged from 2.32 to 2.5. While Item 12 is the easiest to endorse, Item 15 are the hardest to endorse item. Fair average (column 3) is a transformed score of Rasch measures (http://www.winsteps.com/facetman/fairaverage.htm). The items were set to have mean of zero logit. Difference in item measures does not reflect a substantial difference in difficulty. The range is from -.40 to .39 with a mean of .00 and a standard deviation of .25. The fifth column shows the amount of error corresponding to each measure. The error was equal to .03 for all items. Item fit indexes, Infit and Outfit MnSq values are all within acceptable range except Item 10. In fact, Outfit MnSq value is just above acceptable range,

1.77 and Infit MnSq value is in the acceptable range. Linacre argues that it is easy to find misfit when the data size is big enough (http://www.winsteps.com/winman/globalfitstatistics.htm). As a result, we can consider Item 10 has acceptable fit as well.

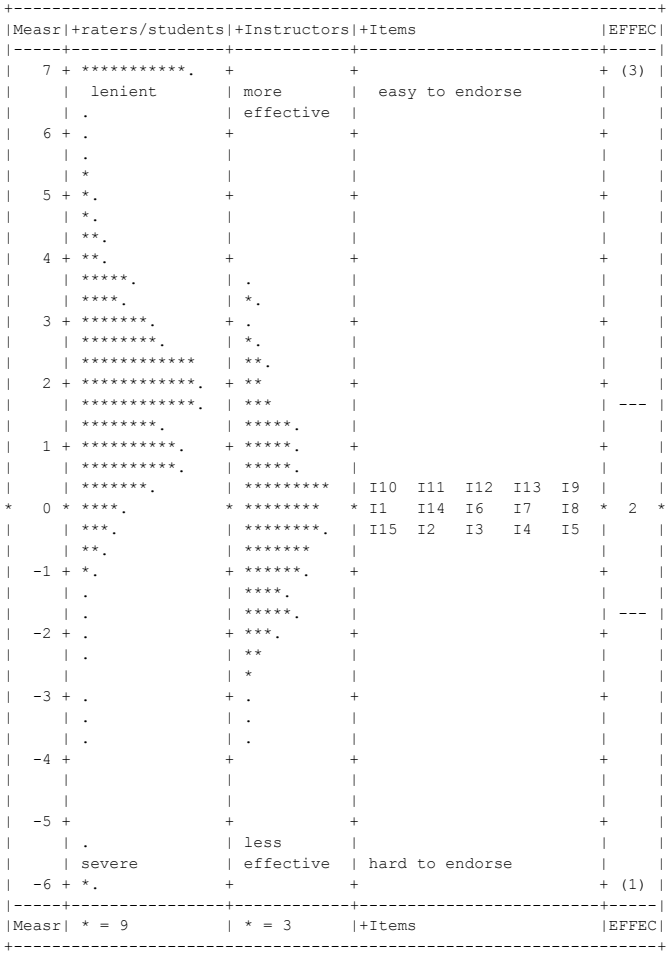

Figure 4. Facets Summary (Rater (students) Severity, Instructor Effectiveness, Item Difficulty)

Table 3. Item Statistics and Estimates

ITEM	Observed average	Fair average	Measure (in logits)	SE	InFit MnSq	OutFit MnSq
15	2.32	2.41	-0.40	0.03	1.31	1.31
5	2.34	2.43	-0.34	0.03	0.87	0.95
2	2.35	2.44	-0.30	0.03	0.91	1.06
4	2.35	2.44	-0.3	0.03	0.9	0.93
3	2.37	2.47	-0.19	0.03	1.05	1.16
14	2.4	2.50	-0.07	0.03	0.86	0.84
7	2.41	2.52	-0.01	0.03	0.98	0.93
1	2.42	2.53	0.01	0.03	0.93	1
8	2.42	2.53	0.02	0.03	1.05	1.06
6	2.45	2.57	0.14	0.03	0.88	0.87
9	2.46	2.58	0.19	0.03	1.04	1.25
10	2.46	2.58	0.20	0.03	1.2	1.77
13	2.48	2.61	0.32	0.03	1	0.98
11	2.49	2.62	0.34	0.03	1.02	1.08
12	2.5	2.63	0.39	0.03	0.98	1.44

Bias Analyses

The second purpose of the study is bias analysis. The students' judging behavior is examined by using the Many-Facet Rasch Model in order to detect any potential source of bias in student evaluation of teaching.

MFRM uses the term 'bias' differently from its meaning in the measurement literature. A bias analysis (an interaction analysis) helps to pinpoint unexpected response pattern by considering more than one facet at the same time. If there is a deviation from what was expected, these patterns point the bias (interactions). Two two-way interaction analyses were conducted; item by course type and item by expected grade.

The course type facet with two elements (required and elective course) was added to my basic three facet model. Bias diagram of course type bias illustrated in Figure 7 in the appendix. As it is seen that students in elective courses rated instructors more positively than the students in required courses. However, a logit measure difference between two course types is not substantial; it is -.07 and .07 for must course and elective course respectively. Interaction analysis indicates that students are able to keep their severity consistently across items on each course type and no evidence of bias were observed in any of the 30 combinations. In other words, instructors of elective courses got always higher ratings on each item than the instructors of the must course did. Figure 8 in the appendix shows bias diagram of t-values which are all within ±2 [χ^2 (30)=7.5, p=1.0].

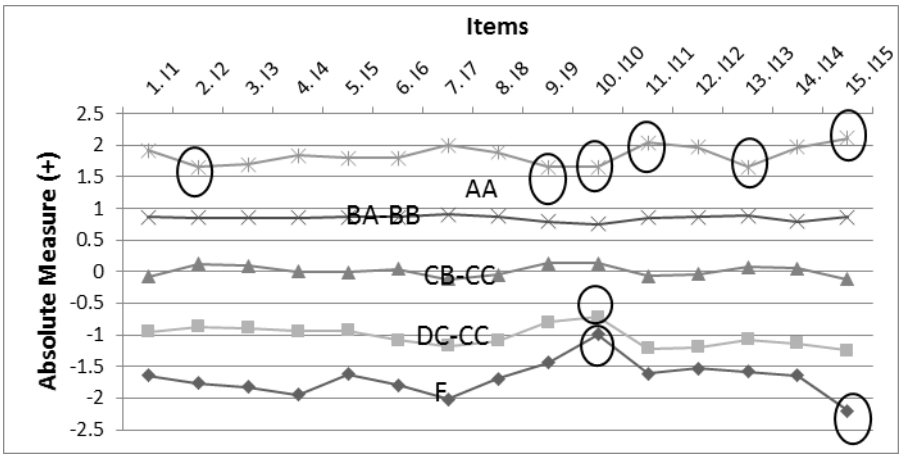

Figure 5. Bias Diagram: Between Items and Expected Grade

Expected grade facet with five elements (AA, BA-BB, CB-CC, DC-DD and F) was added to my basic three facet model. The highest and lowest element difference between measures was 3.52 logits. As the expected grade gets higher, the average rating of the instructors gets higher. Separation reliability of 1.00 shows that this average rating of measures significantly differs across elements of

Observd Score	Expctd Score	Observd Count	Obs-Exp Average	Bias Size	Model S.E.	t	d.f.	Prob.	Infit MnSq	Outfit MnSq	Sq	N	Expected Grade Items	measr	Nu	Ite	measr
183	166.92	89	.18	.69	.21	3.34	88	.0012	1.3	1.3	46	1	F	-1.68	10	I10	-.22
604	582.16	287	.08	.30	.12	2.54	286	.0118	1.1	1.1	47	2	DD DC	-1.02	10	I10	-.22
2538	2489.97	964	.05	.27	.08	3.54	963	.0004	1.3	1.5	75	5	AA	1.84	15	I15	.43
2649	2620.68	964	.03	.20	.09	2.35	963	.0189	1.0	1.5	55	5	AA	1.84	11	I11	-.36
2568	2596.28	964	-.03	-.18	.08	-2.28	963	.0227	1.1	1.1	45	5	AA	1.84	9	I9	-.20
2590	2617.19	964	-.03	-.18	.08	-2.25	963	.0249	1.1	1.1	65	5	AA	1.84	13	I13	-.34
2476	2509.99	964	-.04	-.19	.07	-2.54	963	.0113	1.0	1.9	10	5	AA	1.84	2	I2	.32
2570	2599.14	964	-.03	-.19	.08	-2.36	963	.0185	1.3	2.9	50	5	AA	1.84	10	I10	-.22
141	152.20	89	-.13	-.52	.22	-2.37	88	.0199	1.8	2.2	71	1	F	-1.68	15	I15	.43

Figure 6. Bias/Interaction, Facet Output

the expected grade facet. Expected grade contributes the variance in the model as much as instructors' ability does. Almost 40% of the variance could be explained by the students' expected grade. Bias diagram of expected grade illustrated in Figure 6 shows that there is an interaction between items and expected grades. Nine combination of facet elements out of 75 have a z-score equal or greater than 2. Those elements with statistically significant interactions were shown within a circle in Figure 5 [χ^2 (75)=139.3, p=.00].

Figure 6 provides statistics for nine significant interactions. These elements are shorted according to bias size which is a maximum value of .69 and a minimum value of .18. Students who expected AA from the course rated Item 11 and 15 unexpectedly higher than the models expected. On the contrary, they rated Item 2, 9, 10 and 13 unexpectedly lower. Likewise, students who expected grade lower than CC overrated Item 10 and students expecting to fail the class underrated item 15.

DISCUSSION and CONCLUSION

Economic and political changes in the World have been pushing higher education institutes to exhibit their performances equally well in not only research and but also teaching. European Association for Quality Assurance in Higher Education (2009) emphases that higher education institute should monitor qualification and competence of teaching staff. Although it is widely accepted that SET should not be sole toll to evaluate ones teaching quality, SET result will continue to be used for longer time as internal quality assurance of teaching effectiveness (Penny, 2003) in spite of all arguments against it.

Although student evaluation of teaching has been implementing in western higher education institutes since nearly the beginning of the 20th century, few universities in Turkey have had adopted this evaluation system. This study used the data set obtained from one of these few universities implementing SET. The purpose of this study is to examine what extent do the facets (instructor, student, and rating items) modeled in instructor evaluation contribute to instructors' score variance and examine the students' judging behavior using the MFRM to examine any potential source of bias in student evaluation of teaching effectiveness.

MFRM provides a stronger measurement model than any other method in evaluation of rater mediated assessment. MFRM offers several statistics that helps us examine what extent instructor, student, and rating items in instructor evaluation contribute to instructors' score variance and examine the students' judging behavior using the MFRM to examine any potential source of bias in student evaluation. Rater effect in an evaluation process appears in different forms such as such as severity or leniency, halo or central tendency (Hoyt, 2000; Myford & Wolfe, 2003). Such rater effect introduces a method variance in observed ratings which are associated with the raters and not with examinee. MFRM provides us statistics to evaluate the extent of rater effects. Some of statistics utilized in this study are reliability and separation index, logit measures and Infit and Outfit MnSq. They are discussed respectively in this section.

The interpretation of these statistics depends on the facet considered. Given a small range of logit measures (± 2 logits) separation reliability and separation index for instructor facet is surprisingly high, .95 and 7.69. This result indicates that the spread of the effectiveness measures was considerably much greater than the precision of those measures and most probably big sample sizes resulted high separation among instructors. In general performance assessment it is aimed to differentiate among examinees, therefore, high separation reliability is desired.

The separation reliability and separation index are .92 and 3.93 for student (rater) facet. Unlike instructors (examinee) facet, we do not want high statistics for this facet because ideally we wish equal leniency or severity for raters. When raters practice a highly similar degree, reliability becomes low. Therefore, for this facet low reliability and separation is desired. These statistics showed students differed strongly in the severity with which they rated instructors. In overall, students display a strong leniency effect. This result is supported by Zhao and Gallant (2012).

For item facet, the range of difficulty measures is too small (± .5 logits) and approximately 50% of students answer to items are too predictable, this means that students rated each criteria (item) quite similarly. This result can be interpreted in different ways. These 15 items in the questionnaire is redundant, they almost provide the same information about instructors. Therefore, some items can be eliminated and extra items could be added to widen the separation index among items. A halo effect, another source of rater effect, is signaled by these group level and individual level statistics. "A halo effect refers to a rater's tendency to provide similar ratings of an examinee's performance on conceptually distinct criteria…When the majority of the raters were subject to halo error, the ratings would be highly similar across criteria and, as a result, the criteria showed only little variation in their measures of difficulty. " (Eckes, 2011, p.66). Myford and Wolfe (2004) indicated that rater Infit and Outfit MnSq indices less than one or greater than one, depending on measurement context can be used to diagnose a halo effect. Approximately 50 percent of rater has Infit and Outfit MnSq, values less than one. Low variability in item difficulty measure and large number of MnSq values lower than one draw attention to possible a halo effect. It appears that ratings are highly redundant across criteria.

Another possible rater effect is a central tendency effect. It happens when raters tent to overuse the middle categories of the rating scale. In case of central tendency effect, the scale is only functional for average performing examinees, not with low performing or high performing examinees (Eckes, 2011). This kind of rater effect is not an issue in this study. However, the five point rating scale does not work as expected. Students are clustered at the high end of the five point scale. After rescaling, the results are still similar. Therefore, as a group, students display a strong leniency effect.

So far each facet of the Many Facet Rasch Model was singled out and discussed. The last research question is about potential biasing of the SET questionnaire with respect to course type and expected grade. Exploratory two way interaction analysis, it is also known as bias analysis was used to identify systematic deviations from expectations. The interaction between course type and item facet is not statistically significant. Students in elective courses rated instructors more positively than the students in required courses. Instructors who teach elective courses always get higher average score then instructors teaching required courses. In the second bias analysis, the interaction between students' expected grade facet and item facet was examined. As the expected score gets higher, the instructor score gets higher. Moreover, there is interaction between them. The students expecting AA give the highest score to Item 15 "Overall effectiveness of the instructor" and even higher than the expected score by the model. In contrast, students expecting to fail a course give the lowest average score and even lower than the expected. The finding related to expected grade is supported by previous research (e.g. Dodeen, 2013; Marks, 2000; Marsch, 2007).

In conclusion: Many researchers (e.g. Dodeen (2013), Gursoy and Umbreit (2005), Marks (2000), Marsch, 2007) states that effective teaching is a multidimensional construct. On the other hand, the SET items of this study display unidimensional psychometric structure. Spooren, Brock and Mortelmans (2013) concluded that use of SET in higher education and validity of the scores obtained with SET should continue to be questioned. My conclusion is similar to them. It looks like the most serious threat in SET is halo effect. Halo effect shows that students do not evaluate their instructor as we expected. While evaluating their instructors, they may have different criteria in their minds other than the criteria that the university sets for their instructors.

REFERENCES

Abrami, P. (2001). Improving judgments about teaching effectiveness using teacher rating forms. *New Directions for Institutional Research, 2001*(109), 59-87. http://dx.doi.org/10.1002/ir.4

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 357-74.

Barnes, D. C., Engelland, B. T., Matherine, C. F., Martin, W. C., Orgeron, C. P., Ring, J. K., et al. (2008). Developing a psychometrically sound measure of collegiate teaching proficiency. *College Student Journal, 42*(1), 199-213.

Basow, S. A., & Martin, J. L. (2012). Bias in student ratings. In M.E. Kite (Ed.), *Effective evaluation of teaching: A guide for faculty and administrators*. Retrieved from the Society for the Teaching of Psychology web site: http://teachpsych.org/ebooks/evals2012/index.php

Beran, T., Violato, C., & Kline, D. (2007). What's the 'use' of student ratings of instruction for administrators? One university's experience. *Canadian Journal of Higher Education*, *17*(1), 27-43.

Cashin, W. E. (1995). *Student ratings of teaching: The research revisited. IDEA Paper No. 32.* Retrieved from http://www.faculty.umb.edu/pjt/cashin95.pdf

Centra, J. A. (1993). *Reflective faculty evaluation: Enhancing teaching and determining faculty effectiveness*. San Francisco, CA: Jossey-Bass.

Dodeen, H. (2013). Validity, reliability, and potential bias of short forms of students' evaluation of teaching: The case of UAE University. *Educational Assessment*, *18*, 235-250.

Eckes, T. (2005). Examining rater effects in Testdaf writing and speaking performance assessments: A Many-Facet Rasch analysis. *Language Assessment Quarterly*, *2*(3), 197-221.

Eckes, T. (2009). Many-Facet Rasch measurement. In S. Takala (Ed.), *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment* (Section H). Strasbourg, France: Council of Europe/Language Policy Division.

Eckes, T. (2015). *Introduction to Many-Facet Rasch Measurement. Analyzing and Evaluating Rater-Mediated Assessments*. Frankfurt am Main: Peter Lang.

Ginns, P., Prosser, M., & Barrie, S. (2007). Students' perceptions of teaching quality in higher education: The perspective of currently enrolled students. *Studies in Higher Education*, *32*(5), 603–615.

Gravestock, P. & Gregor-Greenleaf, E. (2008). *Student course evaluations: Research, models, and trends*. Toronto: Higher Education Quality Council of Ontario.

Gursoy, D., & Umbreit, W. T. (2005). Exploring students' evaluations of teaching effectiveness: What factors are important? *Journal of Hospitality & Tourism Research, 29*, 91-109.

Haladyna, T., & Hess, R. K. (1994). The detection and correction of bias in student ratings of instruction. *Research in Higher Education, 35*, 1209–1217.

Hoyt, D. P., Chen, Y., Pallett, W. H., & Gross, A. B. (1999). IDEA Technical Report No. 11: Revising the IDEA systems for obtaining student ratings of instructor and courses. Kansas State University, Manhattan, KS. The IDEA Center. Retrieved from http://ideaedu.org/wp-content/uploads/ 2014/ 11/techreport-11.pdf.

Hoyt, W. (2000). Rater bias in psychological research: When is it a problem and what can we do about it?. *Psychological Methods, 5*(1), 64-86.

Hoyt, W., & Kerns, M. (1999). Magnitude and moderators of bias in observer ratings: A meta-analysis. *Psychological Methods*, *4*(4), 403-424.

Keeley, J. (2012). Course and instructor evaluation. In W. Buskist & V. A. Benassi (Eds.), *Effective college and university teaching. Strategies and tactics for the new professoriate* (pp. 173-180). Thousand Oaks, CA: Sage.

Koh, H., & Tan, T. (1997). Empirical investigation of factor affecting SET results. *International Journal of Educational Management, 11*, 170-208.

Lane, S., & Stone, C. A. (2006). Performance Assessments. In B. Brennan (Ed.), *Educational Measurement*. Westport, CT: American Council on Education & Praeger

Linacre, J. M. (1989). Many-facet Rasch measurement. Chicago: MESA.

Linacre, J. M. (1994). Constructing measurement with a many-facet Rasch model. In M. Wilson (Ed.) *Objective measurement: Theory in practice* (Vol. 2, pp. 129-144) Norwood, NJ: Abex.

Linacre, J. M. (2009). FACETS (Computer program, version 3.66.1). Chicago: MESA.

Linacre, J. M., & Wright, B. D. (2002). Understanding Rasch measurement: Construction of measures from many-facet data. *Journal of Applied Measurement, 3*(4), 486-512.

Lunz, M., Wright, B., & Linacre, J. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education*, *3*(4), 331-345.

Marks, R. (2000). Determinants of student evaluations of global measures of instructor and course value. *Journal of Marketing Education*, *22*(2), 108-119.

Marsh, H. (1982). SEEQ: A reliable, valid, and useful instrument for collecting students' evaluations of university teaching. *British Journal of Educational Psychology*, *52*(1), 77-95.

Marsh, H. W. (1984). Students' evaluation of university teaching: Dimensionality, reliability, validity, potential bias, and utility. *Journal of Educational Psychology, 76*, 707–754.

Marsh, H. (1987). Students' evaluations of University teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research*, *11*(3), 253-388.

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*, 1187–1197.

Messick, S. (1998). Test validity: A matter of consequences. *Social Indicators Research, 45*, 35-44.

Moore, S., & Kuol, N. (2005). Students evaluating teachers: Exploring the importance of faculty reaction to feedback on teaching. *Teaching in Higher Education, 10*(1), 57-73.

Mortelmans, D., & Spooren, P. (2009). A revalidation of the SET37 questionnaire for student evaluations of teaching. *Educational Studies, 35*, 547–52.

Mulqueen, C., Baker, D. P., & Key Dismukes, R. (2002). Pilot instructor training: The utility of the multifacet Item Response Theory model. *International Journal of Aviation Psychology, 12*, 287-303.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4*(4), 386-422.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement, 5*(2), 189-227.

Nelson, J. P., & Lynch, K. A. (1984). Grade inflation, real income, simultaneity, and teaching evaluation. *Journal of Economic Education, 15*, 21–39.

Ory, J. C., & Ryan, K. (2001). How do student ratings measure up to a new validity framework? *New Directions for Teaching and Learning, 5*, 27-44.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. (Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B. D. Wright. Chicago: The University of Chicago Press.

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.

Seldin, P. (1993). The use and abuse of student ratings of professors. *The Chronicle of Higher Education, 39*(46), A40.

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.

Sudweeks, R., Reeve, S., & Bradshaw, W. (2005). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing, 9*(3), 239-261.

Wachtel, H. K. (1998). Student evaluation of college teaching effectiveness: A brief review. *Assessment & Evaluation in Higher Education, 23*(2), 191-211.

Williams, W. M., & Ceci, S. (1997). "How'm i doing?" Problems with student ratings of instructors and courses. *Change: The Magazine of Higher Learning, 29*(5), 12-23.

Wright, R. (1996). A study of the acquisition of verbs of motion by Grade 4/5 early French immersion students. *The Canadian Modern Language Review, 53*(1), 257-280.

Wright, B., & Stone, M. (1979). *Best test design*. Chicago: Mesa Press.

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

Zangenehzadeh, H. (1988). Grade inflation: A way out. *Journal of Economic Education, 19*, 217–230.

Zhao, J. Z., & 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.

Acknowledgements

This work was supported by B.U. Research Fund at Boğaziçi University under contract 14D05P1

UZUN ÖZET

Giriş

Ders değerlendirme anketi (DDA), farklı amaçlar için akademik dönemin herhangi bir noktasında öğrencilerin ders ve öğretim elemanı hakkındaki deneyimlerine dair görüşlerini toplamak amacıyla sıkça kullanılmaktadır (Penny, 2003, Seldin, 1993; Zabaleta, 2007). DDA sonuçları yardımıyla, yüksek öğretim kurumlarında yöneticiler (a) öğretim niteliğini artırmayı, (b) yükseltme veya ödül gibi karar noktalarında veri sağlamayı ve (c) kurumsal hesap verebilmek amaçlı kanıt sağlamayı hedeflemektedirler (Seldin, 1993 ; Spooren, Brock ve Mortelmans, 2013).

DDA yardımıyla yapılan değerlendirmeler bir tür performans değerlendirme olarak kabul edilebilir. Performans değerlendirmede, değerlendirmeye tabi olan kişi bir performans gerçekleştirir ve/veya bir ürün oluşturur ve bu performans veya ürünün kalitesi, en az bir değerlendirici tarafından puanlanır. Performans değerlendirme, bir puanlayıcı aracılı değerlendirme süreci olduğundaysa ekstra önlemler daha güvenilir ve adil bir değerlendirme uygulamaları oluşturmak için dikkate

alınması gerekir. Değerlendirici aracılı değerlendirmede en yaygın tehditlerden birisi 'değerlendirici varyansı"dır. Bu terim, değerlendirilen kişinin performansının kendi beceri seviyesinin yanı sıra değerlendiricinin özelliklerine (katılık/cömertlik, cinsiyet gibi) bağlı olmasına karşılık gelir (Eckes, 2009). Başka bir değişle, ölçmek istediğimiz yapıyla ilgisiz varyanstan oluşan değerlendirici varyansı, performans değerlendirmenin adaletini ve geçerliğini tehdit eder (Messick, 1998; Lane & Stone, 2006).

Alan yazımının gösterdiği gibi, DDA'ya ilişkin teorik ve psikometrik tartışmalar süregelmektedir (Gravestock & Gregor-Greenleaf, 2008). Akademik çalışmalar aslında birbiriyle ilişkili iki ana kaygı etrafında toplanmıştır. Birincisi kaygı elde dilen puanların geçerliğiyle ilgilidir. DDA alan yazınımda "Etkili öğretim" diye tanımlayabileğimiz yapıyı ne derece ölçtümüz soru işaretidir. İkinci kaygıysa elde edilen puanlarda ortaya cıkabilecek ve puanların geçerliğini ve güvenirliğini tehdit edebilecek yanlılık kaynaklarıdır (Gürsoy & Umbreit, 2005). Bütün bu bağlam içinde, bu çalışmanın amacı, a) öğretim elemanlarının puanlarındaki farklılığa/varyansa, değerlendirme sürecindeki elemanların (öğretim elemanı, öğrenci ve değerlendirme maddeleri) ne derece katkı sağladığını ve b) öğrencilerin değerlendirmelerinde yanlılığa yol açacak potansiyel kaynakları çok yüzeyli Rasch modeli yardımıyla incelemektir.

DDA'in, 1920 yılında Washington Üniversitesi'nde ilk kullanımından bu yana oldukça büyük ölçüde alan yazımı oluşmuştur. O zamandan bu yana, DDA ile elde edilen puanların geçerliliği tartışma konusu olduğu gibi, DDA kullanımının eğitimin niteliğini artırıp artırmadığı gibi konular yeni tartışma konuları olarak alana girmiştir. Akademik araştırmaların çoğunluğu Kuzey Amerika, Avustralya ve İngiltereki yüksek öğrenim bağlamı içinde yapılmıştır (Gravestock & Gregor-Greenleaf, 2008; Zabaleta, 2007). Bu çalışmaların çoğu (Abrami, 2001; Beran, Violato & Kline, 2007; Gravestock, & Gregor-Greenleaf, 2008; Marsh, 1987 gibi) DDA'nin kullanımları için genel olarak olumlu bir tutuma sahiptir; Öte yandan, bazı araştırmacılarda nitelikli öğretimle ilgisi olmayan ders ve öğretim elemanı özellikleri yüzünden yanlı sonuçlar verebileceğinden dolayı DDA'nın kullanımında şüpheci bir tutum sergilemektedirler (Dedeen, 2013; Koh & Tan, 1997; Williams & Ceci, 1997). Görüldüğü gibi alan taraması birbirleriyle çelişecek sonuçlar vermektedir. Bu nedenle, DDA sonuçları çok önemli kararlarda, işe alma, promosyon veya yükseltmelerde tek başına kullanılmaması gerektiği düşünülmektedir.

Bir performans değerlendirmede, kişinin puanı bu süreçteki bir grup aktöre bağlıdır. Bunlardan en sık görülenleri; performans görevini alan kişinin beceri seviyesi, performans görevinin zorluk derecesi (Mulqueen, Baker & Dismukes, 2002) ve puanlayıcı etkisidir. Değerlendirme sürecindeki puanlayıcı etkisi farklı sekillerde ortaya çıkabilir. Bunlardan bazıları, katılık/cömertlik, halo etkisi, ve merkezi eğilim etkisidir (Hoyt, 2000; MyFord & Wolfe, 2003). Bu tür puanlayıcı etkileri gözlenen puanlarda metot varyansı oluşturur ve bu varyans performans görevini alan kişiyle ilgili değil, puanlayıcı ile ilgilidir. Farklı bağlamlarda yapılan arastırmalar gösteriyorki puanlayıcı aracılı performans değerlendirmelerinde puanlayıcı etkisi çok fazladır (Eckes, 2005). Örneğin, Hoyt ve Kerns (1999) meta analiz araştırmasında performans görevisi alanların performanlarının %37'si puanlayıcı etkisi ve puanlayıcı-sınavı alan kişi arsındaki etkileşimle açıklanabilir. Birden fazla puanlayıcının olduğu değerlendirme süreçlerinin güvenirliğini test etmek için daha standart prosedürlerden, modern test teorilerinin faydalanıldığı bir dizi yöntem vardır.

Yöntem

Bu çalışmada, büyük bir şehirinde bulunan orta boy bir devlet üniversitenin lisans derslerinden ders değerlendirme anketleriyle toplanan ders ve öğretim elemanı değerlendirme verilerini kullanılmıştır. DDA üç bölümden oluşmaktadır: İlk bölüm ders ve tasarımı hakkında beş madde ve öğretim elemanın etkinliği hakkında ise 10 madde içerir. Her madde beş puanlaı derecelendirme ölçeğine (1, Mükemmel: 5, Kötü) sahiptir. İkinci bölümde ise derse katılım, dersten beklenen not ve dersin programdaki türü (zorunlu/seçmeli) gibi bilgileri ölçen maddeler yer almaktadır. Son bölümde ise öğrencilerin ders ve öğretim elemanıyla ilgili geri bildirimlerini yazabilmeleri için bir metin kutusu sağlanmıştır.

Veri setinde bağlantıyı kurabilmek için tüm üniversiteye çok sayıda ders açan Fen Edebiyat Fakültesi amaçlı olarak seçilmiştir. 2015 Güz Akademik döneminde, DDA uygulandığı ve en az altı öğrencinin katıldığı 254 dersten 1.235 öğrencinin verisi analiz edilmiştir. Rasch analizinden önce, tek boyutluluk varsayımı doğrulayıc ve açımlayıcı faktör analiziyle incelenip, doğrulanmıştır. Rasch analizleri Facet v. 3.71.4 (Linacre , 1987-2014) kullanılarak tamamlanmıştır. Rasch Derecelendirme Ölçeği Modeline (1978 Andrich) dayanan bir üç yüzeyli, iki adet dört yüzeyli modeller kullanılmıştır.

Sonuç ve Tartışma

Analizde kullanılan toplam 63.811 verinin standardlaştırılmış değerinin ±3'den büyük ya da eşit olanlarının sayısı 614 (%0.96), ±2'den büyük olanların sayısı ise 2.870 (%4.49) olarak elde edilmiş ve model veri uyumu sağlanmıştır.

Puanlayıcı yüzeyi incelendiğinde; öğrencilerin farklı katılık derecesine sahip oldukları görülmektedir. Ayırma güvenirliği ve indeksi 0,92 ve 3,93'dür. İdeal durumda puanlıyıcıların eşit katılık derecesine sahip olması beklenir.

Anket maddeleri incelendiğinde maddelerin zorluk dereceleri ±0,5 logit arasında değiştiği ve öğrencilerin yaklaşık %50'sinin cevapları oldukça tahmin edilebilir olduğu görülmüştür. Bu durum iki ayrı şekilde yorumlanabilir. Bu maddeler neredeyse aynı bilgiyi sağlamaktadır bu nedenle bu 15 maddenin hepsi gerekli değildir. Bazı maddeler çıkarılırken, ayırma indeksini yükseltecek şekilde yeni maddeler eklenebilir. Bunun yanı sıra, madde zorluk derecelerindeki düşük varyazyon, grup seviyesindeki muhtemel alo etkisinide de dikkat çekmektedir. Myford and Wolfe (2004) uyguluk içi ve dışı istatistik değerlerinin ölçmede birey seviyesinde halo etkisini belirlemek için kullanılabileceğini belirtmişlerdir. Madde zorluk değerlenlerindeki düşük varyans ve yüksek sayıda birden küçük öğretim elemanı uyum indeks değerleri muhtemel halo etkisine dikkat çekmektedir.

Diğer muhtemel puanlayıcı etkisi ise merkezi eğilim etkisidir. Merkezi eğilim etkisi, puanlayıcı ölçeğin orta kategorilerinin gerektiğinden fazla kullanılmasıyla ortaya çıkar. Bu tür etki gözlenmemektedir. Fakat beş puanlı ölçek istenilen şekilde çalışmamaktadır. Öğrencilerin ölçeğin üst kategori çok kullandığı görülmektedir. Bu durumda öğrencilerin bol notlu davranışa sahip olduklarını göstermiştir.

Farklı yüzeyler arasındaki etkileşime bakıldığında, maddelerle ders tipi (seçmeli/ zorunlu) arasında bir etkileşim olmadığı belirlenmiştir. Fakat maddelerle mğrencinin beklediği not arasında bir etkileşim bulunmuştur. Dersten AA bekleyen bir öğrenci Madde 15'e beklenenden yüksek puan verirken, dersten kalmayı bekleyen öğrenci beklenenden daha düşük puanlama yapmıştır.

APPENDIX

A Three Facet Model

$$\ln\left(\frac{P_{nij}}{P_{ni(k-1)}}\right) = \beta_n - \delta_i - \gamma_j - \tau_k \qquad \text{where}$$

P_{njimk}, = probability of instructor *n* receiving a rating of *k* on criterion *i* from student *j*,
$P_{njim(k-1)}$ = probability of person *n* receiving a rating of *k*-1 on criterion i from rater *j*,
β_n = ability of person *n*,
δ_i = difficulty of criterion (item) *i*,
γ_j = severity of rater *j*,
τ_k = difficulty of receiving a rating of *k* relative to a rating of *k*-1.

The first four Facet model

$$\ln\left(\frac{P_{nij}}{P_{ni(k-1)}}\right) = \beta_n - \delta_i - \gamma_j - \pi_m - \tau_k \qquad \text{where}$$

P_{njimk}, = probability of instructor *n* receiving a rating of *k* on criterion *i* from student *j*,
$P_{njim(k-1)}$ = probability of person *n* receiving a rating of *k*-1 on criterion i from rater *j*,
β_n = ability of person *n*,
δ_i = difficulty of criterion (item) *i*,
γ_j = severity of rater *j*,
π_m = severerity of course type m
τ_k = difficulty of receiving a rating of *k* relative to a rating of *k*-1.

The second four Facet model

$$\ln\left(\frac{Pnij}{Pni(k-1)}\right) = \beta_n - \delta_i - \gamma_j - \pi_m - \tau_k$$ where

P_{njimk}, = probability of instructor *n* receiving a rating of *k* on criterion *i* from student *j*,
$P_{njim(k-1)}$ = probability of person *n* receiving a rating of *k*-1 on criterion i from rater *j*,
β_n = ability of person *n*,
δ_i = difficulty of criterion (item) *i*,
γ_j = severity of rater *j*,
π_m = severity of expected grade m
τ_k = difficulty of receiving a rating of *k* relative to a rating of *k*-1.

SET Questionnaire: The content of the items

1. Course objectives
2. Course design
3. Course materials
4. Course requirements and assignments
5. Overall effectiveness of the course
6. Course materials
7. Awareness of students' comprehension
8. Encouragement of student participation in class
6. Effective use of class time
7. Grading practices
8. Fair grading
9. Fair handling of objections to grades
10. Availability to help
11. Overall effectiveness of the instructor
12. I would choose to take another course with the same instructor

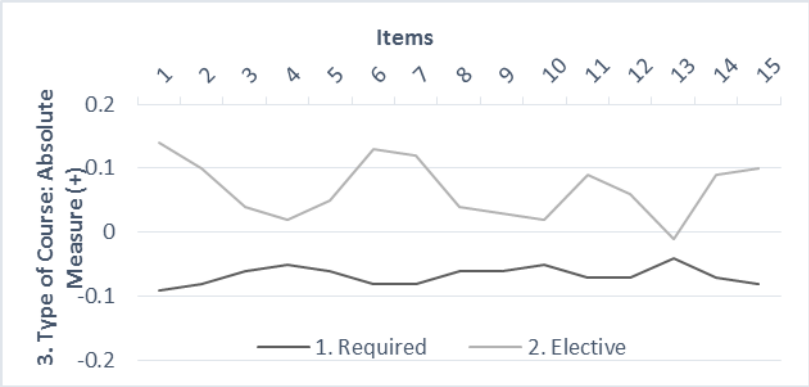

Figure 7. Bias Diagram Showing the Interaction Between Items and Course Type

Figure 8. Bias Diagram Showing the Interaction Between Items and Course Type
Series 1=required 2=elective

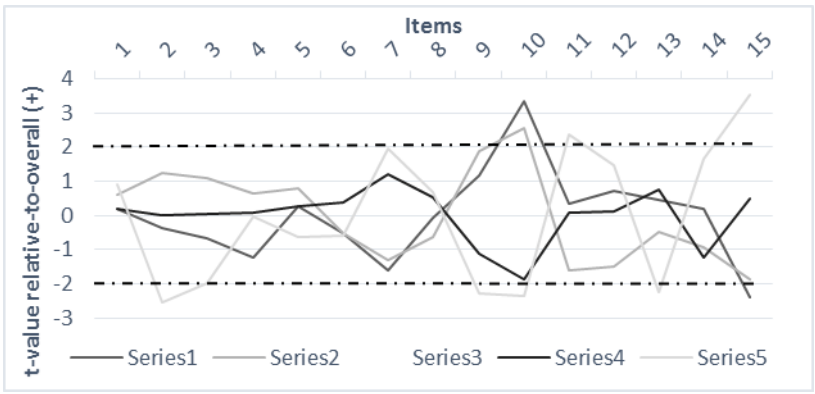

Figure 9: Bias diagram showing the interaction between items and expected grade
Series 1=F, 2=DC-CC, 3=CB-CC, 4=BB-BA, 5=A