RedFame

Journal of Education and Training Studies
Vol. 7, No. 1; January 2019
ISSN 2324-805X E-ISSN 2324-8068
Published by Redfame Publishing
URL: http://jets.redfame.com

Investigation of Interrater Reliability in The Evaluation of Foreign Language Writing Skills With Multigroup Confirmatory Factor Analysis

Emine Önen[1], Melike Kübra Taşdelen Yayvak[1]

[1]Gazi Education Faculty Gazi University, Ankara, Turkey

Correspondence: Emine Önen, Gazi Education Faculty Gazi University, Ankara, Turkey.

Received: July 9, 2018 Accepted: November 26, 2018 Online Published: December 4, 2018

doi:10.11114/jets.v7i1.3421 URL: https://doi.org/10.11114/jets.v7i1.3421

Abstract

In this study, it was aimed to examine the interrater reliability of the scoring of paragraph writing skills on foreign languages with the measurement invariance tests. The study group consists of 267 students studying English at the Preparatory School at Gazi University. In the study, where students write a paragraph on the same topic, the paragraphs are rated separately by three different interrater using the same scoring key. The evidence for the validity measurements was collected with AFA and DFA while the evidence for the reliability measurements was collected by the Cronbach-alpha (α) coefficient. As a result of testing with Multi-Group Confirmatory Factor Analysis within the context of the measurement invariance of the interrater reliability, no evidence of full and partial scalar invariance can be obtained while evidence of formal configural and metric invariance is obtained. As a result, the lack of evidence of scalar invariance means that raters scoring the writing skills do not use the same initial level of performance. In this case, the invariant uniqueness and invariant factor variances could not be tested, and therefore no evidence of reliability between raters could be obtained.

Keywords: interrater reliability, measurement invariance, evaluation of writing skills, multigroup confirmatory factor analysis

1. Introduction

The basic tool that the people need to communicate and express their feelings and thoughts is the language they use. Linguistic performance involves four primary skills: reading, writing, listening, and speaking. Writing skills are of great importance in teaching foreign languages in terms of ensuring correct communication for the individual to convey himself, his thoughts and feelings accurately and clearly to the reader. Measurement and evaluation in the process of teaching foreign language writing skills are crucial. Teaching is an important process involving the planning, implementation, and evaluation stages. From these stages, target-behaviors are tested separately in the measurement-evaluation stage, how much of the terminal behaviors are gained is checked, and quality control of the education is made. It is understood that the measurement and evaluation process benefits teachers, students as well as experts who are making decisions in education program, instructional success and students' learning (Mehrens & Lehman, 1991).

The scoring process can be more subjective since performance assessment methods are generally used in the evaluation of writing skills. In the course of measuring these skills, different kinds of mistakes that may interfere with the scores can lead to inappropriate evaluations of these skills of the individual. Among the sources of errors that may be involved in evaluating the writing skill include the status of the individual who is being rated for the skill, the rater and his scoring attitude, and the scoring key. One of the most important of these error sources is the rater and his attitude. Scoring can be more subjective because it may be influenced by the rater's observations, comments, and personal evaluations. The easiness or toughness of the rater in the scoring, the tendency of the rater to give an average score to performance, and the fact that the rater is scoring differently at different times are the most basic factor that leads to errors in the scores (Coffman, 1971). In this context, measuring errors arising from rater is an important factor affecting the reliability of the scores in particular. All of these aspects emphasize the importance of evaluating the same performance by different raters, that is to say, interrater reliability (Antonioni & Park, 2001; Attali, 2005).

Interrater reliability is defined as the different scoring of the same performance by different raters as a result of subjective judgments affecting the scoring stages (Antonioni & Park, 2001, Attali, 2005). Interrater reliability refers to the consistency between scores given by more than one rater. The differentiation of the scores given by the different raters to the same writing performance will cast a shadow on the correctness of the score for the writing performance. For this reason, it is desirable that the consistency between the scores of the raters to the writing performance, that is, the interrater reliability, is high. In addition, it is essential to evaluate the writing skills accurately in an effective foreign language teaching, since the feedback given to the individuals by evaluating the writing performance helps them to evaluate their writing skills and to see their mistakes.

Especially when evaluating interrater reliability in evaluating foreign writing performance, methods based on the Classical Test Theory (CTT), methods based on the G-theory approach, and methods based on the Multiple Facets Rasch Model (MFRM) approach have been recently used (Huang, 2008, 2011). When looking at the literature (Engelhard, 1994; Kondo-Brown, 2002; Eckes, 2005), it is frequently seen that the Rasch approach is used in studies examining the effect of the rater variable on the scores in performance assesment. Likewise, it is seen that the rater-derived variability in the scores is also examined with the generalizability theory (Barkaoui, 2007; Elorbany & Huang, 2012; Kondo-Brown, 2002; Stuhlmann et al., 1999).

However, there was no study of multi-group CFA in evaluating the writing skill of the interrater reliability in the related literature (Eckes, 2005; Engelhard, 1994; Kondo-Brwon, 2002; Lumley ve McNmara, 1995; Ross fisher, 2005; Weigle, 1998). For this reason, in the present study, the interrater reliability in evaluating foreign language writing performance is examined by multi-group CFA rather than structural equalization models. In this direction, a two-factor measurement model is defined based on the two dimensions of writing performance as "task achievement" and "use of language" and the performance criteria written about them, and the invariance between different raters of these parameter related to this model is tested with the measurement invariance tests.

Measurement Invariance

Measurement invariance is expressed as the mathematical equality of the corresponding measurement parameters for a particular factorial / elementally defined structure in two or more groups (Little, 1997). Kelcey, McGinn, and Hill (2014) refer to the measurement invariance as a condition in which the relationship between a latent variable and its indicators does not change depending on the group from which observations (scores) are obtained. In other words, it is the situation of whether or not the observations obtained by applying the same scale to different groups show a similar structure pattern among the groups. Measurement invariance is considered as a five-stage hypothesis testing process in the literature (Vanderberg and Lance, 2000). Each stage corresponds to a type of the measurement invariance:

1- Configural invariance

2- Metric invariance

3- Scalar invariance

4- Invariant uniqueness

5- Invariant factor variances (Vandenberg and Lance, 2000).

The first of the measurement invariance stages is the configural invariance. Configural invariance is the simplest and most basic type of measurement invariance. In this stage, the aim is to examine whether the items constituting the psychological measuring instrument exhibit the same appearance-pattern in relation to latent variables. In this direction, a hypothesis is tested that there is no difference between the groups in terms of the pattern of free and constant factor loads related to psychological measures. Providing evidence for the measurement invariance would mean that for this study, the raters use the same conceptual point of view in scoring the students' writing performances.

In the metric invariance stage, the factor loads (λ) of the indicators in the model are examined to see whether they are invariable between the raters (Vandenberg and Lance, 1998, 2000). Van de Vijver (1998) addresses the metric invariance as "equality of measuring units". At this stage, a hypothesis is tested that the factor loads (λ) for the indicators in the relevant measurement model are equal/invariable between the groups. Providing evidence of metric invariance will indicate in this study that the definitions of performance criteria (indicators) are similar / the same between raters and that the same measurement units between raters are provided (Salzberger et al., 1999; Vandenberg and Lance, 1998).

After evidence of metric invariance is obtained, the scalar invariance is tested. Scalar invariance means that the intercepts of the measures are equal/invariable between groups (raters in this study). In this study, obtaining evidence for scalar invariance means that score "0" in the scoring key for each of the scoring criteria in the writing skill scorings shows the same level of skill between the raters. Obtaining evidence for both metric and scalar invariance will show that the scores given by the different raters for the same writing performance have the same meaning and it indicates the same level of writing skills (Salzberger et al., 1999; Wicherts, 2007).

In the invariant uniqueness stage, equality/variability of error variances is tested in addition to the invariability of factor loads and intercept values between groups (raters). In the scope of this study, providing evidence of such an invariance would mean that even / equal amount of error will meddle in the scoring in each scoring criterion in the scoring key while raters score the same writing performance. Invariant factor variances, the last step of the measurement invariance, is based on the hypothesis that factor variances are equal/invariable between groups (raters) (Vanderberg and Lance, 2000). In this study, the evidence for this invariance indicates that raters are using equal ranges of the structure size in the scoring based on scoring criteria.

In this context, through the examination of interrater reliability via measurement invariance tests; it would be possible to obtain information about (a) whether the performance criteria definitions in the scoring key are perceived in the same way by the raters, (b) whether these performance metrics is even/equal to the level of representation of the relevant dimension of the writing skills, (c) whether the scaling units are even / equal to the raters, (d) whether the amount of error involved in the rater-derived scores is equal, and (e) whether the raters use the same / equal continuum as the score of the corresponding structure size. From this point of view, it is thought that this study would add a different dimension in examining interrater reliability in evaluating writing skills. In this respect, this research was carried out to examine the reliability between raters through the measurement invariance tests in the evaluation of foreign language skills of university preparatory students.

2. Method

2.1 Study Design

In this study, the interrater reliability was examined by measurement invariance tests in the scoring of writing skills in foreign language. It has been researched whether the psychometric properties of the scores obtained from the scoring of writing skills differ between the raters. This is a survey study as the existing situation being tried to be put forward as it is without any intervention.

2.2 Participant (Subject) Characteristics

The study did not go through the sampling process based on the definition of a population, but instead, a study group was used. The study group consists of volunteer students at level B1 who take English preparatory courses at Gazi University Foreign Languages School. The distribution of students in the study group by sex and faculties is shown in Table 1.

Table 1. Distribution of the students in the study group according to the gender and faculty

Faculty	**Male**		**Female**		**Total**	
	f	**%**	**f**	**%**	**f**	**%**
Engineering	94	64.38	57	47.11	151	56.55
Architecture	23	15.75	26	21.49	49	18.35
Science	10	6.85	13	10.74	23	8.62
Economics	11	7.53	13	10.74	24	8.99
Medical	8	5.48	12	9.92	20	7.49
Total	146	54.68	121	45.32	267	100

A total of 267 volunteer prep school students, 146 of whom are male, and 121 are female, participated in the study. In other words, 54.68% of this group is male students while 45.32% is female students. When we look at the distribution of students in the study group by their faculties, 56.55% of them are engineering faculty, 18.35% of them are architecture faculty, 8.62% of them are science faculty, 8.99% of them are the faculty of economics and administrative sciences, and 7,49% is a medical faculty student. Approximately half of the student groups participating in the study are engineering students because of the fact that the number of engineering students studying in the preparation unit is considerably higher than the other faculties.

2.3 Measures

The data needed in this study were obtained by applying the task of writing a paragraph given to the students in the study group within the framework of the subjects they have seen during an academic year under the English Preparatory Course at the Gazi University Foreign Language School. The writing task has been prepared in accordance with the level of students B1 according to the Common European Framework of Reference for Languages (CEFR) standards.

An analytical scoring rubric has been prepared for the process of scoring foreign paragraph writing skill in the investigation. There are ten performance criteria in the scoring key, four for "task achievement" dimension and six for "linguistic performance" dimension. Performance measures for task achievement dimension are writing topic sentence, writing supporting sentences, giving examples, and writing concluding sentence (Keh, 1990). Performance measures related to the dimension of linguistic performance are vocabulary, grammar, range, linking words, mechanics, such as punctuation and capitalization, and organization & content. All of the performance criteria in the scoring key are rated using a scoring scale ranging from 0-4.

2.4 Data Analysis

In the research, firstly, a two-factor measurement model was defined based on the two dimensions of writing performance as "task achievement" and "use of language", and the performance measures written about them. The fitness level of this model with the data sets obtained from each rater was then tested separately with CFA. In this process, it is necessary to obtain evidence that the sample meets the hypothetical normality assumption in order to decide the parameter estimation method. For this, the multivariable skewness-kurtosis z values and the χ2 and p (probability) values of these data are examined. Since the related data did not show the multivariate normal distribution, it was decided to use the Robust Maximum Likelihood method as the parameter estimation method (Brown, 2006). The two-factor measurement model (Model A) defined for writing skills is presented in Figure 1.

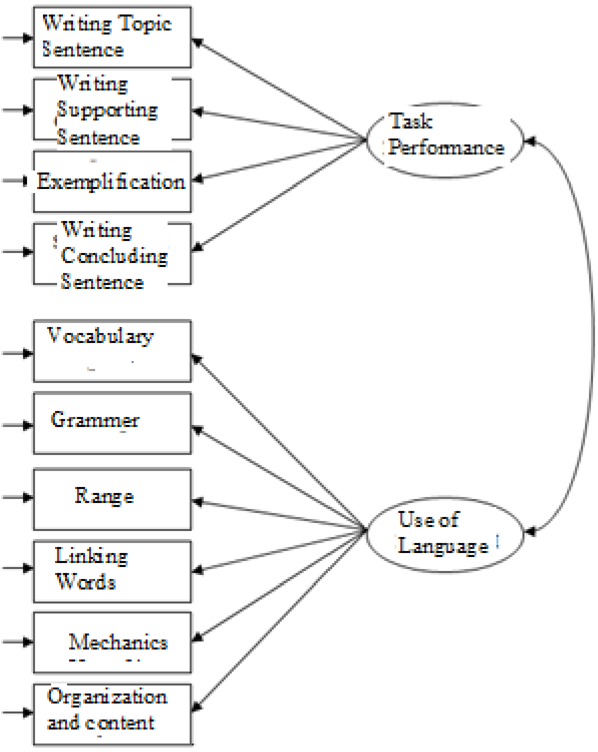

Figure 1. Model A: Base measurement model specified for writing in foreign language skill

After the fitness level of this model to the data sets obtained from each rater was tested separately, it has been tried to obtain evidence on the interrater reliability and the MGCFA for the measurement invariance. In the measurement invariance tests, first, the fitness level of the measurement model to the data is tested at each stage, and then the fitness level with the less limiting model in the stage one before is compared. Since Robust ML is used as a parameter estimation method in this study, Ts statistic is calculated by using scaled differences in chi-squares (SDCS) in comparing the fitness levels of the models. The fact that the value of Ts is non-significant indicates that there is no significant difference in the fitness levels between these models and that the measurement invariance at that level is provided (Brown, 2006).

Since there was no evidence of full scalar invariance in this study, the weaker form of the measurement invariance, partial scalar invariance, was tested. For this, the size of the fixed values for the performance measures for each rater was examined, and the partial scalar invariance was tested by removing the equality limitation for the fixed value with the greatest difference between the raters (Van de Schoot, Lugtig and Hox, 2012). However, although the fixed values of the three performance (MECH, WTS, WSS) measures were freely estimated between raters (by releasing the fixed value for each indicator at a time), there is still no evidence of partial scalar invariance. In this direction, the invariant uniqueness and invariant factor variances at later stages could not be tested.

3. Results

In the study, the fitness level of Model A with the data obtained from the three raters was tested separately. The goodness of fit indices calculated for Model A in this respect are presented in Table 2.

Table 2. Goodness of fit indexes for Model A

	CFI	TLI	RMSEA	SBχ^2(df)	p
Rater 1	0.97	0.97	0.10	129.60 (34)	0.000
Rater 2	0.97	0.97	0.09	119.57 (34)	0.000
Rater 3	0.97	0.97	0.11	145.28 (34)	0.000

Considering the CFI values calculated for the data obtained from each rater, it is thought that the model has a good fit with the data. However, for any of the rater, the RMSEA values indicate that the model does not fit adequately into the database. In this direction, the model was re-specified considering Modification Indices (MI) for the model. The modification indices indicated that the organization & content indicator is related to the task achievement dimension. This indicator indicates how well the learner writes in the given subject in unity by using the relevant vocabulary and grammatical structures. Paragraph unity is related to the basic elements of the paragraph, writing the topic sentence (WTS), writing the supporting sentence (WSS), writing example sentences (WES), and writing concluding sentence (WCS). In this context, it is understood that the raters scored this indicator as a sign of the task achievement of writing a paragraph. Model B, re-specified by making the necessary modifications in this direction, is presented in Figure 2.

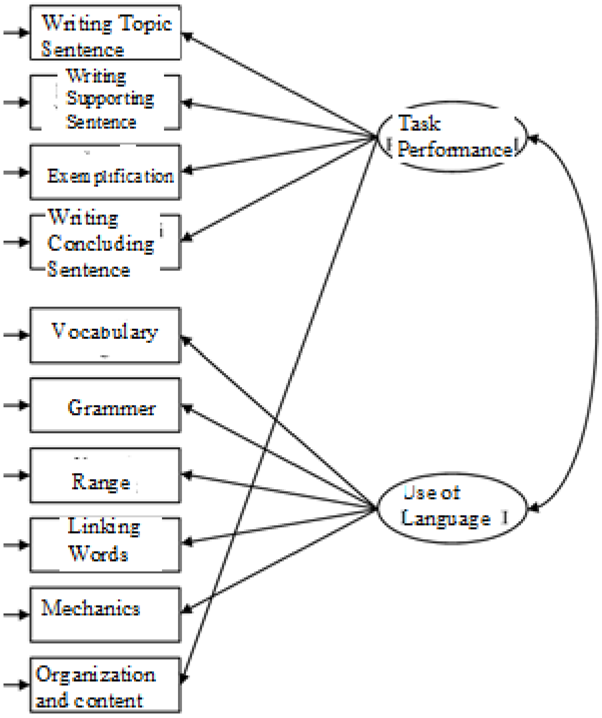

Figure 2. Model B Respecified model for specified for writing in foreign language skill

The calculated goodness of fit indexes for Model B are presented in Table 3.

Table 3. Goodness of fit indexes for Model B

	CFI	TLI	RMSEA	SBχ^2(df)	p
Rater 1	0.99	0.99	0.06	69.89 (34)	0.000
Rater 2	1.00	1.00	0.04	48.60 (34)	0.005
Rater 3	0.99	0.99	0.06	76.96 (34)	0.000

The fitness indices presented in Table 3 indicate that Model B has a good fit with the data obtained from the three raters. When the factor load values for Model B are examined, the values for the first rater are in the range of λ = 0.47 and λ = 0.89; the values for the second rater are calculated at λ = 0.49 and λ = 0.91; and the values for the third rater are in the range between λ = 0.46 and λ = 0.92. When the calculated error variances for the indicators in Model B are considered, the error variances in the values of the first rater are in the range of ε = 0.20 and ε = 0.78; the error variances of the second rater are in the range of ε = 0.21 and ε = 0.77; and the error variances for the third rater are in the range of ε = 0.19 and ε = 0.79. All these findings indicate that Model B has a good fit for individual data sets from each rater, and that the indicators are good representatives of the relevant structure dimension (Kline, 1998). Then Multi-Group Confirmatory Factor Analysis was applied to test the measurement invariance between the raters. The calculated goodness of fit indices is presented in Table 4.

Table 4. Fit indices for meausrement invariance tests

	CFI	TLI	RMSEA	SBχ^2(df)	p
Configural invariance	0.98	0.978	0.06	213.36 (102)	0.000
Metric invariance	0.98	0.98	0.05	224.47 (122)	0.000
Scalar invariance	0.96	0.96	0.07	352.17 (142)	0.000
Partial scalar I. 1	0.97	0.97	0.07	303.60 (140)	0.000
Partial scalar I. 2	0.97	0.97	0.06	287.83 (138)	0.000
Partial scalar I. 3	0.97	0.97	0.06	273.27 (136)	0.000
Partial scalar I. 4	0.97	0.97	0.06	256.65 (134)	0.000

When Table 4 is examined, it is understood that the fitness statistics calculated as a result of testing the configural invariance provide evidence for the configural invariance. This finding means that the three raters score the paragraph writing skills in the foreign language based on the same dimensions of the skill, and the factor structure does not change between the raters. The goodness of fit values obtained from the applied metric invariance test to determine whether the raters use the same unit of measure while scoring the writing performance dimensions indicates a good model fit. Metric invariance can be reached based on the Ts statistic (Ts = 9.86, df = 20) calculated to compare the fitness level of the model tested at the metric invariance stage with the fitness level of the model tested at the configural invariance stage. This means that raters use even / equal units of measurement while scoring writing skills.

The score of the measurement invariance test, which is used to understand whether the raters are using the same initial level of performance while scoring performance scores of students, is pointing to a good model fit. However, the calculated Ts statistic (Ts= 130.45, df = 20) shows that full scalar invariance cannot be achieved. This means that for some performance measures in the scoring key, the fixed values are non-invariant among the raters. In this direction, the fixed values of the indicators calculated for the three raters were compared one by one, and the indicator that the most difference between the three raters was determined. When the fixed values of the Mechanism indicator are examined; the fixed value of the first scorer is τ = 3.06, the fixed value of the second scorer is τ = 2.76, and the fixed value of the third scorer is τ = 3.19. It has been determined that the most differentiation is this indicator. Partial scalar invariance (1) was tested by releasing this parameter of this indicator. Then, the fitness level of the scalar invariance model and the fitness level of the partial scalar invariance model were compared. Based on the calculated Ts (Ts= 48,48; df = 2) statistic, the result of the partial scalar invariance cannot be achieved.

Then the fixed value for the writing topic sentence, which its fixed value differs the second most between the raters, was also released and the partial scalar invariance was tested for the second time. The calculated Ts (Ts = 15.79; df = 2) statistic indicates that partial scalar invariance cannot be achieved at this stage too. Subsequently, the fixed value for the "writing supporting sentence" indicator was freely estimated, and the partial scalar invariance was tested for the third time. Again, the calculated Ts (Ts= 15.51; df = 2) indicates that partial scalar invariance cannot be achieved at this stage. Finally, partial scalar invariance was tested for the last time by releasing the fixed parameter for the "writing example sentence" indicator. The Ts (Ts = 15.83; df = 2) statistic calculated at this stage also indicated that partial scalar invariance was not achieved. As a result of these analyzes, no evidence of partial scalar invariance can be obtained. This means that the raters do not use the same starting level of performance while scoring the writing skills. Invariant uniqueness and invariant factor variances could not be tested because no evidence of partial scalar invariance was found.

4. Discussion

There is no evidence of scalar invariance in this study, where interrater reliability is examined through measurement invariance tests while providing evidence of configural and metric invariance. Evidence of configural invariance indicates that the raters scored with a similar conceptual point of view in scoring the writing skill, in other words, this skill has a similar meaning to all raters. Providing evidence of metric invariance means that the three raters score the writing skill using the same unit of measure. This indicates that a change in a unit in writing skills -among the raters- leads to a statistically even/equivalent change in terms of points scored by the scoring key. Since there is no evidence of full scalar invariance in this study, partial scale invariance is tested by removing equality limits for the four parameters in the model. However, this also doesn’t provide evidence for partial scale invariance, nor is there a sufficient number of indefinite indications for each factor, which is why the scalar invariance is not achieved. Failure to provide evidence for scalar invariance implies that the origins of the variables observed in the measurement model defined for writing skill differ between raters. This situation indicates that if the individual's writing performance is scored with this scoring key, he will get different scores from different raters meaning different writing skill scores. In this respect, in terms of the dimensions of the individual's writing skill in this study, when the situation is assessed by different raters using this scoring key, it will lead to different deductions about the individual's writing skills. When this scoring key is used in this direction, it is possible for individuals to make inaccurate evaluations of their writing skills and take incorrect decisions. Based on the findings from this study, it is concluded that there is no evidence of interrater reliability when writing skills are assessed using this scoring key (Salzberger et al., 1999, Vandenberg and Lance, 2000; Wicherts, 2007).

In the study, it is concluded that the same characteristics are scored differently among the raters. Similar to this result, Aktaş (2013), Atmaz (2009); Barkoui (2007); Eckes (2005); Kondo-Brown (2002); Korenovska (2013); Matsuno (2009); Sudweeks, Reeve & Bradshaw (2004) points out that the scores differ originating from the raters in studies where the interrater reliability are examined with the approaches such as MFRM and Generalizability. However, it differs in some ways from the point of the information obtained from these studies and the information obtained by the measurement invariance tests.

In these studies, where the interrater reliability was examined, MFRM was used when the effect of toughness and easiness of the rater on the scores and the effect of other variability sources on the scores were examined; Generalizability theory can be used when it is desired to generalize all raters in the population with the scores in the sample. Apart from these studies, the measurement invariance tests provide information on whether or not raters use the same conceptual framework, score the performance criteria in the same way and understand these criteria in the same way, and whether they make the same amount of error in scoring. Thus, whether the scores between the raters are consistent can be determined, and if it is not consistent, this inconsistency about the scoring key can be clarified more clearly. In this way, necessary regulations can be made, and more reliable scorecards can be reached.

There was no evidence of interrater reliability for scores obtained from the analytical scoring rubric examined in this study. In this respect, it is clear that using this scoring key cannot reliably measure the writing performance of the individual.

In order to correct this situation, the rater can be instructed about the scoring key before the scoring is done, and it can be examined how this situation will change the findings at hand. Further, evidence for scalar invariance can be reworked. Another recommendation that can be presented in the literature of the relevant field except for the suggestions based on the research results is to examine the interrater reliability with the Multiple Indicators Multiple Causes (MIMIC) model, where the rater is considered as a latent variable. In the MIMIC model, detailed information on how well the different rater estimate latent variables can be provided because of the path coefficients indicating the direct effect of the rater variable. In this way, the effect of the rater on the indicator, intercept (τ) and factor load (λ) values can be examined and compared with the results obtained with the MGCFA.

References

Aktaş, M. (2013). Aynı performans görevinin farklı sayıda puanlayıcılar tarafından üç farklı teknikle puanlanmasından elde edilen puanların güvenirliklerinin genellenebilirlik kuramına göre incelenmesi. (Unpublished Master thesis). Mersin University, Educational Sciences Institues, Mersin, Türkiye.

Antonioni, D., & Park, H. (2001). The relationship between rater affect and three sources of 360-degree feedback ratings. *Journal of Management, 27*, 479–495. https://doi.org/10.1177/014920630102700405

Atmaz, G. (2009). Puanlama yönergesi (rubrik) kullanılması durumunda puanlayıcı güvenirliğinin incelenmesi. (Yüksek lisans tezi). Unpublished Master thesis). Mersin University, Educational Sciences Institues, Mersin, Türkiye.

Attali, Y. (2015). Reliability-based feature weighting for automated essay scoring. *Applied Psychological Measurement. 39*(4). https://doi.org/10.1177/0146621614561630

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed method study. *Assessing Writing, 12*(2), 86-107. https://doi.org/10.1016/j.asw.2007.07.001

Brown, T. A. (2006). *Confirmatory factor analsis for applied research*. New York: The Guilford Press.

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. https://doi.org/10.1207/s15434311laq0203_2

Elorbany, R., & Huang, J. (2012). Examining the impact of rater educational background on ESL writing assessment: A generalizability theory approach. Language and Communication Quarterly, 1, 2-24. Retrieved from http://www.untestedideas.com/lcq.html.

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement, 31*(2), 93-112. https://doi.org/10.1111/j.1745-3984.1994.tb00436.x

Huang, J. (2008). How accurate are ESL students' holistic writing scores on large-scale assessments? A generalizability theory approach. *Assessing Writing, 13*(3), 201-218. https://doi.org/10.1016/j.asw.2008.10.002

Huang, J. (2011). Generalizability theory as evidence of concerns about fairness in large-scale ESL writing assessments. *TESOL Journal, 2*(4), 423-443. https://doi.org/10.5054/tj.2011.269751

Keh, C. L. (1990). Feedback in the writing process: a model and methods for implementation. *ELT Journal, 44*(4). https://doi.org/10.1093/elt/44.4.294

Kelcey, B., McGinn, D. & Hill, H. (2014). Approximate measurement invariance in cross-classified rater-mediated assessments. *Frontiers in Psychology*, 5. https://doi.org/10.3389/fpsyg.2014.01469

Kline, R. B. (1998). *Principals and Practice of Structural Equation Modeling*. New York. The Guilford Press.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing, 19*(1), 3–31. https://doi.org/10.1191/0265532202lt218oa

Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behaviorial Research, 32*(1), 53-76. https://doi.org/10.1207/s15327906mbr3201_3

Lumley, T., & McNamara, T. F. (1995). *Rater characteristics and rater bias: implications for training*. Paper presented in Language Testing Research Colloquium, Cambridge. https://doi.org/10.1177/026553229501200104

Matsuno, S. (2009). Self-, peer-, and teacher-assessments in Japanese university EFL writing classrooms. *Language Testing, 26*(1). https://doi.org/ 10.1177/0265532208097337

Mehrens, W. A., & Lehmann, I. J. (1991). *Measurement and evaluation in education and psychology.* New York: Harcourt Brace College Publishers.

Ross, F. R. L. (2005). Developing effective success rubrics. *Kappa Delta Pi, 41*(3), 131-135. https://doi.org/10.1080/00228958.2005.10518823

Salzberger, T., Sinkovics, R. R., & Schlgelmich, B. B. (1999). Data Equivalence in Cross-Cultural Research: A Comparison of Classical Test Theory and Latent Trait Theory Based Approaches. *Australasian Marketing Journal, 7* (2), 23-38. https://doi.org/10.1016/S1441-3582(99)70213-2

Stuhlmann, J., Daniel, C., Dellinger, A., Denny, R. K., & Powers, T. (1999). A generalizability study of the effects of training on teachers' abilities to rate children's writing using a rubric. *Journal of Reading Psychology, 20*, 107-127. https://doi.org/10.1080/027027199278439

Van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Pschology, 9* (4), 486-492. https://doi.org/10.1080/17405629.2012.686740

Van de Vijver, F. J. R. (1998). Towards A Theory of Bias and Equivalence.*ZUMA-Nachrichten Spezial*, 41-65. Retrieved from Web:http://www.gesis.org/Publikationen/Zeitschriften/ZUMA_ Nachrichten_spezial.

Vandenberg, R. J., & Lance, C. E. (1998). A summary of the issues underlying measurement equivalence and their implications for interpreting group differences. Research Methods Forum. Retrieved from Web: http://www.aom.pace.edu/rmd/1998_forum_equiv_group_differences .html

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. *Organizational Research Methods 3*, 4–70. https://doi.org/10.1177/109442810031002

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2). https://doi.org/10.1177/026553229801500205

Wicherts, J. M. (2007). Group Differences in Intelligence Test Performance. Unpublished dissertation, University of Amsterdam. (Universiteit van Amsterdam). Retrieved from Web:
http://www.repository.naturalis.nl/document/44999

Copyrights