ISSN 1799-2591
Theory and Practice in Language Studies, Vol. 1, No. 11, pp. 1531-1540, November 2011

doi:10.4304/tpls.1.11.1531-1540

A Many-facet Rasch Model to Detect Halo Effect in Three Types of Raters

Farahman Farrokhi
Department of English, University of Tabriz, Tabriz, Iran
Email: ffarrokhi20@yahoo.co.uk

Rajab Esfandiari (Corresponding author)
Department of English, University of Tabriz, Tabriz, Iran
Email: rbesfandiari@gmail.com

***Abstract*—Raters play a central role in rater-mediated assessment, and rater variability manifested in various forms including rater errors contributes to construct-irrelevant variance which can adversely affect an examinee's test score. Halo effect as a subcomponent of rater errors is one of the most pervasive errors which, if not detected, can result in obscuring an examinee's score and threatening validity and fairness of second language performance assessment. To that end, the present study is an endeavor to detect halo effect in L2 essays, using a relatively newly employed methodology, a many-facet Rasch model (MFRM) in language assessment. The participants in this study consisted of 194 raters—subdivided into self-rater, peer-rater, and teacher rater—who rated 188 essays written by 188 undergraduate Iranian English majors at two state-run universities in Iran. The collected data were rated using a 6-point analytic rating scale and were analyzed using the latest version of Facets 3.68.0 to answer the research question of the study. The results of facets analysis showed that, at group level, the raters did not exhibit any sign of halo effect, but, at individual level, all rater types displayed considerable halo effect. Further analysis revealed that rater types were unanimous about halo effect on four items and that self-rater showed more of a halo effect compared to the other two rater types.**

***Index Terms*—rater variability, rating scale, MFRM, raters, halo effect**

I. Introduction

Rasch models comprise a growing family of models (Fisher, 2007). The basic Rasch model was introduced by the Danish mathematician and statistician George Rasch (1960, 1980) and was later developed and made widely known by Wright (Wright & Stone, 1979). The basic Rasch model was a probabilistic model premised on the assumption that a person's score on a test is the result of the person ability and task difficulty and is mathematically expressed as the probability or likelihood of a particular person with a certain ability on a particular item to get a score right or wrong is the function of the ability of that person and the difficulty of that item (Baghaei, 2009). The basic Rasch model or standard dichotomous model handles dichotomously scored items such as multiple choice items.

Andrich's (1978) rating scale analysis extended the basic Rasch model in which data from likert-type scales could be handled. For Andrich's model, the estimation is based on the probability of a certain candidate with a certain ability of getting a certain score on a scale for an item. Masters's (1982) partial credit model was an extension of Andrich's model in which a range of marks could be awarded to a response depending on its quality. Partial credit model provides information for individual items on rating scale thresholds or steps; in other words, it states how easy or difficult it is for an individual candidate with a certain ability to move from one score point on the scale thresholds to another for an individual item (Bond & Fox, 2007) To put in a nutshell, "Rasch models allow one to make probabilistic statements about item difficulty, candidate ability, and rating scale thresholds. Such statements are expressed in terms of units called logits, the logarithm of the odds of a certain outcome" (McNamara & Adams, 1991, p. 3).

Technically speaking, we still have two facets: ability and difficulty. Linacre (1989/1994) went a step further. He added leniency/severity of judges as a third facet, hence a many-facet Rasch model. The many-facet Rasch model is an extension and generalization of the partial credit model.

In what follows, we will provide the requirements of the many-facet Rasch measurement followed empirical studies done on halo effect. We will then embark upon reviewing halo effect.

II. Review of Related Literature

A. Requirements and Assumptions of the Many-facet Rasch Measurement

As was mentioned earlier, the many-facet Rash measurement belongs to a growing family of Rasch models, which applies to a class of measurement models that aim at providing a fine-grained analysis of multiple variables potentially having an impact on test or assessment outcomes and is usually used for performances that are awarded subjective

ratings, such as essays or speaking assessments (Eckes, 2009). As such, it is no different from other Rasch models in that it must meet certain requirements and assumptions. These requirements include unidimentionality, invariance, additivity, ordering, and fit statistics. Our discussion will be very brief. For a detailed explanation, readers are referred to Baghaei (2009) and Baghae and Amrahi (2009).

The many-facet Rasch model is unidimentional. In other words, all items on a test should measure the same single underlying variable or construct (Eckes, 2009). The many-facet Rasch measurement is invariant, which means that "parameters—person agility, item difficulty, and judge severity—should be independent of each other' (Linacre, 1989/1994, p. 42). In other words, examinee measures are invariant across different sets of items or tasks or raters and item, task, or rater measures are invariant across different groups of examinees. Invariance is also called specific objectivity (Rasch, 1960), measurement invariance (Bond & Fox, 2007), and parameter separation (Smith, Jr., 2004), which all express more or less the same conceptual notion in that the ability estimates of persons are freed from the distributional properties of the specific item attempted. Likewise, the estimated difficulties of items are freed from the distributional properties of specific people used in the calibration.

The many-facet Rasch measurement is additive, which refers to the properties of the measurement units expressed in logits which are equal-interval over the entire continuum of a scale. In Rasch models, persons and items conjoin to define the common interval scale and item characteristics curves which are based on the estimation and calibration of these two factors should never cross each other, but they should be parallel, the violation of which is un-unidimentionality (Smith Jr., 2004). This is technically known as ordering. Fit statistics which act as quality control indicators show how closely the data fit the model (Baghaei &Amrahi, 2009). Rasch models are prescriptive, ideal robust models, but data are messy and incomplete and may never perfectly fit the model. As far as the data fit the model usefully, it shows good data-to-model fit.

B. Conceptual Definitions of Halo Effect

Wells (1907) is generally credited with first identifying the effect (Myford & Wolfe, 2004a). Thorndike (1920, p. 25) coined the term and defined it "as a marked tendency to think of the person in general as either good or rather inferior and to color the judgments of the qualities by their general feeling. This same constant error toward suffusing ratings of special features with a halo belonging to the individual as a whole."

In the field of language testing, Yorozuya and Oller, Jr. (1980) were probably the first researchers to investigate this "judgmental bias," as they would prefer to call it, and defined it as "a tendency for judges to assign similar scores across the various scales … For instance, a judge rating an interviewee high on, say, the Vocabulary scale might also assign a high rating on Grammar and each of the other scales quite independently of the constructs supposedly underlying the scales. This kind of judgmental bias could be called a halo effect—a kind of spillover across scales causing them to be more strongly correlated with each other" (p. 136).

In the context of MFRM analysis, the halo effect is defined as "a rater's tendency to assign ratees similar ratings on conceptually distinct traits" (Myford & Wolf, 2004b, p. 209). In other words, raters fail to distinguish between conceptually distinct and independent aspects of ratees' performances and give them similar ratings across those traits (Engelhard, 2002). More recently and in line with the above-two mentioned definitions in light of MFRM, Eckes (2009, p. 5) rightly noted that "this effect [halo effect] manifests itself when raters fail to distinguish between conceptually distinct features of examinee performance, but rather provide highly similar ratings across those features; for example, ratings may be influenced by an overall impression of a given performance or by a single feature viewed as highly important." In this paper, when it comes to halo effect, it is this definition which we will be working with.

C. Approaches to Minimize the Halo Effect

There are many ways to employ to reduce the halo effect as it affects the learner's performance. Myford and Wolfe (2004a) propose six approaches to minimize it, but we only quote one of them which we used in our study. To minimize halo effect, Myford and Wolfe (2004a, p. 396) recommend researchers to "train raters to be aware of the halo effect and the impact it can have on their ratings so that they can attempt to guard against this tendency."

D. Studies Done on Halo Effect Using MFRM

Studies having employed MFRM to investigate halo effect are very rare. Below we will mention some of the studies conducted in this area. In a study of Georgia state writing assessment program for 8th-grade students in the USA, Engelhard (1994) employed 15 highly experienced, trained raters to rate 264 compositions, using a 4-point analytic scale on five domains or categories including. He investigated four rater effects: severity or leniency, halo, central tendency, and restriction of range. The results of the Facets analysis of mean square fit statistics showed that two out of 15 raters' ratings were muted, indicating the presence of halo effect, which means that they tended to rate holistically and failed to differentiate among students. One of these raters rated 8 out of 17 students with uniform rating patterns and the other rater rated 14 out of 21 students with uniform rating patterns.

In L2 field, there are a couple of studies which have investigated halo effect. As early as 1980s in a pioneering work, Yorozuya and Oller, Jr. (1980) used 15 trained graduate and undergraduate native English speakers to rate the tape recorded performances of 10 students on an interview. The ratings were done on two different conditions, using four 10-point scales of grammar, vocabulary, pronunciation, and fluency. Under condition one, the raters rated only one of

the four scales. Under condition two, rated each interview on four scales at one hearing. Stronger higher correlations across scales rated on the same occasion than on separate occasions were indicative of halo effect. Halo effect was manifested under condition one in which all four scales were marked at a single hearing of each interview. Furthermore, the mean squared loading under condition one was .89 and under condition two it was .79. "[This] 10 (10%) difference can be read as a halo effect" (p. 146). Yorozuya and Oller, Jr. (1980, p. 146) concluded that "scales rated on the same occasion are contaminated by a sizable halo effect."

In Japan, in an endeavour to develop a criterion to recommend competent students as professional medical translators to a translation agency for which the translator training program was offered, Kozaki (2004) employed GENOVA and FACETS to set multiple standards on performance assessment. To that end, 9 performances of 20 performances of adult native Japanese speakers with mixed levels of performance in translation from Japanese medical papers to English were chosen. All nine examinees were either professionally or amateurishly involved in translation of Japanese into English. Four professionally and bilingual judges literate in both Japanese and English were chosen to judge the performance of the examinees on a 4-point rating scale on seven assessment categories. The judging consisted of three procedures to arrive at a cut-off score: assessing the categories independently, rank ordering the performances, and pass-fail ratings by at least three judges. Judge behavior was analyzed by Facets to examine judge severity, central tendency, and halo effect. The author showed that raters 1 and 4 assigned unexpectedly harsh ratings to the weakest examinee and unexpectedly lenient ratings to the most able examinee. She interpreted such unexpected ratings to be signs of halo effect, which "judges carry over the impression of competence… creating non-independence of assessment categories… grammar or vocabulary or both" (Kozaki, 2004, pp. 21-22).

In New Zealand, Knock, Read, and von Randow (2007) compared the effectiveness of on line and face-to- face feedback to individual raters within the context of a large-scale academic writing assessment of students entering a major English–medium university. Sixteen highly trained experienced native and non native English as a second language teachers were equally divided into online and face-to-face group. The study was done in four phases: pre-training, training, post-training rating, and post-training feedback. The raters used a 4-9-band rating scale on three categories fluency, context, and form to rate 70 candidates' scripts of the writing sections of the low-stakes test, DELNA. Four rater effects were investigated: rater severity, internal consistency, central tendency and halo effect. Using group and individual statistics indicators of Facets, the authors claimed that at group level there was no sign of halo effect either before or after training, but, at individual level, they argued that very low rater fit mean square indices could be an indication of halo effect. Three raters displayed a halo effect after training and three others rated in a more differentiated fashion after training.

III. The Present Study

The present study uses many-facet Rasch measurement (hereafter MFRM) to detect halo effect in self-raters, peer-raters, and teacher raters. The present study compensates the limitations of the previous studies in the following ways: it incorporates 194 raters, with 188 acting as students raters and six others as teacher raters, it employs English major students, it is conducted in a different EFL setting, Iran, and finally it employs a fully-crossed design which was lacking in the previous studies.

A. Research Question

In the present study, we were interested in how three rater types, namely, self-rater, peer-rater, and teacher rater, showed variability in terms of halo effect in relation to each other. The following research question was, therefore, generated: To what extent do self-rater, peer-rater, and teacher rater display halo effect when rating the essays of students using an analytic scale?

B. Participants

The participants in the present study consisted of 194 raters, who were subdivided into student raters and teacher raters. Student raters were 188 undergraduate Iranian English majors enrolled in Advanced Writing classes in two state-run universities in Iran, comprising three fields of study: English Literature, Translation Studies, and English Language Teaching. The student raters were labeled either self-assessors or peer-assessors. Teacher assessors were six Iranian teachers of English.

Student raters ranged in age from 18 to 29, with one over 30, and another with unidentified age. One hundred and thirty one student raters (69.7%) were female and 57 (30.3%) were male. Eighty-six (45.7%) were native Farsi-speakers, 68 (36.2 %) were native-Turkish speakers, four (2.1%) were native-Kurdish speakers and another four (2.1%) were grouped as "Other". Ninety-five (50.5%) were sophomores, 29 (15.4%) were juniors, and 64 (34.0%) were seniors. Only three of them (1.6%) had the experience of living in an English-speaking country. The number of years they had studied English ranged from 1 to 24 years and most of them (61.7%) had studied the English language in language institutes before entering the university.

Teacher assessors were all male. They came from two language backgrounds: four teacher assessors were native-Farsi speakers, and the other two were native-Turkish speakers. They ranged in age from 23 to 36. None of them had the experience of living in an English-speaking country. They had taught writing courses from one to seven years.

Three of them were affiliated with a national university, one of them with a private university, and two of them were classified as "Other". All of them had a degree in English: three of them were PhD students in ELT, two had MAs in ELT, and one had a BA in English literature.

C. The Rating Scale

For the purposes of the present study, we chose an analytic rating scale. The scale we developed for the present study is based on Jacobs Jacobs, Zinkgraf, Wormuth, Hartfiel and Hughey's (1981) ESL Composition Profile, but differs from it in many aspects (See the appendix).

To develop our rating scale, we also referred to writing textbooks in the literature because we wanted the scale to reflect the structure of a standard five-paragraph essay, so the following three books were consulted as a guide to compose the scale categories: Composing with confidence: Writing effective paragraphs and essays, Refining composition skills: Grammar and rhetoric, and The practical writer with readings. The scale contains fifteen items which comprise the elemental features of a five paragraph essay: substance, thesis development, topic relevance, introduction, coherent support, conclusion, logical sequencing, range, word choice, word form, sentence variety, overall grammar, spelling, essay format, and punctuation, capitalization, and handwriting.

A six-point rating scale was chosen because these are "the most common number of scale steps in college writing tests, and a larger number of steps may provide a degree of step separation difficult to achieve as well as placing too great a cognitive burden on raters, while a lower number may not allow for enough variation among the multifaceted elements of writing skills" (Schaefer, 2008,p. 473).

D. Data Collection

One hundred and eighty five-paragraph essays were collected over a span of a year and a half from 188 students enrolled in advanced writing courses in two state-run prestigious universities in two different cities in Iran. The students came from six classes taught by four instructors. The students in advanced writing classes are taught punctuation, expression, features of a well-written paragraph, and principles of a one-paragraph and five-paragraph essay. The students were taught these principles of writing for eight weekly meetings. Immediately after the eight weekly meetings, they were told by their respective teachers they would have to sit the midterm exam the following week.

At the exam they were given 90 minutes to write a five-paragraph essay ranging in length from five hundred to seven hundred words on the following topic: **In your opinion, what is the best way to choose a marriage partner? Use specific reasons and examples why you think this approach is the best.** This topic was chosen from a list of TOEFL TWE topics. All the students were given one topic in order to control for topic effect. Following the data collection, a rating session was held with all the student raters and teacher raters, in which they were fully instructed how to rate the essays. The latest version of Facets 3.68.0 (Linacre, 2011) was used to analyze data.

E. Data Analysis

To analyze the data from the rater-type mediated data in the present study, we employed the latest version of Facets 3.68.0, a computer software package designed and developed by Linacre (2011). We followed Myford and Wolfe (2004b) and Engelhard (2002) procedures for analyzing data on halo effect when MFRM is used. Engelhard (2002) proposes both individual and group indices as units of analysis, but he takes into account only mean square fit statistics (outfit mean square and infit mean square) to detect halo effect for both individual and group raters. Like Engelhard (2002), Myford and Wolfe (2004b) also consider two types of indices, but, unlike him, their methods of detection are more elaborative and all inclusive. Myford and Wolfe (2004b) recommend researchers to use both individual-level statistics indicators and group-level statistics indicators.

To determine the halo effect via individual-level statistics indicators, Myford and Wolfe (2004b) propose the following procedures: (1) the researcher should first look at fit indices—infit mean square less than 1 and outfit mean square greater than 1 indicate halo effect; (2) the researcher should always have an eye for rater's observed ratings and the model's expected ratings—any mismatch between the rater's and the model's is a sign of halo effect in light of fit indices; (3) the researcher performs a Rater x Trait bias-interaction analysis—t. score either greater than 2 or smaller than -2 indicates rater misfit; (4) and finally the researcher examines the observed and expected ratings for raters flagged as misfit.

To determine halo effect via group-level statistics indicators, Myford and Wolfe (2004b) propose the following four indicators. The **fixed chi-square** tests the "fixed effect" hypothesis that all traits share the same degree of difficulty measure. A non-significant value may indicate a halo effect in ratings of all raters. The **Trait Separation Ratio** is an index of the spread of the trait difficulty measures relative to their precision. A low trait separation measure suggests halo in the ratings. **Trait Separation Index** implies the number of measurably different levels, or strata, of trait difficulty. A low trait separation index suggests halo in the ratings. As Myford and Wolfe (2004b) state, this index may be large when the number of raters or ratees is large. **Reliability of the Trait Separation Index** provides information about how well one can differentiate among the items in terms of their levels of difficulty. Again a low trait separation may connote a halo effect. This index should be ideally 1.

IV. RESULTS

The present study employs a fully crossed design in which all raters rated all essays. The data was analyzed with Facets 3.68.0, a software program for MFRM (Linacre, 2011).Three facets were specified for this study: students, rater type, and items. The mathematical formula for facets is given below:

$\text{Log}\ (P_{nirk}/P_{nir}(k\text{-}1)) = B_n - D_i - T_r - F_k$

Where:

P_{nirk} = the probability of student n being rated k on item i by rater type r,

$P_{nir}(k\text{-}1)$ = the probability of student n being rated k-1 on item i by rater type r,

B_n = the proficiency of student n,

D_i = the difficulty of item i,

T_r = the severity of rater type r, and

F_k = the difficulty of scale category k, relative to scale category k-1.

A. Initial Analysis

Before answering the research question, we did a preliminary Facets run to test for data-model fit. The results of the analysis showed that Students 94, 101, and 160, Raters 22, 24, 27, 48, 74, 76, 95, 145, and 176, and Item 7(logical sequencing) were misfits. The common practice in the literature (See McNamara 1996) is to delete the misfitting elements. However, as the purpose of this study is to examine rater effects, and not to refine a test instrument, this approach was regarded as inappropriate, as we might end up throwing out the baby with the bath water. Valid but unexpected ratings may reveal valuable insights into rater behavior, and so a different approach was adopted, which may be called a "lazer strategy" rather than a "scalpel strategy" (Myford, personal communication).

First we identified and deleted individual cases of highly unexpected ratings. We then reran the analysis and this time found no misfitting elements. According to Linacre (2011), satisfactory model fit is indicated when about 5% or less of (absolute) standardized residuals are ≥ 2, and about 1% or less of (absolute) standardized residuals are ≥ 3. In our data, there were a total of 19,699 valid responses, that is, responses used for estimation of model parameters. Of these, 697 responses were associated with (absolute) standardized residuals ≥ 2, and 45 responses were associated with (absolute) standardized residuals ≥ 3, so the number of unexpected responses is much smaller than Linacre considers, indicating satisfactory model fit.

B. Reliability and Validity of Rating Scale

The category statistics (Figure 1) and the probability curves (Figure 2) provide the necessary information about the rating scale, indicating that the rating scale functioned reliably and validly with rater type. According to Linacre (2004), in order for a rating scale to function effectively, there should be at least ten observations in each category, average measures should advance monotonically with counts, outfit-mean squares should be less than two, and step difficulty or step calibration should advance by 1.4, but less than 5 logits. Figure 1 shows that the rating scale meets all these guidelines: there are more than ten observations at each point, average measures advance monotonically, outfit mean squares are almost perfect (1.0), and the categories are the most probable ones, showing that the steps are appropriately ordered and function well (the guideline that step calibrations should advance by 1.4 is only true when the categories are dichotomous; otherwise, it can be ignored, Linacre, personal communication). Myford also claims that for student achievement the levels are appropriately ordered (Myford, personal communication).

Category Statistics
Model = ?,?,?,R6

DATA			QUALITY CONTROL			RASCH-ANDRICH		EXPECTATION		MOST	RASCH-
Category	Counts	Cum.	Avge	Exp.	OUTFIT	Thresholds		Measure at		PROBABLE	THURSTONE
Score	Used	%	Meas	Meas	MnSq	Measure	S.E.	Category	-0.5	from	Thresholds
1	875	4%	-.03	-.08	1.0			(-2.25)		low	low
2	1877	10%	.02	.05	.9	-.78	.04	-1.03	-1.64	-.78	-1.27
3	3491	18%	.17	.18	1.0	-.51	.02	-.34	-.65	-.51	-.58
4	5132	26%	.31	.31	1.0	-.14	.02	.24	-.06	-.14	-.08
5	5317	27%	.44	.44	1.0	.34	.02	1.03	.58	.34	.48
6	3007	15%	.60	.59	1.0	1.09	.02	(2.44)	1.76	1.09	1.43
		100%						(Mean)		(Modal)	(Median)

Figure 1 shows the reliability and validity of the rating scale.

Figure 2 shows the student probability curves and is a graphic illustration of Figure 1. The probability curves show the threshold at which students are likely to be scored at the next highest level. These should resemble a range of hills (Linacre, 2004). That is, as ability level increases on the logit scale, the probability increases of achieving the next highest score ranking. Figure 2 is also another means to help us decide whether the rating scale functions well.

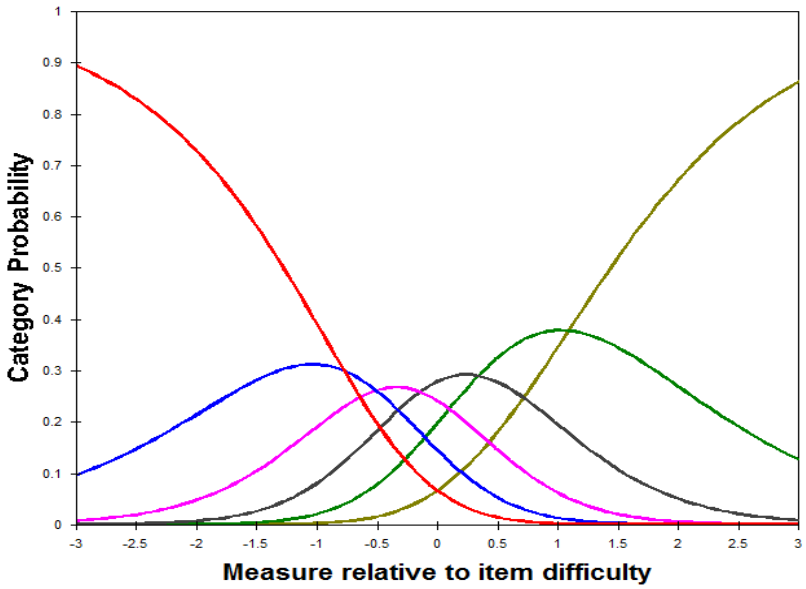

Figure 2 shows probability curves for students.

C. Research Question Analysis

To answer our research question (To what extent do self-rater, peer-rater, and teacher rater display halo effect when rating the essays of students using an analytic scale?), we first present the group-level statistics indicators and then go for individual-level statistics indicators as discussed in the data analysis section. Group-level statistics indicators could be obtained from table 7 of Facets output. Figure 3 shows the group-level statistics indicators for the assessment criteria or items of our rating scale. As can be seen, the figure contains a lot of information, but for our own study purposes, we only need the bottom of the table where those four statistics are shown.

Total Score	Total Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Corr. PtBis	Nu	Items
5800	1320	4.4	4.53	-.21	.02	.87	-3.6	.87	-3.6	1.13	.21	1	Substance
5223	1312	4.0	4.13	.08	.02	1.06	1.6	1.05	1.4	.99	.33	2	Thesis development
4829	1309	3.7	3.84	.26	.02	1.37	9.0	1.37	9.0	.44	.23	3	Topic relevance
5301	1323	4.0	4.16	.06	.02	.94	-1.7	.94	-1.7	1.14	.33	4	Introduction
5522	1310	4.2	4.36	-.08	.02	.82	-5.2	.82	-5.1	1.22	.26	5	Coherent support
5252	1312	4.0	4.15	.06	.02	1.13	3.4	1.13	3.4	.76	.21	6	Conclusion
5866	1291	4.5	4.68	-.32	.02	1.28	6.5	1.23	5.4	.73	.22	7	Logical sequencing
5148	1320	3.9	4.05	.13	.02	.71	-9.0	.72	-9.0	1.40	.20	8	Range
4241	1319	3.2	3.34	.54	.02	.96	-1.1	.96	-1.1	.98	.18	9	Word choice
5726	1299	4.4	4.55	-.22	.02	.87	-3.6	.85	-4.0	1.17	.20	10	Word form
4687	1318	3.6	3.70	.34	.02	.77	-7.6	.78	-7.3	1.34	.23	11	Sentence variety
5383	1308	4.1	4.26	-.01	.02	.76	-7.5	.75	-7.7	1.35	.25	12	Overall granmmar
6393	1319	4.8	4.96	-.59	.03	1.15	3.3	1.13	3.0	.88	.14	13	Spelling
5797	1315	4.4	4.55	-.22	.02	1.16	4.0	1.15	3.6	.81	.25	14	Essay format
5089	1324	3.8	3.99	.16	.02	1.22	6.1	1.23	6.2	.68	.18	15	Punctuation
5350.5	1313.3	4.1	4.22	.00	.02	1.00	-.4	1.00	-.5		.23	Mean (Count: 15)	
516.8	8.8	.4	.40	.27	.00	.20	5.5	.20	5.4		.05	S.D. (Population)	
535.0	9.1	.4	.41	.28	.00	.21	5.7	.20	5.6		.05	S.D. (Sample)	

Model, Populn: RMSE .02 Adj (True) S.D. .27 Separation 11.91 Strata 16.21 Reliability .99
Model, Sample: RMSE .02 Adj (True) S.D. .28 Separation 12.33 Strata 16.77 Reliability .99
Model, Fixed (all same) chi-square: 2034.6 d.f.: 14 significance (probability): .00
Model, Random (normal) chi-square: 13.9 d.f.: 13 significance (probability): .38

Figure 3 shows the group-level statistics indicators.

The results from the **Fixed Chi Square** test for the items are shown in the bottom line of Figure 3. The chi-square value of 2034.6 with 14 degrees of freedom is statistically significant ($p < .05$), indicating that at least two items are significantly different in terms of their difficulty. These results suggest that there is not a group-level halo effect present in our data set. As was stated in the data analysis section, a non-significant chi square means halo effect, but in our study chi square is statistically significant.

The **Trait Separation Ratio** of 12.33 as shown in the second line from the bottom of our Figure 3 signals that the spread of the item difficulty measures is about 13 times larger than the precision of those measures. This indicator does not suggest a group-level halo effect, because as was discussed in data section, a low trait separation ration was suggestive of a halo effect.

The **Trait Separation Index** of 16.17 suggests that there are over 17 statistically distinct strata of item difficulty in this sample of items. This index used to be calculated manually in the earlier versions of Facets and the formula for its calculation is (4G + 1) / 3, but in the later versions Facets gives us this index in its output. There is no evidence here of a group-level halo effect. In order for ratings to display a halo effect, this index should be low, but in our analysis it is large enough.

The **Reliability of the Trait Separation Index** is .99, which is very close to ideal 1. This high degree of reliability of the trait separation index implies that rater type could reliably distinguish among the items. Therefore, this indicator does not suggest a group-level halo effect in this data set because as we discussed in the data analysis section, the closer the reliability to 1, the more reliably the rater type could distinguish among items.

The group-level statistics indicators as was explained and illustrated in this section showed no sign of halo effect for rater type, namely, self-rater, peer-rater, and teacher rater. This might be an indication of good results, but at group level many individual characteristics are not shown and the ratings are averaged over the students across items, so individual-level statistics should be closely scrutinized to see if the same results will be obtained or not. Unfortunately in some studies, these individual-level statistics indicators are ignored to interpret the findings of their studies positively and favorably, hence giving an incomplete picture of the reality.

To determine the halo effect at an individual level, we performed a rater type x item bias-interaction analysis and the results are shown in Table 1. Such an analysis is in line with the explanations given at the data analysis section.

TABLE 1
SHOWS RATER TYPE X ITEM REPORT FROM FACETS OUTPUT ANALYSIS FOR HALO EFFECT.

Rater Type	Logits	Items	Logits	Obs. Score	Exp. Score	Obs-Exp. Average	Bias Size	Model S. E.	t.score	Infit MnSq	Outfit MnSq
Self	-.17	2	.08	704	625.2	.55	-.47	.08	-5.56	1.0	1.0
Self	-.17	3	.26	715	584.5	.92	-.77	.09	-8.63	.9	.9
Self	-.17	4	.06	711	628.9	.57	-.50	.09	-5.81	.9	.9
Self	-.17	6	.06	661	628.0	.23	-.18	.08	-2.36	.8	.8
Self	-.17	7	-.32	620	674.3	-.39	.33	.07	4.48	.5	.6
Self	-.17	8	.13	549	619.9	-.49	.32	.07	4.94	.6	.6
Self	-.17	9	.54	460	519.9	-.42	.25	.06	3.90	1.0	1.0
Self	-.17	10	-.22	615	668.3	-.38	.30	.07	4.21	.7	.7
Self	-.17	13	-.59	689	724.1	-.25	.26	.08	3.20	1.1	1.1
Peer	.05	2	.08	617	563.2	.39	-.28	.07	-3.70	.9	.9
Peer	.05	3	.26	621	515.0	.78	-.54	.08	-6.99	1.1	1.1
Peer	.05	7	-.32	594	642.4	-.35	.26	.07	3.71	.6	.6
Peer	.05	10	-.22	589	628.9	-.28	.21	.07	2.95	.7	.7
Peer	.05	15	.16	516	552.1	-.26	.16	.07	2.43	.9	1.0
Teacher	.12	2	.08	3902	4034.7	-.13	.08	.02	3.29	1.1	1.1
Teacher	.12	3	.26	3493	3729.5	-.23	.14	.02	5.72	1.4	1.4
Teacher	.12	4	.06	3989	4093.8	-.10	.06	.02	2.60	.9	.9
Teacher	.12	7	-.32	4652	4549.4	.10	-.08	.03	-2.83	1.5	1.5
Teacher	.12	10	-.22	4522	4428.8	.09	-.07	.03	-2.49	.9	.9

Fixed (all = 0) chi-square: 414.6 d.f.: 45 significance: .00: p□.00

Note: Items: 1=Substance, 2=Thesis development, 3=Topic relevance, 4=Introduction, 5=Coherent support, 6=Conclusion, 7=Logical sequencing, 8=Range, 9=Word choice, 10=Word form, 11=Sentence variety, 12=Overall grammar, 13=Spelling, 14=Essay format, 15=Punctuation.

Table 1 shows the individual-level statistics indicators. Column one shows the rater type—self-rater, peer-rater, and teacher rater. Column two shows logits or measure for rater type. Column three shows the items of the rating scale. Column four shows logits or measure for these items. Column five shows the observed scores, which are the sum of ratings the rater type gave across students on that particular item. Column six is the expected cores, which are the sum of the ratings across students on that particular item the model gives us. Column seven is the average between observed scores and expected scores. Column eight shows the bias size, which is the size of bias measure in logits relative to overall measures. Column nine shows standard errors which are low, indicating the good precision of our measurement. Column 10 is t.score, whose value greater than 2or smaller than -2 shows bias. Columns 11 and 12 show fit indices: Infit MnSq, an abbreviation for infit mean square and Outfit MnSq, an abbreviation for outfit mean square. These should be ideally 1, below which shows overfit and above which shows underfit.

As is evident in the table, the standard errors (SEs) are low, and the mean square fit statistics are good, with no cases of misfit (for the purposes of the present study, following Wright, Linacre, Gustafson and Martin-Lof, (1994), we chose .5-1.5 range). Out of 45 bias terms, only 19 were statistically significant, with t-scores either greater than +2 or smaller than -2. Eleven of the significant interactions are positive (showing severity), and eight of the significant interactions are negative (showing leniency). Rater type showed statistically significant bias toward only 10 out of 15 items (items 2,3,4,6,7,8,9, 10, 13 and 15). Self-rater shows nine statistically significant interactions, teacher rater shows five, and peer-rater also shows five.

We can now very easily determine that at individual level rater type shows signs of halo effect. We first look at the t. score, which is either greater than 2 or smaller than -2 for ten out of 15 items. Next we look at the observed ratings and

expected ratings. As you can see, for negative t.score, observed scores are higher than expected scores, which implies that the rater type assigns higher ratings to students on those items, and for positive t.score, observed scores are lower than expected scores, which suggests that the rater type assigns lower ratings to students on those items.

On closer inspection, we discern that self-rater showed more halo effect, because this rater type showed nine cases of positive and negative t.score; the other two rater types—peer-rater and teacher rater—showed equal halo effect because they showed equal number of cases of either positive or negative t.score. The interesting point concerning this halo effect among rater type is that although rater type varied in terms of showing halo effect, the rater type unanimously displayed halo effect toward four items of the rating scales—items 2, 3, 7, and 10. This common pattern is worthy of attention.

V. Discussion, Conclusions, and Pedagogical Implications

The present study set out to detect halo effect in three rater types—self-rater, peer-rater, and teacher rater, employing a relatively newly employed methodology, MFRM, in language assessment. To our knowledge, this is the first study which ventured to examine halo effect in a new way with distinct rater types. The results of our Facets analysis at group-level statistics indicators including Fixed Chi Square, Trait Separation Ratio, Trait Separation Index and Reliability of the Trait Separation Index showed that rater type displayed no sign of halo effect, but the analysis at individual-level statistics indicators including the rater type x bias analysis revealed that all rater types showed considerable halo effect. The results also showed that self-assessor on the whole displayed more of a halo effect compared to peer-assessor and teacher assessor. The analysis of findings further revealed that the rater type displayed unanimous halo effect toward four items of the rating scale, providing us with a pattern concerning halo effect.

The halo effect as defined in the literature is the carry-over from one judgment to another, that is, assigning similar ratings to ratees across items. That rater type did not show halo effect reflects the very fact that at the group level rater type can distinguish between conceptually distinct items of the rating scale, which connotes that the rating scale as whole was functioning properly with rater type. Another plausible explanation is that at group level, ratings are averaged across students and across items of the rating scale. This summing up disguises many idiosyncratic features which might be disclosed at individual level. Group level analysis findings, though tentative, might not be suggestive.

Halo effect was shown at individual level among rater type in our study. Generally, our findings confirm those of previous studies (Engelhard, 1994; Kozaki, 2004; Knock, Read, and von Randow, 2007; Yorozuya and Oller, Jr., 1980). Engelhard (1994) found that two out of his 15 highly trained raters displayed halo effect, but he failed to explain why such halo effect occurred with two of his so highly trained raters. Comparing generalizability theory with many-facet Rasch measurement in determining halo effect, Kozaki (2004) also found that two out of his four professionally and bilingual judges showed signs of halo effect on two categories of grammar and vocabulary. She attributed this halo effect to powerful roles these two categories played in the assessment and judges carried over the impression of competence. Knock, Read, and von Randow (2007) attributed halo effect to lack of training and feedback to her raters because after training and feed back at least some of the raters did not show halo effect in the face-to-face group, but halo effect remained with raters in the on line group even after training and feedback. In Yorozuya and Oller, Jr.'s (1980) study, all 15 raters showed halo effect and the halo effect, according to the authors' explanations, could be accounted for the conditions under which the raters were rating. There were two conditions and the raters in condition one rated all four scales—grammar, vocabulary, pronunciation, and fluency—at a single hearing rather than at separate hearings and this led to higher intercorrelations of scales, hence the appearance of halo effect.

We could argue for three main reasons why rater type exhibited halo effect in the present study. Firstly, it should be noted that in our study all the raters were inexperienced, having no rating experience in other contexts. This especially holds true about the self-rater and peer-rater who for the first time rated their own essays and those of their peers' essays. Therefore, it is no surprise that they would exhibit halo effect, given the highly experienced raters in previous studies did display it. Secondly, the amount of training which the rater type in our study received was very slight compared to other raters in other studies (cf. Knock, Read, & von Randow, 2007). Our raters got only one-hour training, but in the previous studies sometimes the raters were given hours of training; besides, some of the raters in the previous studies, although they were rater staff for prestigious corporations such as ETS, underwent retraining. Thirdly, lack of feedback could also account for the emergence of halo effect. Training followed by feedback could have helped raters in this study to improve and show less sign of halo effect just as it did in Knock, Read, and von Randow's (2007) study. Although group training has turned out to work to reduce rater errors such as halo effect, the effectiveness of individualized feedback is yet to be fully investigated, especially in longitudinal studies (knock, 2011).

Another pressing issue relates to the halo effect at individual level in the present study and previous studies, using MFRM. What does it connote? To answer this question, we should compare generalizability theory (hereafter G. Theory), which entered the scene to obviate the shortcomings of classical test theory and many-facet Rasch measurement, which was an extension of the basic Rasch model. These models have been compared and contrasted to examine if they could yield the same results and what differences, if any, could be revealed in employing these two techniques. (cf. Bachman, L.F., Lynch, B.K, & Mason, M, 1995; Lynch & McNamara 1998). Both G. Theory and MFRM strive for identifying the relative effects of variance attributable to facets and interaction among the facets, but while G. Theory does so at group level, MFRM not only does so at group level but also at individual level. Although, as

Kozaki (2004) claims that these two techniques are complementary, Linacre (1993) concludes that G. Theory provides a general summary with no implications for individual examines apart from the number of observations that are made because for each examinee it is the raw score that matters most, but MFRM focuses on the individual examinees and for each examinee a measure is estimated which is statistically as independent as possible of the particularities of the raters, items, tasks and so on. G. theory is not capable of producing "a linear measure of each examinee's performance level, qualified by its standard error and quality-control fit statistics" Linacre, 1993, p. 3). This is a quality of MFRM which makes it distinct from G. Theory.

Halo effect is definitely detrimental to the students' test scores and could potentially distort their final test scores, especially when it comes to rater-mediated assessment (McNamara, 2000) or judge-awarded ratings (Linacre, 2004). Halo effect as a subcomponent of rater variability is a characteristic of raters, not tests and contributes to construct-irrelevant variance which adversely affects examinees' scores. It has to be minimized as far as possible; otherwise, the students might be victim of this rater error rather than their true ability because when raters fail to distinguish conceptually distinct items or traits which have been designed to measure examinees' abilities, they indirectly obscure these measures or abilities. The studies reviewed in this paper all unanimously emphasize the role of experience, training, and feedback which could reduce, if not eliminate, this potential error.

The present study has a number of implications. The first implication goes to the syllabus designers at higher education to introduce self-assessment and peer-assessment into higher education based on recent research. Hasty introduction of these assessment tools will have dire consequences because of the unwanted variability which they might bring into testing situation. This concerns mostly self-assessor who exhibited a larger number of halo effect than that of either peer-assessor or teacher assessor. Self-assessment should be used with special care because it is so idiosyncratic.

The second implication relates to rating purposes for summative purposes in language assessment because both self-assessors and peer assessors have been employed recently to rate their own products, so if they are going to be accurate and reliable raters alongside teachers, they should be given training. Rater training which includes familiarization activities, practice rating, feedback, discussion and monitoring, could lead to the fairness and reliability of these student raters. Although both self-assessor and peer-assessor showed halo effect in the present study, this does not preclude us from not allowing them to assess their own writing because we did not include all the elements of rater training.

The present study has some limitations too. The first one is the small number of teacher assessors, although our study employed more teacher assessors compared to previous studies which employed from one to four. Still this size is small, and future studies should strive for more teacher assessors. The second limitation concerns the number of essays each self-assessor and peer-assessor rated. This produces more model standard error which reduces reliability of ratings. Future studies should undertake to have self-assessors and peer assessors more writing products.

APPENDIX ESSAY RATING SHEET

Essay number:							
Rater's name:							
		Very poor	Poor	Fair	Good	Very good	Excellent
1.	Substance	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐
2.	Thesis development	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐
3.	Topic relevance	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐
4.	Introduction	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐
5.	Coherent support	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐
6.	Conclusion	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐
7.	Logical sequencing	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐
8.	Range	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐
9.	Word choice	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐
10.	Word form	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐
11.	Sentence variety	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐
12.	Overall grammar	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐
13.	Spelling	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐
14.	Essay format	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐
15.	Punctuation/capitalization/handwriting	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐

ACKNOWLEDGMENTS

This study is based on the final report of a research project toward the PhD thesis of the second author and was financially supported by the Research Office of the University of Tabriz. The grant was given to both authors.

REFERENCES

[1] Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika* 43.9, 561–573.

[2] Bachman, L.F., B.K.. Lynch, and M. Mason. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing* 12.2, 238–257.
[3] Baghaei, P. &. N. Amrahi. (2009). Introduction to Rasch Measurement. *The Iranian EFL Journal* 5, 139-154
[4] Baghaei, P. (2009).Understanding the Rasch model. Mashhad: Mashhad Islamic Azad University Press.
[5] Bond, T. G., & C. M. Fox. (2007). Applying the Rasch model: Fundamental measurement in the human sciences (2nd edn.). Mahwah, NJ: Erlbaum.
[6] Eckes, T. (2009). *Many-facet Rasch measurement*. Retrieved June 1, 2011, from http://www.coe.int/t/dg4/linguistic/Source/CEF-refSupp-SectionH.pdf.
[7] Engelhard, G. Jr. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement* 31.2, 93–112.
[8] Engelhard, G., Jr. 2002. Monitoring raters in performance assessments. In G. Tindal & T. Haladyna (Eds.) *Large-scale assessment programs for ALL students: Development, implementation, and analysis*. Mahwah, NJ: Lawrence Erlbaum Associates, 261-287.
[9] Fischer, G. H. (2007). Rasch models. In C. R. Rao & S. Sinharay (Eds.), *Psychometrics: Handbook of statistics* (Vol. 26). Amsterdam: Elsevier, 515–585
[10] Jacobs, H. L., S. A., Zinkgraf, D. R., Wormuth, V. F., Hartfiel and J. B. Hughey. (1981). *Testing ESL composition: A practical approach*. Rowley, MA: Newbury House.
[11] Knoch, U., Read, J., & von Randow, T. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing* 12.2, 26–43.
[12] Kozaki, Y. (2004). Using GENOVA and FACETS to set multiple standards on performance assessment for certification in medical translation from Japanese into English. *Language Testing*, 21.1, 1–27.
[13] Linacre, J. M. (1993). Generalizability Theory and many-facet Rasch measurement. Paper presented at the annual meeting of the American educational research association, 1993, Atlanta.
[14] Linacre, J. M. (2004). Optimizing rating scale effectiveness. In., E. V., Smith, Jr, & R. M. Smith. (Eds.). *Introduction to Rasch model*. Maple Grove, Minnesota: JAM press, 258-278.
[15] Linacre, M. (1989/1994). Many-facet Rasch measurement. Chicago: MESA press.
[16] Lynch, B.K. & T.F. McNamara. (1998) Using G-theory and Many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing* 15.2, 158–180.
[17] Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika* 47.2, 149–174.
[18] McNamara, T. F. (2000). Language testing. Oxford, UK: Oxford University Press.
[19] McNamara, T. F. (1996). Measuring second language performance. New York: Longman.
[20] McNamara, T. F., and R. J. Adams. (1991). Exploring rater behavior with Rasch techniques. Paper presented at the annual testing research colloquium, March 21-23, in Princeton, NJ. (ERIC Document Reproduction Service No. ED345498).
[21] Myford, C. M. and E. W. Wolfe. (2004a). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. In E. V. Smith and R. M Smith (Eds). *Introduction to Rasch measurement*. Maple Grove, MI: JAM Press, 518-574.
[22] Myford, C. M. and E. W. Wolfe. (2004b). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. In E. V. Smith and R. M Smith eds. *Introduction to Rasch measurement*. Maple Grove, MI: JAM Press, 460–517.
[23] Rasch, G. (1960/1980). Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago Press. (Original work published 1960)
[24] Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing* 25.4, 465–93.
[25] Smith, Jr. E. V. (2004). Evidence for the reliability of measures and validity of measure interpretation: A Rasch measurement perspective. In E. V. Smith, Jr & R. M Smith (eds.), *Introduction to Rasch model*. Maple Grove, Minnesota: JAM Press, 93-122.
[26] Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology 4*, 25-29.
[27] Wells, F. L. (1907). A statistical study of literary merit. *Archives of Psychology* 1, (Monograph No. 7).
[28] Wright, B. D., & Stone, M. H. (1979). Best test design. Chicago: MESA Press.
[29] Wright, B. D., J. M. Linacre, J. E. Gustafson, & P. Martin-Lof. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*, 370. Retrieved June 1, 2011, from http://rasch.org/rmt /rmt83b.htm
[30] Yorozuya, R. & J. W. Oller, Jr. (1980). Oral proficiency scales: Construct validity and the halo effect. *Language Learning*, 30.1, 135-153.

Farahman Farrokhi received his PhD in English Language Teaching from Leeds University, England. Currently, he is an associate professor at the University of Tabriz. His research interests include classroom discourse analysis, EFL teachers' perceptions of different feedback types, negative and positive evidence in EFL classroom context and language testing.

Rajab Esfandiari is currently a PhD candidate at the University of Tabriz and a visiting scholar at Ochanomizu University. He has been teaching undergraduate English majors for the past few years. His research interests include teaching and assessing writing, researching L2 classroom assessment, and employing many-faceted Rasch measurement model in language assessment.