Psicológica (2014), 35, 385-397.

Analysis of rater severity on written expression exam using Many Faceted Rasch Measurement

Gerardo Prieto* and Eloísa Nieto

Universidad de Salamanca, Spain

This paper describes how a Many Faceted Rasch Measurement (MFRM) approach can be applied to performance assessment focusing on rater analysis. The article provides an introduction to MFRM, a description of MFRM analysis procedures, and an example to illustrate how to examine the effects of various sources of variability on test takers' performance on a writing test by means of a MFRM analysis. Results highlight the usefulness of the MFRM to detect raters that have extreme values on the severity continuum. MRFM provides a common metric for the facet scores (test takers, tasks, raters). This is advantageous because it facilitates understanding of the assessment process as well as providing objective measurement of facet elements.

Ratings that raters assign to test takers' responses to constructed-response tasks do not depend solely on the respondents' level of performance. Other facets that may affect their ratings must also be taken into account, such as task difficulty, the severity of the rater, and the appropriate use of *scoring rubrics* that are composed of categories which describe various levels of performance. Rater behavior must undoubtedly be taken into consideration in order to validly assess the construct in question (Lane & Stone, 2006). Raters can differ in their interpretation of tasks and/or the categories included in rubrics, in the level of severity they exercise, in the extent to which they are influenced by their general impression of a test taker (e.g., a halo effect) or by test taker background characteristics such as gender or cultural background. These differences all

* Acknowledgements: This article would not have been possible without the contributions of the *Instituto Cervantes* and *Cursos Internacionales* of the University of Salamanca, S.A. Correspondence should be sent to Gerardo Prieto, Facultad de Psicología, Universidad de Salamanca, Avenida de la Merced, 109-131, 37005 Salamanca, Spain. E-mail: gprieto@usal.es

contribute to measurement error, to invalidity, and to the lack of fairness of assessment.

The aim of this paper is to analyze rater differences in the levels of severity/leniency they exercised using the *Many Faceted Rasch Measurement Model* (Linacre, 1989). In this study the raters in question were assessing a written expression tasks forming part of the Diploma in Spanish as a Foreign Language (DELE) Level C1 Exam. These exams are administered by the Instituto Cervantes (Spain). Level C1 is described as *Effective operational proficiency*. To attain this level of command of the language, respondents must be able to write clear and well-structured texts about complex topics, highlighting main ideas, defending points of view with complementary ideas, motives and examples, and provide a suitable conclusion. The data analyzed in this paper were collected from test takers at a first sitting of the DELE-C1 exam in written expression (writing).

The need to establish a system of quality control in subjective scoring led us to implement the *Many Faceted Rasch Measurement* (Linacre, 1989) model. As we shall describe below, this model permits analysis of the actions of different raters on different tasks, and enables us to determine, in part, whether the scoring categories appearing on rubrics must be adjusted or changed in order to obtain more consistent or valid scores. A MFRM analysis allows us to obtain measures from raw scores on many of the factors affecting the quality of a writing exam. The MFRM model, an extension of one of the most well-known Rasch models (i.e., the Partial Credit Model, Wright & Masters, 1982), has been shown to be useful in assessing the behavior of raters when scoring exams on speaking and writing in English, German and Spanish (Eckes, 2011; Kondo-Brown, 2002; Park, 2004; Prieto, 2011; Tyndall & Kenyon, 1996).

By using a MFRM approach to analyze rating data, we obtain on a common equal-interval logit scale the estimations of the parameters of the facet elements involved in the assessment (the performance of the test takers, task and criterion difficulty and rater severity in the variable). In the present study emphasis was placed on assessing the severity of the raters. In the context of a MFRM analysis, rater *severity* is defined as the tendency of a rater to assign scores to respondents that on average are lower than expected if the scores given by other raters to the same group of test takers are taken into consideration. Similarly, rater *leniency* can be understood as a rater's tendency to assign on average higher scores than expected if we take into account the scores given by other raters to the same respondents (Myford & Wolfe, 2004b).

Many Faceted Rasch Measurement (MFRM)

The MFRM model is an extension of the Partial Credit Model for polytomous items in which a test taker's performance is scored using one or more rubrics, each of which is composed of a set of ordered categories. The model can be applied to cases in which there are diverse measurement factors (test takers, tasks, raters, criteria, etc.) that can contribute to measurement error. This model allows us to represent, controlling for measurement error, the additive contribution of each facet to the *logit* or logarithm of the ratio between the probability that an respondent will receive one score on the task (for example, 3) and the probability of that same respondent receiving the immediately lower score (2).

To wit,

$$\log (P_{nijlk} / P_{nijl(k-1)}) = B_n - D_i - R_j - C_l - F_{jk} \quad (1)$$

when

P_{nijlk} = the probability that test taker *n* will receive score *k* on criterion *l* for task *i* from rater *j*,

$P_{nijl(k-1)}$ = the probability that test taker *n* will receive the next lower score (*k-1*) on criterion *l* for task *i* from rater *j*,

B_n = competency of test taker *n*,

D_i = difficulty of task *i*,

R_j = severity of rater *j*,

C_l = difficulty of criterion *l*, and

F_{jk} = difficulty of receiving from rater *j* a rating of *k* relative to *k -1*.

In Equation 1, the logistic transformation of ratios of successive category probabilities ($\log (P_{nijlk} / P_{nijl(k-1)})$) is the dependent variable and the different facets (test takers, tasks, raters, criteria, etc) are the independent variables. That is, the model specifies that the likelihood of rater *j* giving a respondent *n* a score (*k*) instead of a lower score (*k-1*) on criterion *l* for task *i* will depend on the additive effects of the difficulty of the task (D_i), the severity of the rater (R_j), the respondent's level of performance (B_n), the difficulty of the criterion *l,* and the relative difficulty of scale category *k* when compared to category *k-1* across all tasks and criteria (F_{jk}). Since what we want to determine is whether the raters differ in the manner in which they apply the rubrics, this formulation of the MFRM model allows us to investigate whether that is the case. With the MFRM model, the parameters of each facet can be estimated independently of the rest of the facets and are calibrated jointly onto a single linear scale (i.e., the logit scale). For each element of each facet, the analysis provides a measure in logits, a standard

error of measurement (SE=the accuracy of the value estimated) and fit indices that describe the degree to which between the observed responses match those predicted by the model. Besides these statistics at the individual level, it is possible to obtain group statistics indicative of the average fit, the mean, the variability and reliability of the measures of the test takers, tasks, criteria and raters (Myford & Wolfe, 2004a).

The analyses with the MFRM model were run using the FACETS computer program (Linacre, 2009). The properties and resources of MFRM are those of the Rasch model: conjoint measurement, sufficient statistics for parameter estimation, interval measures, specific objectivity, estimation of accuracy of each measure and analysis of the fit of respondents, tasks, raters and assessment criteria to the model (Prieto & Delgado, 2003).

Basic statistics

Fit indices. These indicate the degree to which the observed scores differed from expected scores. An observed score is the one given by a rater to a test taker on one criterion for one task. An expected score is the one predicted by the model given the level of performance of the test taker, the severity of the rater, the difficulty of the criterion, and the difficulty of the task. The fit indices are the averages of the squared standardized residuals, called *Infit* and *Outfit*. A mean-square outfit value is the non-weighted mean of these squared standardized residuals, while a mean-square infit value is the information-weighted mean of the squared standardized residuals (Wolfe, 2009). Both statistics have an expected value of 1 and can vary between 0 and infinity. Values lower than 1 reveal that the residuals (the differences between observed and expected scores) are lower than expected randomly (that is, they can be interpreted as an *overfitting* the model). Values higher than 1 are those that show greater misfit than expected. Conventionally, values over 2 are considered to reveal a severe misfit that degrades the measurements (Linacre, 2009). FACETS yields individual fit values for the test takers, raters, tasks and assessment criteria.

Single rater-rest of the rater correlation ($R_{c,rc}$).This statistic quantifies the extent to which the assessments of each rater are consistent with the assessments of the other raters. Conventionally, values lower than .30 allow us to identify raters whose ratings are not consistent with the ratings of the rest of the raters (i.e. a rater's ordering of respondents by their levels of competency differs from the ordering of the rest of the raters).

The reliability of separation index (SR). Besides providing estimates of the accuracy of the measurements of individual elements of each facet (i.e., a standard error estimate for each test taker, each rater, each task, each

criterion), FACETS provides reliability assessments at the group level. SR is an index used to evaluate the reliability of the measures of the elements of the different facets (respondents, tasks, criteria or raters) and it reflects an estimate of the ratio of "true" score to observed score variance of the measures; the index can vary between 0 and 1. The substantive interpretations of SR differed among the facets (Myford & Wolfe, 2004a). In the case of the test takers, TSR (*Test taker separation reliability*) is comparable to coefficient alpha reported in studies of Classical Test Theory-based analyses of rating data, and it indicates the proportion of true variance with respect to the variance observed in the test takers assessed. In this case, high values of TSR are expected when the measures reliably reflect the variability of the persons in the variable. As regards the reliability of the measures of rater severity, the *Rater separation reliability* (RSR) statistic should not be interpreted as the degree of agreement among the raters, but as the degree to which they differ in severity. High values of the RSR indicate real differences among the raters (i.e., not attributable to measurement error). Since it is usually desirable not to have substantial variations among the raters as regards severity, a low values of the RSR (close to 0) is the goal.

Statistics of the categories used in scoring rubrics. To determine whether the categories included in a scoring rubric are empirically functional (ordered and distinguishable) several indicators are taken into consideration: the average of the respondent competency measures that went into calculation of each category calibration measure, outfit mean-square indices for each category, and the ordering of the steps between categories (Linacre, 2002). If the categories on a rubric are functioning properly, the averages of the competency measures for the respondents receiving a score in each of the successive categories must be ordered monotonically. This pattern of outcomes reveals that the higher the score received, the higher the level of the respondents in the latent variable (Park, 2004). The outfit mean-square values for the categories are also indicators of their functionality. For each assessment category, FACETS calculates the average of the competency measures for the respondents included in the category (observed measure) and an expected measure (i.e., the average test taker competency measure that the measurement model would predict for that category if the data were to fit the model perfectly). As indicated previously, if the observed value and the expected value are very similar, the outfit mean-square value will be close to 1.0. Outfit mean-square values greater than 2.0 indicate that the assessment category has not been adequately used. Finally, it can be observed whether the steps between categories are monotonically ordered and sufficiently separated. Disorder in

the steps indicates that categories exist that are not ones of most likely use in any range of the variable measured.

METHOD

Participants. Our sample comprised 943 test takers who completed the written expression part of the DELE exam for obtaining the Diploma of Spanish as a Foreign Language Level C1.

Instrument. The written expression part of the exam consists of two tasks. In the first one, after listening to a lecture, the respondent must write a 220- to 250-word composition expressing his or her point of view regarding the conflicting positions on the topic that were presented in the lecture. In the second task the respondents had to write a text of similar length choosing one of the following options: write a report on the functioning of a library you have been working in, or write a letter of complaint to a newspaper.

Procedure. A total of nine raters participated in the assessment of the respondent answers. The texts of each respondent were assessed independently by two raters, who gave their assessment on five criteria: an overall (holistic) score that assesses the test taker's performance as a whole, and four analytical scores, in which the rater analyzes different aspects separately giving a score to each criterion. These criteria are: *range* (balance between the lexical resources used and the topics and communication situations involved), *appropiacy* (adaptation of the text to the context), *accuracy* (knowledge and ability in using grammatical categories and the rules of morphology and syntax) and *coherence* (control of the resources necessary to establish relations between the discourse and the communication situation). Ratings were provided on a four-category rating scale (0,1,2,3). The direct score of a respondent on an exam is the sum of the values assigned to the criteria in the two tasks. The fact that not all the raters assess all the respondents poses a limitation for joint scaling of the differences in rater severity. For this reason, emphasis was placed on comparing the raters who assessed the same groups of respondents.

RESULTS

Figure 1 shows the "Wright Map like", a very useful resource for visualizing the joint measurement of the test takers' competency, the severity of the raters and the difficulty of the tasks, of the criteria and of the steps between assessment categories. Calibration of all the facet elements on the same interval scale (logit) allows the results to be interpreted in the same frame of reference. The *Respondent* column in Figure 1 contains the distribution of the respondents on the logit scale. Each asterisk (*) represents 11 individuals and each point, a lower frequency. Respondents with higher measures are located in the upper part of the column and those with the lower measures, in the lower part. Great variability in the competency of the test takers can be observed (between 4.41 and -4.42 logits). The rater severity values appear in the *Rater* column, with Rater 1 being the most severe (1.06) and Rater 3, the most lenient (-0.93 logits). The variability in rater severity is moderately high and greater than would be desirable. Ideally, raters should be observed to differ very little from each other in the levels of severity they exercise, as this would indicate that the criteria for assigning measures were used uniformly by the raters. The *Task* column shows the level of difficulty of the tasks comprising the exam. It can be seen that there is scarcely any difference in difficulty between the two tasks. The *Criterion* column depicts the relative difficulty values of the variables used to measure the tasks. It should be noted that, even though the differences in difficulty are small, the variables *accuracy* and *holistic* are more difficult than the rest. Finally, the *R.1-R.9* columns show in lines the situation of the values in logits of the *steps* between the categories used (0 to 3) by the raters to score the respondents' responses. It can be observed that the steps between the categories differ among the raters. This indicates that the scoring rubrics were not used in a uniform way. Below we discuss in more in detail the results from our investigation of rater functioning.

In Table 1 we present the raw score averages of the ratings that each rater assigned (on a scale of 0 to 3), the raters' severity measures (in logit), their measures of accuracy (standard error), their mean-square fit statistics and the correlation between the scores of each rater and the rest of the raters (inter-rater agreement). The raters differed a great deal in the levels of severity they exercised, which is undesirable. Ideally, the observed variations in severity should be negligible and attributable to standard error, and the RSR (*Rater separation reliability*) would show a low value. In this case, the RSR is very high (0.99); it is actually high enough to suggest that the observed differences in severity among the raters are very reliable. In fact, the accuracy of the severity estimation is high (i.e., standard errors of

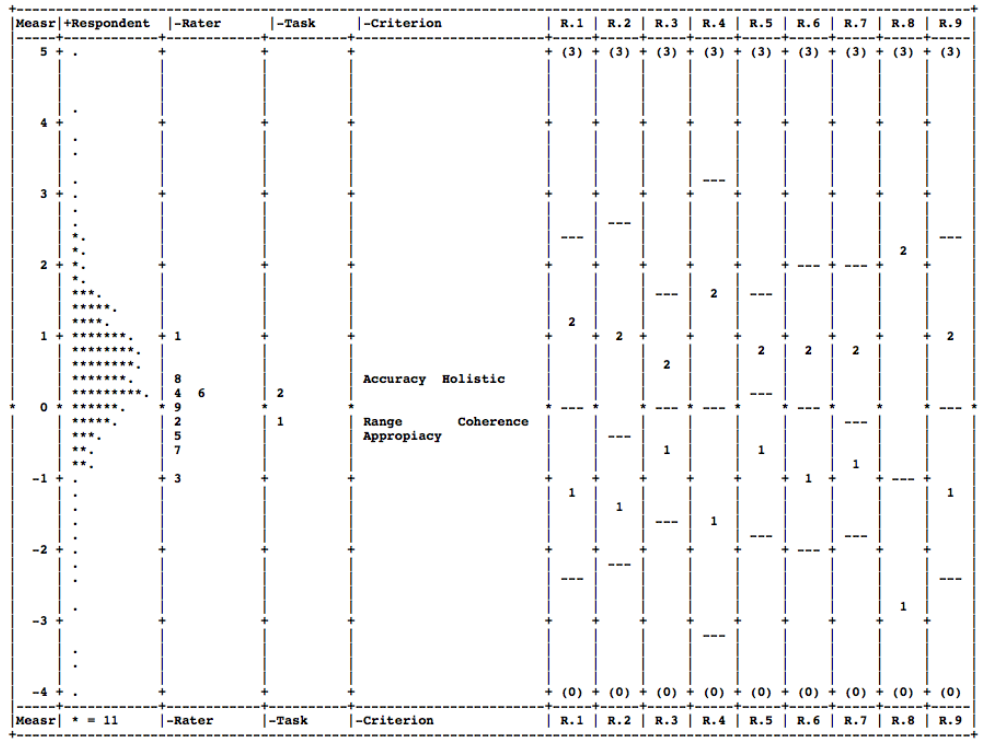

Figure 1. Map containing the measures of the facets analyzed. The horizontal dashed lines in the columns R.1-R.9 indicate the category threshold measures corresponding to the raters.

the rater severity measures range between 0.03 and 0.08). The mean-square fit statistics (which range between 0.60 and 1.64) indicate that all the raters show acceptable intra-rater consistency in their assessments. The correlations of each rater with the others ($R_{c,rc}$) ranged between 0.57 and 0.73, indicating an adequate agreement among the raters in the *ordering* of the test takers by their levels of competency. Nonetheless, it was observed that some of the raters differed substantially in severity. Thus, even though their assessments show an acceptable correlation with one of the other raters, the tendency to increase or decrease scores systematically can increase or decrease the likelihood that a test taker will pass the cut-off point.

Table 1. Rater values and statistics (N=9). Raters that share a letter assessed the same set of respondents.

Rater	Average Score	Severity	S.E.	Infit	Outfit	$R_{c,rc}$
1a (id: 5)	1.2	1 .06	0.03	0.64	0.65	0.73
2b (id: 9)	2.1	-0.11	0.03	0.74	0.77	0.69
3a (id: 20)	2.4	-0.93	0.03	0.65	0.60	0.69
4c (id: 146)	1.6	0.25	0.03	0.75	0.75	0.69
5b (id: 445)	2.3	-0.42	0.03	0.82	0.76	0.64
6c (id: 801)	1.6	0.30	0.03	0.73	0.74	0.70
7c (id: 847)	2.3	-0.52	0.07	0.72	0.73	0.65
8d (id: 9013)	1.5	0.45	0.08	1.64	1.55	0.57
9d (id: 9447)	1.9	-0.06	0.05	0.77	0.78	0.65

Average Score: Average Raw Non Linear Score
Severity: Severity Measure (in logits)
S.E.: Standard Error of Severity Measure (in logits)
Infit: Infit MnSq; Outfit: Outfit MnSq
RSR (Rater Separation Reliability) = .99.
Id: Rater ID Code

Since all the raters did not assess all the exams, it is advisable to interpret the differences in severity by comparing the assessments of raters who assessed the same respondents. Table 2 shows a detailed analysis of the raters who, when scoring the same set of test takers, showed the largest differences in severity. Rater 1 was the most severe (1.06 logits) and Rater 3, the most lenient (-0.93 logits). This difference is statistically significant. The averages of the direct scores of both raters differed by 1.20 points, a notable difference taking into account that the range of the scale was 0 to 3. Despite having assessed the same respondents, the two raters differed notably in the percentages of assignments to each category and in the values of the steps of the category characteristic curves (Table 2). Observe, for instance, that 50% of the scores given by Rater 1 were 1's, whereas 59% of the time Rater 3 gave scores of 3's. This aspect shows that the scoring rubrics were not used by these two raters in a uniform way. The step between categories 0 and 3 in Rater 1 (2.17) indicates that only the respondents with a value higher than this step in the variable have a high likelihood of receiving a 3. However, the respondent only needs more than 0.97 in the variable to have a high probability of receiving a score of 3 from Rater 3. Moderate and statistically significant difference in severity can also

be observed between Raters 7c and 4c (.77). In this case, Rater 7c was more lenient. There is a slighter difference between Raters 8d and 9d (.51).

Table 2. Statistics of the categories used by the raters with extreme values in severity.

Rater	Category	% of Ratings	Step
1a	0	18	--
	1	50	-2.18
(1.06)	2	29	0.01
	3	3	2.17
3a	0	3	--
	1	11	-1.04
(-0.93)	2	28	0.07
	3	59	0.97
7c	0	2	--
	1	10	-1.19
(-0.52)	2	44	- 0.46
	3	44	1.65
4c	0	2	--
	1	39	-3.17
(0.25)	2	53	0.03
	3	6	3.15
8d	0	3	--
	1	45	-4.20
(0.45)	2	52	- 1.21
	3	0	5.41
9d	0	2	--
	1	23	-2.03
(-0.06)	2	58	- 0.21
	3	18	2.23

Rater: Identification and severity value in parentheses.
% of Ratings: Percentage of ratings of each rater in each category.
Step: Value in logits between thresholds of successive categories.

DISCUSSION AND CONCLUSION

Although greater emphasis has been placed on the evaluation of differences in rater severity, also of interest in this paper is that we employed the MFRM model to measure the facet elements that affect the grading of the written expression tasks on the exam. We observed that the test takers attained a medium to high level of performance (0.55 logits on average) and high variability (SD= 0.96). The reliability of the estimations of the respondents was adequate (PSR = 0.84). Only 2.0% showed a severe misfit with the model. The tasks and criteria differed little in difficulty. Although two of the raters (Raters 1 and 3) scoring the exams showed a clear difference in their degree of severity, the rest of the raters showed minor differences. The differences in severity between Raters 4 and 7 and Raters 8 and 9 were moderate. According to the fit statistics, all the raters showed high intra-rater consistency in their assessments. Inter-rater agreement was acceptable, given that the correlations of each rater with the other of the raters ($R_{c,rc}$) showed sufficient consistency among the raters when ordering the respondents according to their competency. Even though the ordering of the test takers was similar, differences in severity can affect the likelihood of a respondent's surpassing the cut-off point for a passing grade. The difference between the raters with extreme values in their degree of severity may be the result of different ways of interpreting the scoring rubrics. Our results suggest that raters should be trained to use the scoring rubrics in a similar way.

The analysis we present here also demonstrates the utility of the Many Faceted Rasch Measurement Model to measure on a common scale the element parameters of the different facets involved in measuring performance on constructed response tests. This procedure contributes additional information to the methodology derived from Generalizability Theory (Cronbach et al., 1972), as it is not limited to the quantification of the different sources of error that affect observed scores. Being able to measure the elements of each facet using a common metric facilitates understanding of the different aspects that influence assessments and allows us to obtain measures of facet elements that are independent of the rest, and correct their idiosyncratic influence (Park, 2004).

RESUMEN

Análisis de la severidad de los calificadores de un examen de expresión escrita mediante el modelo *Many Faceted Rasch Measurement*. En este trabajo se describe cómo se puede aplicar el modelo Many Faceted Rasch Measurement (MFRM) para analizar la evaluación del rendimiento mediante calificadores. El manuscrito presenta una introducción al modelo MFRM, una descripción de los procedimientos de análisis y un ejemplo para ilustrar cómo se analizan los efectos de diversos factores en el rendimiento de los examinados en un test de expresión escrita. Los resultados ilustran la utilidad del modelo para detectar los calificadores que presentan valores extremos en el continuo de severidad. El modelo MFRM aporta puntuaciones en una métrica común de los diversos elementos de las facetas integradas en el proceso de medición (examinados, tareas, calificadores). Esta integración aporta ventajas para comprender el marco de la evaluación.

REFERENCES

Cronbach, L.J., Gleser, G.C., Nanda, H. and Rajaratnam, N. (1972). *The dependability of behavioural measurements*. New York: Wiley.

Eckes, T. (2011). *Introduction to Many-Facet Rasch Measurement. Analyzing and Evaluating Rater-Mediated Assessments*. Frankfurt am Main: Peter Lang.

Kondo-Brown, K. (2002). An analysis of rater bias with FACETS in measuring Japanese L2 writing performance. *Language Testing, 19*, 1-29. http://dx.doi.org/10.1191/02655322o2it218oa

Lane, S. and Stone, C.A. (2006). Performance Assessment. In R. L. Brennan (Ed.): *Educational Measurement* (pp 387-431). Wesport, CT: ACE/Praeger.

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

Linacre, J.M. (2002). Optimizing Rating Scale Category Effectiveness. *Journal of Applied Measurement, 3,* 85-106.

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

Linacre, J.M. and Wright, B.D. (2002). Construction of Measures from Many-Facet Data. *Journal of Applied Measurement, 3*, 484-509.

Myford, C.M. and Wolfe, E.W. (2004a). Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement: Part I. In E. V. Smith y R.M. Smith (Eds.*). Introduction to Rasch Measurement* (pp. 460-515). Maple Grove, MN: JAM Press.

Myford, C.M. and Wolfe, E.W. (2004b). Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement: Part II. En E. V. Smith y R.M. Smith (Eds.*). Introduction to Rasch Measurement* (pp. 518-574). Maple Grove, MN: JAM Press.

Park, T. (2004). An Investigation of an ESL Placement Test of Writing Using Many- facet Rasch Measurement, *Papers in TESOL & Applied Linguistics, 4*, 1-21.

Prieto, G. (2011). Evaluación de la ejecución mediante el modelo Many-Facet Rasch Measurement. *Psicothema, 23,* 233-238. [Performance assessment using Many-Facet Rasch Measurement].

Prieto, G. and Delgado, A.R. (2003). Análisis de un test mediante el modelo de Rasch. *Psicothema, 15,* 94-100. [Test analysis using Rasch model].

Tyndall, B. and Kenyon, D. M. (1996) Validation of a new holistic rating scale using Rasch multi- faceted analysis. En A. Cumming y R. Berwick (Eds.), *Validation in language testing* (pp. 39-57). Clevedon: Multilingual Matters.

Wolfe, E.W. (2009). Item and Rater Analysis of Constructed Response Items via the Multi-Faceted Rasch Model. *Journal of Applied Measurement, 10*, 335-347.

Wright, B.D. and Masters, G.N. (1982). *Rating scale analysis*. Chicago: MESA Press.

(Manuscript received: 22 February 2014; accepted: 19 May 2014)