Validation of Writing Scales for Turkish as a Second Language through Many-Facet Rasch Measurement

Fatma Küçük Üçpınar and Aylin Ünaldı

Abstract
Rating scales and the extent to which raters use them effectively are two important factors that influence scoring validity of language tests when open-ended writing tasks are concerned. Research regarding rating scale development and validation in the assessment of English is ample; however, there has been no research on scale validation in the assessment of Turkish as a Second Language (TSL) to this date. This study reports on the development of two analytical rating scales used to assess academic writing skills of test takers in TSL, and presents quantitative evidence on the rating scale validation. For this purpose, texts written by 39 TSL students were scored by three raters. The analyses were conducted using Many-facet Rasch Measurement. Results indicate that empirically-developed analytical rating scales were used consistently and appropriately by the raters providing evidence for the reliability and effectiveness of the rating scales.

Key Words: Scale development, writing assessment of Turkish as a second language, many-facet Rasch measurement

Literature Review

Establishing the scoring validity of language tests is claimed to be critical in test validation as tasks that are valid in terms of cognitive and contextual parameters are of little value if the marking of exam scripts is not reliable (Shaw & Weir, 2007). Therefore, Shaw and Weir (2007) describe scoring validity as the superordinate term encompassing all the aspects of the testing procedures that are likely to influence the reliability of test scores. In high-stakes tests such as language tests taken by students before pursuing academic studies in second language (L2)-medium universities it is even more critical to ensure that test score interpretations are meaningful and appropriate as test results are used to take important decisions about students' future. (Mendoza & Knoch, 2018). When students are evaluated through open-ended writing tasks, obtaining valid scores that sufficiently reflect students' writing ability is a major concern (East, 2009). Rating scale efficacy and the extent to which raters can apply the scale accurately to score test takers' responses are two important issues to be taken care of in order to achieve scoring validity.

In this study, we aim to present evidence for scoring validity of a newly developed Academic Writing Test of Turkish as a Second Language (TSL) by describing the development and validation of two analytical rating scales used in its scoring. As Mendoza and Knoch (2018) argue most language assessment studies including rating scale development, rating procedures and rater effects have been published in the assessment of English whereas very few studies have been conducted in

Fatma K. Üçpınar, Fatih Sultan Mehmet Vakıf University, School of Foreign Languages, ftmkk58@gmail.com
Aylin Ünaldı, Boğaziçi University, Department of Foreign Language Education, aunaldi@boun.edu.tr

the assessment of other languages. In this respect, this study is of importance as it reports on the assessment of academic writing of TSL for the first time; thus, it may encourage further studies on this line of research

Literature Review

Rating Scale Development

Davies et al. (1999) define rating scale as "[an assessment tool] consisting of a series of constructed levels against which a language learner's performance is judged" (p. 153). The levels usually range between no mastery to full mastery of specific language skills. Each of the levels identified in the scale typically consists of verbal descriptions so that raters can determine which specific levels the written performances correspond to. Weigle (2002) argues that there are several decisions that should be made in rating scale development, and these considerations have been studied in the literature extensively.

What Type of Rating Scale Is Desired?

According to Weigle (2002), the first decision to be made in rating scale development pertains to the type of the rating scale to be used. Two main types of rating scales are typically distinguished in the literature: holistic and analytical (Jonsson & Svingby, 2007). These major types of scales basically have to do with whether a single score or multiple scores will be assigned to each writing scripts (Weigle, 2002). Holistic scoring involves "the assigning of a single score to a script based on the overall impression of the script" (Weigle, 2002, p.112). The major idea behind this approach is that writing is treated as a single entity in real life; therefore assigning a single score is the best way to capture the integrated qualities of writing (Knoch, 2009). Although holistic scoring has been widely preferred in assessment of writing because it is seen as time-saving and authentic, it has also been criticized (Knoch, 2009). First, it is argued that a single score is unlikely to provide diagnostic information about candidates' ability, which is a problem especially in L2 assessment. A single score will not help to identify aspects of writing which tend to develop at different rates in L2 learners. Furthermore, when a single score is assigned for the performances of these candidates, it is difficult to identify how raters arrive at their decisions about candidates' abilities. Sakyi (2000), for example, investigated the raters' decision-making processes when evaluating writing scripts with holistic scoring procedures through verbal protocols. Sakyi (2000) identified four distinct rating styles among six raters, and it was found that some raters did not use the scoring criteria at all, instead relied on personal judgments.. While some raters focused on errors, others focused on the development of ideas; or when they attempted to use scoring criteria, they were able to consider only one or two features to distinguish between levels of ability. This finding underlines the fact that raters may use various criteria including their personal judgments in holistic scoring and this may lead to inconsistent ratings due to construct irrelevant variance adversely affecting scoring validity.

One way of overcoming possible limitations of holistic scoring is the use of analytic scoring. Analytic scoring involves assigning separate scores for various traits of writing such as content, organization, register, vocabulary, grammar; and these traits might differ based on the purpose of the assessment (Weigle, 2002). Analytic scoring provides more detailed information about candidates' writing abilities compared to holistic scoring. This aspect of analytic scoring is of value especially in diagnostic tests, in which detailed analysis of candidates' performance is needed for the purpose of giving feedback. Moreover, through analytic scoring, it is possible to capture uneven aspects of L2 learners' writing performances. Citing the works of Adams (1981) and Francis (1977), Shaw and Weir (2007) maintain that differentiation between multiple components might be particularly useful for training inexperienced raters, as analytic scoring allows raters to focus on one aspect at a time, and it is easier than assigning an impressionistic score. On the other hand, analytic scoring is not without disadvantages. A major problem has to do with practicality. Because more than one decision needs to be made during analytic rating procedures, it takes longer time than holistic scoring. Secondly, as Knoch (2009) maintains, there is no guarantee that raters will distinguish between multiple traits of the scoring rubric. Crucially, as discussed in Myford and Wolfe (2003), it is very common that rating of one aspect may affect ratings of other aspects, creating a halo effect.

Barkaoui (2007, 2010, 2011) investigated the role of the rating scale type (holistic and analytic) on rating processes and score variability. The findings showed that the raters using holistic scoring tended to go back to the text itself to make their decisions while raters using analytic scoring referred to the scoring rubric more frequently; in other words, raters using analytic scoring were more attentive to the evaluative criteria in rating scale. Moreover, it was found out that raters were likely to be less severe with analytic marking. When it comes to reliability, holistic scoring resulted in higher inter-rater reliability whereas analytic scoring led to higher self-consistency.

Studies on the type of rating scale suggest that both holistic and analytic scoring methods have their own strengths and limitations. The testing purpose and contextual variables should determine the choice of the rating scale.

What Are the Criteria Based on?

When the type of rating scale to be used is decided on, the next consideration will be how to design the scale itself (Weigle, 2002). The ways in which rating criteria are constructed are of great importance because the wording of the scale is considered to represent test developers' view of writing construct to be tested. Weigle (2002) discusses two methods of constructing rating scale: a priori method and empirically-based methods. Knoch (2007) suggests that a priori method is carried out through intuitive judgments of experts about the nature of language development in order to construct scale descriptors. Such intuitive methods typically involve "…develop[ing] a rating scale based on pre-existing scales, teaching syllabus or a needs analysis" (Knoch, 2009, p. 43). This is generally done by a group of experienced teachers or language testers taking the role of experts. Although it is argued that most rating scales are constructed intuitively, such type of scales has been criticized by several researchers (e.g., Fulcher, 2003; Knoch, 2007; Turner & Upshur, 2002). Turner and Upshur (2002) summarize the criticisms that rating scales constructed through a priori method have received: a) the ordering of the criteria do not generally reflect the findings of second language acquisition (SLA) research; b) the criteria are mostly irrelevant to the characteristics of task response and the context; c) the criteria are not grouped at relevant descriptor levels properly; d) the wording of scale descriptors are often ambiguous, and this causes raters to interpret the scale in a different way. These concerns raised against intuitively designed scales seem to be valuable especially for reliability of test scores and validation of rating scales.

One means of responding to the criticisms of intuition-based rating scales is to construct scale descriptors empirically, which involves examination of students' actual written responses and defining the characteristics that differentiate responses and the levels of rating scales (Knoch, 2007). With this method of scale development, it is claimed that the descriptors are more likely to reflect the features of test takers' performances at different proficiency levels; and describing real features of writing at each level may solve the problem of relative wording used in scale descriptors. As a result, raters are expected to apply the rating scale more consistently and efficiently. What is more, empirically-developed descriptors are believed to reflect the natural order of writing acquisition, and all these features are claimed to increase score reliability as raters are able to base their decisions on more explicit and realistic evidence (Knoch, 2007). Despite its promises, empirically-based rating scales have also been criticized. Fulcher, Davidson and Kemp (2011), for example, maintain that as empirically-developed rating scales involve descriptions of performance in specific genres or contexts, they may not be applicable to rate performances in other contexts. This in turn may affect the generalizability of inferences that are made based on ratings. Knoch (2007) conducted a study with ten trained raters to investigate whether an empirically developed rating scale functioned differently from an intuition-based analytic rating scale. The results indicated that the individual components of the empirically developed scale (pilot scale) were more discriminating than the intuition-based scale (existing scale). Moreover, the pilot scale resulted in higher inter-rater reliability whereas the raters tended to differ more in terms of severity when they used the existing rating scale. The researcher concluded that analytic rating scales may not necessarily function in an analytic manner, if scale categories are not described explicitly and detailed enough.

How Many Points or Scoring Levels Will Be Used?

Another important consideration with regard to scale construction is to decide on the number of scoring points to be used to distinguish between different ability levels. It is claimed that the number of distinctions that raters can make is limited (Weigle, 2002). Myford (2002) claims that there has not been a consensus on the optimal number of scale points although it seems that a scale with points ranging between 4 and 9 is ideal for raters to be able to discriminate between test takers' various proficiency levels, and not to be overwhelmed with too many scale points, at the same time.

Besides the number of scale points, it is also important to consider the number of categories if analytic scores are used. It is suggested in the Common European Framework of Reference for Languages (CEFR) that "more than 4 or 5 categories start to cause cognitive overload" and that "7 categories is psychologically an upper limit" (Council of Europe, 2001, p.123). Therefore, even though many aspects of a specific skill are defined, it may be more reasonable to select only the ones that are important for the purpose and the context of the test (Weigle, 2002).

Raters

Just as decisions regarding the rating scales should be made carefully in order to prevent potential source of error that may threaten score validity, raters as another source of error should also be carefully monitored as Popham (1990 as cited in Myford & Wolfe, 2003) argued. A number of studies have shown considerable rater effects as a source of systematic variance in the ratings of written performance (i.e., Hoyt, 2000; Lumley & McNamara, 1995; Myford & Wolfe, 2003). This kind of variability is principally unwanted as "[it] is associated with characteristics of raters and not with the performance of examinees" (Eckes, 2008, p. 156). Major rater effects which are considered to be sources of systematic error variance are identified as severity/leniency, halo, central tendency, inconsistency/randomness, and bias (Knoch, 2009). These rater effects have raised concerns about validity of ratings and, thus, have been the focus of researchers (Eckes, 2005; Lumley & McNamara, 1995; Myford & Wolf, 2003; Wolfe, 2004). To illustrate, Engelhard (1994) worked with 15 raters to investigate rater effects on the quality of ratings by using many-faceted Rasch measurement model. The data revealed significant differences in rater severity. Besides, two raters rated the compositions holistically rather than analytically, and this was seen as the evidence for halo effect. Moreover, nearly 80% of ratings were in the two middle categories of the rating scale, displaying the presence of central tendency effect. Similarly, Eckes (2005) conducted a study to investigate rater severity and bias effects towards examinees, the efficiency of the rating criteria and the tasks in writing and speaking sections of the test of German as a Foreign Language through many-faceted Rasch measurement. The results revealed substantial variability in raters' level of severity. Although the raters were consistent in their overall ratings, they were significantly less consistent in relation to criteria and tasks (for speaking test) than in relation to examinees. In other words, the raters were biased towards certain criteria and tasks, which led them to display more severity or leniency with them.

Another line of research has investigated decision making behavior of raters with different personal background, rating background and work experience (Knoch, 2009). Cumming (1990) examined the decision making processes of expert and novice raters and found out that expert raters used a wide range of criteria, self-control strategies and knowledge sources while reading and judging student compositions whereas novice raters tended to use much fewer of these criteria and skills probably derived from their general reading strategies, and they relied on online corrections of student texts to make their judgments. In a similar vein, Wolfe, Kao and Ranney (1998) investigated cognitive differences of proficient and non-proficient raters. The results indicated that proficient raters tended to use a top-down approach through which they focused on general features of texts and made an overall judgment of writing quality. Less proficient scorers, on the other hand, seemed to use a bottom-up approach focusing on more specific features of the essay and interrupting their reading process to see if the text so far satisfies the scoring rubric.

With regard to rater occupation, O'Loughlin (1992, as cited in Shaw & Weir, 2007) compared the rating behavior of teachers from different subject areas who rated essays produced by native-speaker students and EFL students. Findings showed that language teachers did not pay as much attention to content as teachers of other academic subjects and EFL teachers were more attentive to grammar and cohesion than mainstream English teachers. Similarly, Weigle, Boldt and Valsecchi (2003) examined how ESL, English and other content area instructors perceive and evaluate ESL student writing. They concluded that raters from different disciplines bring their own expectations of what constitutes a good writing based on the conventions of their discourse community, which consequently influence their way of using assessment criteria. For example, instructors of English departments were more concerned with grammar than other raters whereas psychology department raters devoted their primary focus for content.

Weigle (1999) suggests that rater expectations are another factor that may influence test scores. In a study investigating rater-prompt interactions, Hamp-Lyons and Mathias (1994) found that raters awarded higher scores to the performances in response to the tasks that were judged as difficult by the experts. Hamp-Lyons and Mathias suggested that this unexpected finding might have resulted from the compensatory strategies employed by the raters in order to negate the effect of prompt difficulty and reward students who went for the difficult tasks. Eckes (2012) examined the relationship between raters' perception of criterion importance and their rating behavior by conducting bias analysis with multi-faceted Rasch measurement. He found that the criteria that were perceived as important received more severe ratings than the ones considered as less important.

The studies mentioned above suggest that raters might vary in terms of their decision making processes that seem to vary based on their personal background, rating experience, professional training, and expectations. Research has shown that differences in rating behavior may lead to considerable variability in scores that are not related with examinees' performance, thus threaten score validity. In an attempt to eliminate rater effects such as severity, leniency, halo, central tendency and bias, it is now obvious that one needs to construct detailed scoring criteria with unambiguous and explicit descriptors. However, a well-constructed scoring rubric may not be sufficient by itself to eradicate errors associated with rater characteristics.

Rater Training

In an attempt to minimize variability of raters and improve the reliability of rating process, rater training is crucial. Jacob et al. (1984, as cited in Weigle, 1994) argue that training aims to "ensure more consistent interpretation and application of the criteria and standards for determining communicative effectiveness of writers" (p. 43). Shaw and Weir (2007) suggest that no mark scheme can capture the definition of a level in a way that raters could apply consistently unless each level is exemplified with benchmark scripts during rater training. Research has shown the effectiveness of rater training. Weigle (1994) investigated the effect of training on inexperienced raters of ESL compositions based on verbal protocols. The findings revealed that training helped clarify the comprehension of intended rating criteria and modify raters' expectations in terms of writer characteristics and task demands. Weigle (1998) compared the ratings of experienced and inexperienced ratings before and after training using many-facet Rasch measurement. The results demonstrated that inexperienced raters were more severe and inconsistent than experienced raters before training. Rater training proved to be successful in improving consistency of raters and reducing rater severity although significant severity was still present. In other words, rater training contributed to intra-rater reliability rather than inter-rater reliability. This finding is in line with Lumley and McNamara (1995) who found that rater training made raters more self-consistent, but not eliminated rater harshness.

In the light of these studies, the present study attempts to provide evidence on the points discussed above to support the valid interpretation and use of the rating scales developed to assess academic writing skills in TSL. Two rating scales, one for the graph-interpretation and one for argumentative essay task, were developed for the TSL test. Two research questions are addressed regarding the reliability and the validity of the rating scales and possible involvement of rater effects in the assigned scores:

1. How reliably does the rating scale function?
2. To what extent is the quality of ratings influenced by rater effects?

Methodology

Two rating scales were developed to assess two open-ended writing tasks which were constructed for the purpose of the study: 1) a graph interpretation task and 2) an argumentative essay geared at B2 and C1 levels, respectively (See Küçük, 2017 for a detailed discussion on task development). These tasks were given to a group of L2 learners of Turkish; two raters were specifically trained for scoring and the results were analyzed through Many-facet Rasch Measurement. The details are given below.

Rating Scale Development

The TSL writing rating scale was developed through an iterative process consisting of several stages such as trialing, revising, and multiple drafting. It was developed as an analytical rubric with four assessment criteria: content, organization, language use and vocabulary, each criterion having four levels and each level having two categories (see Küçük, 2017 for the English version of the scales and appendix for the Turkish version). The total score that can be assigned for an essay ranged from 4-36.

The rating scale was initially constructed based on the adaptations from the ESL Composition Profile by Jacobs et al. (1981, as cited in Weigle, 2002), IELTS band descriptors for Task 1 and Task 2 and Written Assessment Criteria Grid by Council of Europe (2009). Several researchers made use of Jacobs et al.'s ESL Composition profile in their study directly or by adapting it to rate the writing tasks (i.e., Bacha, 2001; Delaney, 2008; East, 2009; Ong & Zhang, 2010). IELTS band descriptors for Task 1 and Task 2 were considered relevant to the rating scales developed for this study due to the similarity of academic writing abilities tested in the two tests (a graph interpretation task and an independent argumentative essay task). The descriptors of Written Assessment Criteria Grid from CEFR (Council of Europe, 2009, p.187), especially the ones on range, accuracy and coherence, were also analyzed and incorporated where necessary.

However, after the first trial, it was observed that the descriptors of the initial draft scale did not adequately capture the features of student responses from two different task types. It was observed that the task requirements might vary across different task types, and hence the necessity to clarify them in the scale. The initial draft scale went through considerable revision through the analysis of student responses from the first and the second pilot testing and certain features from sample essays were identified and added to the scale. Special attention was paid to make the wording of the descriptors as explicit as possible.

The final draft consisted of two distinct rating scales to be used for the graph interpretation task and the argumentative essay task in an attempt to reflect the relevant constructs that each task was intended to represent (see appendix for the final version of the scale). It was hoped that this empirically-based improvement in the scale would make it more discriminating and result in higher inter-rater reliability and self-consistency among raters (Knoch, 2007).

Participants

The test was taken by 47 students who came to Turkey through the Erasmus International Student Exchange Program. The students were registered in courses at different levels of Turkish for Foreigner (TKF) classes (A2-B2). In terms of their country of birth, language background and age, the participants constituted a diverse group. Most participants had been learning Turkish for more than one year at the time of testing.

The two raters in the study were working as research assistants at the Department of Foreign Language Education of Boğaziçi University. They were native speakers of Turkish but did not have much experience in rating writing scripts by using a rating scale. Neither of the raters was experienced in rating responses that are written in Turkish by L2 learners of Turkish.

Data Collection Procedures

The data was collected in usual class time. The participants had 50 minutes to complete the two writing tasks (20 minutes for task 1 and 30 minutes for task 2). The participants were informed about the test by their instructors before the administration and the participation was voluntary.

Scoring Procedures

An intensive rater training session was conducted by one of the authors before the actual rating session. During the training session, the raters were first informed about task demands, writing construct and rating procedures. They then familiarized themselves with the scale descriptors and marked a bunch of benchmark scripts that represent different scale levels for each task. These were previously chosen and rated by the authors; an assessment expert and an experienced rater. Through extended discussions, a mutual understanding of the rating scale and score meaning was established. Each script was double-scored by the two raters, the scores were compared for discrepancy and when significant differences were observed between the first and the second rater, the raters were asked to score the scripts for the second time without seeing the initial scores.

Analyses

The scale efficiency is investigated using many-facet Rasch measurement (MFRM), which was implemented using Minifac (FACETS) software, version 3.71.4 developed by Linacre (2014). The MFRM analysis provides a variable map in which students, raters, tasks and criteria are calibrated on the same logit-scale with equal intervals. The variable map gives information about student distribution based on their proficiency, rater severity, and task and criteria difficulty. Along with the variable map, the MFRM analysis produces a measurement report (i.e., rater measurement report, examinee measurement report) for each facet involved in the analysis (Eckes, 2009). These measurement reports include statistics such as fit indices (infit and outfit mean square values), fixed effect chi-square tests, and two different separation statistics: The separation index and the reliability of separation index (Myford & Wolfe, 2003).

Fit indices show the extent to which the observed measures of students, raters and tasks match with the expected measures that are estimated by the MFRM model (Myford & Wolfe, 2003). Fit indices consist of infit and outfit mean square values. Possible mean square values for infit and outfit indices range between 0 and 1. Values between 0.5 and 1.5 are "…productive for measurement or …indicative of useful fit" (Linacre, 2008 as cited in Eckes, 2009, p. 18).

Fixed effect chi-square test indicates "[whether] the fixed effect hypothesis that the estimates of all the elements within a given facet can be viewed as sharing a common parameter, after allowing for measurement error" is true (Myford & Wolfe, 2003, p. 409). For example, the fixed effect chi-square test for tasks tests the hypothesis that all the tasks in the study are of equal difficulty. Similarly, 'the fixed effect chi-square test for the raters facet tests the hypothesis that all raters exercised the same level of severity when evaluating ratees, after accounting for measurement error' (Myford & Wolfe, 2003, p. 409). The significance value reported shows the probability of whether the fixed effect hypothesis should be kept or rejected.

The separation statistics reports "the amount of variability (or spread) in the measures estimated by the MFRM model for the various elements in the specified facet relative to the precision by which those measures are estimated" (Sudweeks, Reeve & Bradshaw, 2005, p. 245). The reliability of separation index can range between 0 and 1, whereas the value of the separation index ranges between 1 to infinity.

For the investigation of the reliability of the scale (the first research question), selected statistics from criterion measurement report, inter-rater and intra-rater reliability statistics, and category statistics that were produced by the MFRM analysis were used: Statistics of criteria measures, the criteria separation index, the reliability of criteria separation index and fit indices for criterion were reported. Criteria measures (in logits) are given in the variable map, which includes a column on criteria difficulty, criteria being content, organization, language and vocabulary in this case. It is harder for students to receive high scores on the scale criterion that appears higher in the column than on the criterion appearing lower. Similarly, the higher criterion measures show the more difficult criteria for students to get high scores on. The criteria separation index is used to identify the number of statistically different strata of criteria difficulty, which might be used to determine if the raters actually apply the rating scales in an analytical way or not (Myford & Wolfe, 2004). The reliability of separation index for criteria is expected to be closer to 1 as the criteria in a scale are supposed to be of differing difficulty as an indication of analytical functioning. The infit and outfit mean square values are expected to be close to 1 between the range of 0.5 and 1.5 in order to argue that they all relate to the same construct (unidimensionality) (Eckes, 2009).

For the estimation of inter-rater reliability, the point biserial correlation indices were used, and rater fit statistics were examined to find evidence for intra-rater reliability. The point biserial correlation measures need to be close to 1 for high inter-rater reliability (Knoch, 2007).

The MFRM also provides several statistics in order to evaluate the effectiveness of the rating scale. The average student measures, outfit mean-square values and Rasch Andrich Threshold measures were used to examine scale effectiveness (Eckes, 2009). The average student measures are required to advance monotonically as the scale categories increase from 1 to 8 in order to claim that the scale categories are appropriately ordered and meaningfully applied. Similarly, Rasch Andrich Threshold measures need to increase with the category scales. Outfit mean square values are supposed to be smaller than 2.0 to argue that the categories are used appropriately by the raters.

Raters are known as another potential source of variability that lead to construct irrelevant variance and lower score validity. The second question, therefore, is concerned with the potential involvement of rater effects in the ratings of student responses. Rater infit and outfit mean square values are expected to be between 0.5 and 1.5 in order to claim that raters used the rating scale consistently (intra-rater reliability) (Eckes, 2009). On the other hand, variable rater behaviour such as rater severity, halo effect, central tendency and inconsistency may endanger scoring validity; therefore, should also be investigated to show that no significant rater effect is involved in the ratings of student responses. The following statistics from the MFRM analysis were used for data analysis:

Rater severity: Rater severity measures, rater separation index and the reliability of rater separation index from the rater measurement report were used to examine differences in the severity of raters. Raters with higher measures (in logits) appear to have exercised higher levels of severity than the ones with lower severity measures. To be able to claim that raters involved in the ratings exercised similar levels of severity, the difference between the most severe and the least severe rater should be as small as possible. In addition, rater separation index is expected to be close to 1 to argue that the raters were interchangeable as the index shows the number of statistically distinct groups in terms of severity (Eckes, 2009). Finally, the reliability of rater separation index need to be close to 0 as this index indicates how separate the raters are in terms of severity they exercised (Myford &Wolfe, 2003). The closer the reliability index to 1, the more different the raters are in terms of severity.

Inconsistency: Rater fit indices allowed to investigate raters' inconsistency in their ratings. If infit and outfit mean square values are within the range of 0.5 and 1.5, one can argue that the raters were self-consistent in their ratings.

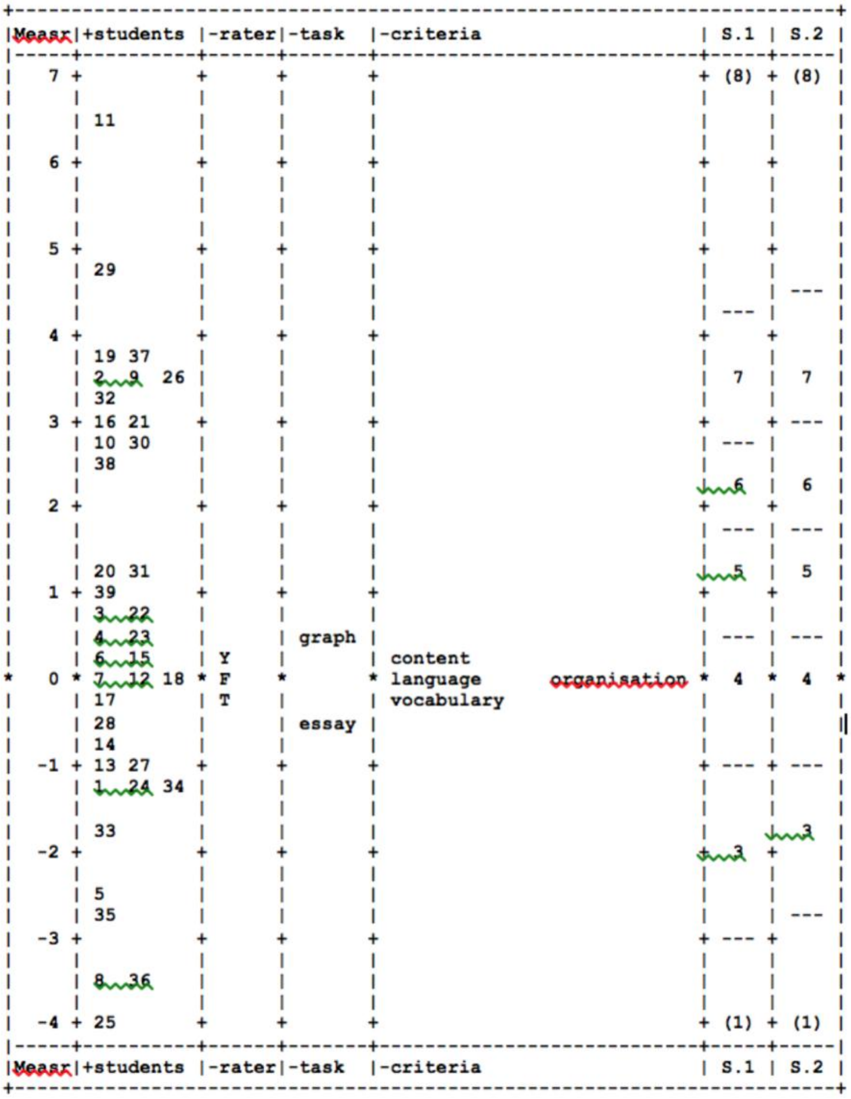
Central tendency: Rater fit indices and student separation index were used to investigate central tendency. When the infit and outfit mean square values are lower than 0.5 (overfitting raters), one can conclude that raters tended to overuse certain categories, providing evidence for central tendency when raters overuse the middle categories (Knoch, 2007; Myford & Wolfe, 2004). Student separation index is another statistics that can show the existence or absence of central tendency. A high student separation index means that students were well discriminated in terms of their levels of proficiency by the raters, and thus it can be used as another piece of evidence for the absence of central tendency (Sudweeks et al., 2005).

Halo effect: The criteria separation index from the criteria measurement report was used to examine halo effect. In order to argue that raters were able to distinguish between conceptually different aspects of rating scale (scale criteria), the index is expected to correspond to the number of criteria in the rating scale.

Results

Research Question 1: How Reliably Does the Rating Scale Function?

Figure 1 below shows the variable map which portrays graphically the measures of four facets specified in the analysis (students, raters, tasks, criteria). The first column in the map shows the logit scale, which is a true interval scale, as opposed to raw scores in which distances between intervals may be different (Park, 2004). The second column shows estimates of student proficiency. Each number represents one student, and higher scoring students appear at the top of the column whereas lower scoring students appear at the bottom, logit 0 being the average. The distribution of student proficiency measures is quite wide, ranging from a high of 6.48 logits to a low of -3.92 logits. The third column compares the raters in terms of their severity levels. More severe raters appear higher in the column and more lenient ones appear lower. Figure 1 shows the most severe rater (Rater Y) has a measure of 0.18 logit, while the most lenient rater (Rater T) has a measure of -0.15 logit, indicating that raters did not differ much in the levels of severity. The fourth column compares the two tasks in terms of their difficulty estimates. The more difficult task appears higher in the column whereas the easier task appears lower. Accordingly, the two tasks are not of equal difficulty, the graph interpretation task being relatively more difficult than the essay task. The fifth column compares the four scoring criteria in terms of their relative difficulties. Criteria appearing higher in the column were more difficult for the students to receive high ratings on than the criteria appearing lower in the column. Thus, it was somewhat most difficult to get high ratings on *content* (i.e., the most difficult criterion) than on *vocabulary* (i.e., the easiest criterion). The last two columns depict the eight-point scoring scales used to rate students' responses on the graph interpretation and essay task, respectively.



```
+-------------------------------------------------------------------------+
|Measr|+students |-rater|-task  |-criteria                   | S.1 | S.2 |
|-----+----------+------+-------+----------------------------+-----+-----|
|   7 +          +      +       +                            + (8) + (8) |
|     |          |      |       |                            |     |     |
|     | 11       |      |       |                            |     |     |
|     |          |      |       |                            |     |     |
|   6 +          +      +       +                            +     +     |
|     |          |      |       |                            |     |     |
|     |          |      |       |                            |     |     |
|     |          |      |       |                            |     |     |
|   5 +          +      +       +                            +     +     |
|     | 29       |      |       |                            |     |     |
|     |          |      |       |                            |     | --- |
|     |          |      |       |                            | --- |     |
|   4 +          +      +       +                            +     +     |
|     | 19 37    |      |       |                            |     |     |
|     | 2  9  26 |      |       |                            |  7  |  7  |
|     | 32       |      |       |                            |     |     |
|   3 + 16 21    +      +       +                            +     + --- |
|     | 10 30    |      |       |                            | --- |     |
|     | 38       |      |       |                            |     |     |
|     |          |      |       |                            |  6  |  6  |
|   2 +          +      +       +                            +     +     |
|     |          |      |       |                            | --- | --- |
|     |          |      |       |                            |     |     |
|     | 20 31    |      |       |                            |  5  |  5  |
|   1 + 39       +      +       +                            +     +     |
|     | 3  22    |      |       |                            |     |     |
|     | 4  23    |      | graph |                            | --- | --- |
|     | 6  15    | Y    |       | content                    |     |     |
*   0 * 7  12 18 * F    *       * language     organisation  *  4  *  4  *
|     | 17       | T    |       | vocabulary                 |     |     |
|     | 28       |      | essay |                            |     |     |
|     | 14       |      |       |                            |     |     |
|  -1 + 13 27    +      +       +                            + --- + --- |
|     | 1  24 34 |      |       |                            |     |     |
|     |          |      |       |                            |     |     |
|     | 33       |      |       |                            |     |  3  |
|  -2 +          +      +       +                            +  3  +     |
|     |          |      |       |                            |     |     |
|     | 5        |      |       |                            |     |     |
|     | 35       |      |       |                            |     | --- |
|  -3 +          +      +       +                            + --- +     |
|     |          |      |       |                            |     |     |
|     | 8  36    |      |       |                            |     |     |
|     |          |      |       |                            |     |     |
|  -4 + 25       +      +       +                            + (1) + (1) |
|-----+----------+------+-------+----------------------------+-----+-----|
|Measr|+students |-rater|-task  |-criteria                   | S.1 | S.2 |
+-------------------------------------------------------------------------+
```

Note: S.1 = Scoring rubric used for graph interpretation task; S.2 = Scoring rubric used for essay task

Figure 1. FACETS summary (student proficiency, rater severity, task and criteria difficulty)

To investigate the first research question in detail, selected statistics from the MFRM analysis were reported. Specifically, selected statistics from the criteria measurement report, rater point biserial correlation indices and rater fit indices from the rater measurement report and selected statistics from the category statistics were used to examine the effectiveness of the rating scales used in the present study.

Criteria

A summary of selected statistics included in the criteria measurement report is provided in Table 1.

Table 1. Summary of statistics included in the criteria measurement report

Criterion	Difficulty Measure	Standard Error	Infit Mean-Square Index	Outfit Mean-Square Index
Content	0.31	0.10	1.40	1.34
Organization	0.03	0.10	0.93	0.99
Language Use	-0.07	0.10	0.82	0.81
Vocabulary	-0.28	0.10	0.67	0.64
Mean	0.00	0.10	0.95	0.94
S.D.	0.21	0.00	0.27	0.26

Note: Reliability of separation index = 0.78; separation index = 2.82; fixed chi- square:17.9, *df*:3, *p* = .00

Table 1 indicates that the four criteria differed somewhat in difficulty as suggested by the criteria separation index (2.82) and the reliability of criteria separation (0.78). Among the four criteria, there were nearly three statistically distinct levels of difficulty. Specifically, the hardest criterion to get high ratings on was content (0.31 logit). By contrast, the easiest criterion to get high ratings on was vocabulary (-0.28 logit). The difficulty measures for organization (0.03 logit) and language use (-0.07 logit) were very similar. These results may provide evidence for the effective functioning of the criteria although one could suspect that two of the criteria (i.e., language use and organization) may be functioning somewhat similarly.

The infit and outfit mean-square values for the criteria were within the acceptable range of 0.5 to 1.5, indicating that there were no overfitting or misfitting criteria. The fact that there were no overfitting criteria suggests that the four criteria were not scored too similarly, and the fact that there is no misfitting criterion provides evidence for psychometric unidimensionality of the four criteria, suggesting that they might all be associated with the same underlying construct (Eckes, 2009). In other words, ratings on one criterion agree well with the ratings on other criteria, leading to a single pattern of proficiency across all four criteria (Park, 2004).

Inter-rater and intra-rater reliability estimates

To measure inter-rater reliability, FACETS provides two measures of rater reliability: the rater point biserial correlation index and the percentage of exact rater agreement. The former is a measure of how similar the raters are in their rankings of students and the latter shows the percentage of how many times the raters assigned exactly the same score as another rater (Knoch, 2007). Table 2 provides the summary of these two rater reliability measures.

Table 2. Summary of rater reliability measures

Rater	Rater Point Biserial Measure	Percentage of Exact Agreement
F	0.90	48.1 %
T	0.91	42.7 %
Y	0.88	37.5 %

Myford and Wolfe (2003) use the term "single rater-rest of raters correlations" for this type of correlation index, which means that each correlation index indicates the correlation measure of one rater with the other two raters within this group of raters (p. 416). Accordingly, the single rater-rest of rater correlations seem to be substantial, which were 0.90, 0.91, and 0.88 for Rater #F, Rater #T and Rater #Y, respectively, suggesting a significant level of agreement between the raters; therefore, high reliability in the scoring. The third column indicates that Rater #F has the highest exact agreement percentage (48.1%), suggesting that Rater #F awarded exactly the same scores 48.1% of times as the other raters under the same conditions, while Rater #Y has the lowest agreement percentage (37.5%). For intra-rater reliability, rater infit and outfit mean square values are provided by the rater measurement report (see Table 4).

Category Statistics

In order to examine whether eight-point rating scales which were used to score students' responses for graph interpretation and essay tasks functioned as intended, a summary of selected category statistics are given in Table 3.

Table 3. Category statistics for rating scales

	Graph				Essay			
Cat.	AvMeas	OFit	Threshold	SE	AvMeas	Ofit	Threshold	SE
1	-3.53	1.4			-2.77	1.0		
2	-2.98	1.2	-5.6	0.43	-2.72	0.6	-5.85	1.01
3	-1.36	1.1	-3.11	0.24	-1.17	0.9	-2.61	0.28
4	-0.46	1.2	-0.74	0.18	0.09	0.8	-0.82	0.20
5	0.68	0.9	0.94	0.22	0.95	0.9	0.72	0.18
6	2.19	1.2	1.46	0.24	2.09	0.8	1.82	0.22
7	3.16	0.8	3.01	0.22	3.75	0.6	2.90	0.23
8	4.14	0.9	3.50	0.28	4.62	1.0	3.83	0.22

Note: Cat. = Category, AvMeas _ Average Measure, Ofit = Outfit, Thresholds = Rasch-Andrich thresholds,

The first column in Table 3 shows category labels as appeared in scoring scales ranging from 1 and 8. The second and the sixth column indicate the average student proficiency measure by rating scale category. Linacre (2002) suggests that the average measures should advance monotonically as the categories increase. For both tasks, this seems to be the case since the average measures for both scales increase as the categories increase (from -3.53 to 4.14 for the graph interpretation task and from -2.77 to 4.62 for the essay task). As such, the categories for both tasks were ordered appropriately and meaningfully.

Outfit mean-square index (the third column for the graph interpretation task and the seventh column for the essay task) is another indicator of rating scale functionality (Linacre, 2002). FACETS computes average student proficiency measure and an expected student proficiency measure. The larger the discrepancy between the average and expected measures, the larger the outfit mean-square index will be (Eckes, 2009). Linacre (2002) suggests that outfit mean-square index should be less than 2.0 as a high value of mean-square related to a category is evidence for the fact that the category has been used in unexpected contexts. As shown in Table 3, all the outfit mean-square values were less than 2.0. This suggests that the categories for both rating scales seemed to function as intended.

The category thresholds (columns 4 and 8) can also provide information on the quality of a rating scale. It is expected that these thresholds advance monotonically with categories. Otherwise, it means that they are disordered suggesting low probability of occurrence of certain categories due to the rating behavior in which those categories are employed (Linacre, 2002). Table 3 shows that threshold measures advance monotonically as the categories increase (i.e., from -5.6 to 3.50 for the first scale, and from -5.85 to 3.83 for the second scale).

Research Question 2: To What Extent is the Quality of Ratings Influenced by Rater Effects?

This question was examined in terms of rater severity, rater inconsistency, central tendency and halo effect through selected statistics from the rater measurement report provided by many-facet Rasch measurement analysis and selected statistics from student measurement report.

Rater Severity

The rater measurement report reports a measure of the level of severity each rater exercised, as well as measures of each rater’s ability to use the rating scales in a consistent manner when evaluating multiple students’ responses (see Table 4)

Table 4. Summary of statistics included in the rater measurement report

Rater	Severity	Standard error	Infit Mean-Square	Outfit Mean-
F	-0.03	0.08	0.92	0.90
T	-0.15	0.09	0.82	0.86
Y	0.18	0.09	1.14	1.09
Mean	0.00	0.09	0.96	0.95
S.D.	0.14	0.00	0.13	0.10

Note: Reliability of separation index = 0.61, separation index = 2.01

The first column shows the rater IDs. The second column shows that the difference between the severity measures of the most severe (Rater #Y) rater and the most lenient (Rater #T) rater was 0.33 logits, indicating that the three raters appeared to exercise similar levels of severity when rating students' responses. The rater separation index indicates the number of statistically distinct groups in terms of rater severity. Thus, the separation index of 2.01 suggests that there were about two statistically distinct strata of rater severity within this small group of raters. The reliability of rater separation index indicates how different the raters are in their severity measures unlike inter-rater reliability, which is a measure of how similar the raters are in their severity measures (Eckes, 2009, p. 20). In other words, when raters display similar measures of severity, the reliability of separation index is expected to be close to 0. Therefore, a low separation reliability index is desirable for raters. The rater separation reliability index for this analysis was 0.61, indicating that the raters differed somewhat in their severity.

Rater Inconsistency

To examine rater inconsistency, rater fit indices were used. The fourth and fifth columns in Table 4 show rater fit statistics. One examines rater fit statistics to determine whether raters used the rating scales in a consistent manner (Eckes, 2009). The infit and outfit mean-square values for all three raters were within the range of 0.5 and 1.5, which means that none of them were misfitting. That is to say, all of the raters were self-consistent in their ratings.

Central Tendency

Central tendency was examined through rater fit indices (Table 4) and student separation index (Table 5). The fact that there were no overfitting raters (i.e., no infit mean-square values lower than 0.5) suggests that the raters did not tend to overuse certain (generally middle) scale categories, which could lead the raters to appear as too consistent. An overfitting rater is one who has assigned ratings that are closer to the expected ratings than the measurement model predicts. This was not the case with this particular group of raters (Knoch, 2007).

The student proficiency measures ranged from -3.92 to 6.48 logits, with a mean of 0.70 logit (SD = 2.44). The student separation index was 9.46, with a reliability index of 0.98. The separation index is an estimate of the number of distinguishable levels of proficiency among the students. The separation index of 9.46 indicates that there were about nine statistically distinct strata among the 39 student proficiency measures.

Table 5. Summary of results for students (*N* = 39)

Mean of the proficiency measures	0.70
Standard deviation of proficiency measures	2.44
Student separation index	9.46
Reliability of student separation	0.98
Fixed (all same) chi-square	1771.8 (*df* = 38, *p* = .00)

The reliability of the student separation is the Rasch equivalent of KR20 or Cronbach Alpha statistics (O'Sullivan, 2005). A reliability coefficient of 0.98 indicates that the raters' ratings on the two tasks reliably separated students into different levels of proficiency. It also suggests that those ratings did not show evidence of central tendency error.

Halo Effect

The criteria separation index from the criteria measurement report was used to investigate whether the raters were able to distinguish between different aspects of the rating scales (see Table 1). The fact that there were nearly three distinct levels of difficulty among the four criteria provides evidence for the absence of halo effect. This finding suggests that the raters were able to discriminate among the three criteria in the rating scales.

Discussion

Rating scales and their effective use by raters are two important facets that contribute to the validity of our decisions based on the writing test scores. This study focused on the quantitative validation of the TSL academic writing rating scale and the raters' use of the scale in assessing the academic writing skills of TSL learners.

The first issue handled was how reliably the rating scale functions for its intended purpose. East (2009) citing Cherry and Meyer (1993) notes that "the more pieces of information available, the more reliable will be the conclusions drawn from the data" (p. 92). In other words, multiple ways used to gather evidence for the reliability of the rating are likely to increase the trustworthiness of the conclusions. One of these ways is to examine the results of the criteria measurement report generated by MFRM analysis. The statistics of criteria measures showed that the most difficult criterion was content, whereas vocabulary was the easiest criterion for the students to get high ratings. In other words, the students had the tendency to score higher on vocabulary than on the other criteria, and they scored lowest on content. Organization and language use were of similar difficulty although they received slightly higher scores on language use. This finding suggests that learners of TSL may have differing proficiency levels in different aspects of writing ability, and the analytic rating scales used in the study were able to reflect the uneven profile of L2 learner's writing proficiency just as Weigle (2002) points out with regard to advantages of using an analytic rating scale. In addition, the findings of the study seem to confirm previous research on differential performance in writing output (i.e. Bacha, 2001). Bacha (2001) found that the analytic ratings assigned for different components of writing were significantly different from each other with L1 Arabic students of English. Similarly, the TSL students in the present study had the necessary vocabulary repertoire and linguistic structures to complete the academic tasks, but they generally had problems generating ideas on the given subject (content) and organizing their ideas as required by each task (organization). This problem was more salient with the graph interpretation task. The analytic scores helped to understand that some students were unfamiliar with the content requirements of the graph interpretation task despite the fact that they had a good control of grammatical structures and vocabulary (See Küçük Üçpınar & Ünaldı (in preparation) for further discussion of the issue).

The criteria separation index generated by the criteria measurement report indicated nearly three statistically distinct levels of difficulty among the four criteria. This finding suggests that raters were able to distinguish at least three criteria in the rating scales, and thus the rating scales functioned analytically as intended. This finding may provide support for the effectiveness of empirically developed rating scales as suggested by various researchers (i.e., Knoch, 2007; Turner & Upshur, 2002). The descriptors of empirically-based rating scales are based on the analysis of actual student responses; therefore, they are argued to be more discriminating and explicit in terms of their level descriptors than intuitively developed rating scales (Turner & Upshur, 2002; Knoch, 2007). The raters in the present study were able to distinguish between the criteria successfully in spite of their inexperience in rating and the scale's empirically derived descriptors seemed to help overcome their lack of experience.

The fact that the infit and outfit mean square values of the criteria were within the acceptable levels suggests that these four criteria (i.e., content, organization, language use and vocabulary) relate to the same general dimension, the general writing construct, giving support to the assumption of psychometric unidimensionality (Eckes, 2009). The fact that different traits of the rating scale were all related to the same underlying writing construct may provide further evidence for the validity of the rating scales used in the study (Park, 2004).

The findings from category statistics provided empirical evidence for the effectiveness of rating scale categories, as suggested by Linacre (2002). Eight categories of the rating scales were appropriately ordered and satisfactorily distinguishable. The fact that raters were able to distinguish between 8 categories as identified in the rating scales and used them appropriately may suggest another evidence for validity of the rating scales. However, as Myford (personal e-mail communication, November 8, 2016) cautioned, the rating scales should be used with much larger number of students and raters in order to obtain more accurate and stable values of category statistics and to be able to make sound claims about how well rating scale categories function.

The MFRM analysis provided useful information to detect and evaluate rater effects that might have been involved in the ratings of student responses. Selected statistics from the rater measurement report suggested that the raters exercised similar levels of severity although they were not interchangeable. Rater fit indices indicated the raters were consistent in the way they applied the scoring criteria. These may provide further evidence for the effect of rater training in improving raters' self-consistency rather than eliminating the differences in rater severity (i.e., Weigle, 1998). Lumley and McNamara (1995) argue that the main concern of the rater training should be to minimize "the random error in rater judgments" due to the fact that a lack of self-consistency in the ratings makes it impossible to carry out an orderly process of measurement (p.57). Similarly, McNamara (1996) stated,

> To accept that the most appropriate aim of rater training is to make raters internally consistent so as to make statistical modelling of their characteristics possible, but beyond this to accept variability in stable rater characteristics as a fact of life, which must be compensated for in some way. (p. 127)

The point biserial correlation measures in Table 2 indicated a considerable degree of consistency among the raters in the current study although some levels of difference in severity existed. Therefore, the raters in the current study seemed to have applied the rating scales consistently and similarly, and a lack of inter-rater reliability was not an issue in the study, either. Evidence for the absence of central tendency effect was obtained from fit statistics and the measure of student separation index (Myford & Wolfe, 2004). The findings suggest that raters made use of all the scale categories in their ratings and they were able to distinguish between the performances of students that displayed different levels of proficiency. The fact that halo effect did not appear with this group of raters was already discussed in the first question to investigate the reliability of the rating scales. The criteria separation index suggested that raters applied the rating scales analytically as intended.

When the four types of rater effects examined, it seems that there is evidence for the lack of strong rater-related variance. One possible explanation for these findings might be the overall effectiveness of the rater training despite the fact that it was brief. It is often stressed in the literature that having a well-developed rating scale with clear and explicit level descriptors would be insufficient without exemplifying those level descriptors with actual student responses through rater training (e.g.., Shaw & Weir, 2007). Along with the effectiveness of rater training, the findings seem to provide evidence for the effectiveness of the empirically developed rating scales and purposefully selected benchmark essays. Knoch (2007) argues that intuitively developed rating scales have the potential to cause various rater effects. As the descriptors of such scales may not reflect the characteristics of learners' actual language use explicitly, raters tend to create their individual interpretations of the descriptors. This in turn may cause rater severity, inconsistencies, halo effect or central tendency if raters simply choose 'the play-it safe method' and use the middle categories. The current study seems to provide support for the efficiency of empirically developed rating scales. All in all, the fact that serious rater effects were not involved in the ratings and the scales work efficiently provides evidence for the score validity of the TSL Academic Writing Test.

Conclusion

Overall, this study suggests that empirically-developed analytical rating scales used to assess students' academic writing skills in TSL operate reliably. Multiple types of evidence collected through MRFM analyses helped establish the validity of the rating scales, which in turn contributed to overall validity of the academic writing test for TSL. As well as emphasising the importance of developing effective rating scales, the study revealed the significance of rater training to familiarize them with the scale descriptors in order to avoid potential involvement of raters' personal judgements in the scores, which can lead to inconsistent ratings within and among the raters. Finally, it should be noted that the conclusions drawn here are preliminary and tentative based on a small number of student responses and a small group of raters. In order to make stronger validity claims, rating scales need to be further investigated by extending their use in new tasks with higher number of students, and raters. We are hoping that the scales developed in this study will further be used in other TSL assessment contexts and this study will help the TSL researchers in the development of other scales.

References

Bacha, N. (2001). Writing evaluation: What can analytic versus holistic essay scoring tell us? *System*, *29*(3), 371-383.

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing*, *12*(2), 86-107.

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly*, *7*(1), 54-74.

Barkaoui, K. (2011). Effects of marking method and rater experience on ESL essay scores and rater performance. *Assessment in Education: Principles, Policy & Practice*, *18*(3), 279-293.

Council of Europe (2001). *Common European reference framework for languages*. Strasbourg, France: Author. Retrieved from http://www.coe.int/T/DG4/Portfolio/?L=E&M=/documents_intro/common_framework.html

Council of Europe (2009). *Manual for Relating Language Examinations to the Common European Framework of Reference for Languages (CEFR)*. Strasbourg, France: Author. Retrieved from http://www.coe.int/t/dg4/linguistic/Manuel1_EN.asp

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing*, *7*(1), 31-51.

Delaney, Y. A. (2008). Investigating the reading-to-write construct. *Journal of English for Academic Purposes*, *7*(3), 140-150.

Davies A, Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of Language Testing*. Cambridge: Cambridge University Press.

East, M. (2009). Evaluating the reliability of a detailed analytic scoring rubric for foreign language writing. *Assessing Writing*, *14*(2), 88-115.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly: An International Journal*, *2*(3), 197-221.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, *25*(2), 155-185.

Eckes, T. (2009). Many-facet Rasch measurement. *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment*. (Section H). Strasbourg, France: Council of Europe/Language Policy Division. Retrieved from http://www.coe.int/t/dg4/Linguistic/CEF-refSupp-SectionH.pdf

Eckes, T. (2012). Operational rater types in writing assessment: Linking rater cognition to rater behavior. *Language Assessment Quarterly*, *9*(3), 270-292.

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.

Fulcher, G. (2003). *Testing second language speaking*. Harlow: Pearson Education.

Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing*, *28*(1), 5-29.

Hamp-Lyons, L., & Mathias, S. P. (1994). Examining expert judgments of task difficulty on essay tests. *Journal of Second Language Writing*, *3*(1), 49-68.

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

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, *2*(2), 130-144.

Knoch, U. (2007). Do empirically developed rating scales function differently to conventional rating scales for academic writing? *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, *5*, 1–36.

Knoch, U. (2009). *Diagnostic writing assessment: The development and validation of a rating scale.* Frankfurt, Germany: Peter Lang.

Küçük, F. (2017). *Assessing academic writing skills in Turkish as a foreign language*. (Unpublished master's thesis). Boğaziçi University, Istanbul, Turkey.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness, *Journal of Applied Measurement*, *3*(1), 85-106.

Linacre, J. M. (2014). *FACETS* (Version 3.71.4) [Computer software]. Chicago, IL: MESA Press.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, *12*(1), 54-71.

Mendoza, A., & Knoch, U. (2018). Examining the validity of an analytic rating scale for a Spanish test for academic purposes using the argument-based approach to validation. *Assessing Writing*, *35,* 41-55.

McNamara, T. F. (1996). *Measuring second language performance*. London and New York: Longman.

Myford, C. M. (2002). Investigating design features of descriptive graphic rating scales. *Applied Measurement in Education*, *15*(2), 187-215.

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

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

O'Sullivan, B. (2005). *A practical introduction to using FACETS in language testing research*. Unpublished manuscript, University of Roehampton, London, UK.

Ong, J., & Zhang, L. J. (2010). Effects of task complexity on the fluency and lexical complexity in EFL students' argumentative writing. *Journal of Second Language Writing*, *19*(4), 218-233.

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.

Sakyi, A. (2000). Validation of holistic scoring for ESL writing assessment: A study of how raters evaluate ESL compositions on a holistic scale. In A. Kunnan (Ed.), *Fairness and validation in language assessment* (pp. 130–153). Cambridge: Cambridge University Press.

Shaw, S. D., & Weir, C. J. (2007). *Examining writing: Research and practice in assessing second language writing* (Vol. 26). Cambridge: Cambridge University Press.

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

Turner, C. E., & Upshur, J. A. (2002). Rating scales derived from student samples: Effects of the scale maker and the student sample on scale content and student scores. *TESOL Quarterly*, *36*(1), 49–70.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing*, *11*(2), 197-223.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, *15*(2), 263-287.

Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing*, *6*(2), 145-178.

Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.

Weigle, S. C., Boldt, H., & Valsecchi, M. I. (2003). Effects of task and rater background on the evaluation of ESL student writing: A pilot study. *TESOL Quarterly*, *37*(2), 345-354.

Wolfe, E. W., Kao, C. W., & Ranney, M. (1998). Cognitive differences in proficient and nonproficient essay scorers. *Written Communication*, *15*(4), 465-492.

Wolfe, E. W. (2004). Identifying rater effects using latent trait models. *Psychology Science*, *46*, 35–51.

İkinci Dil Olarak Türkçede Yazma Becerisi Değerlendirme Ölçeklerinin Geçerliğinin Çok Yönlü Rasch Ölçüm Modeliyle İncelenmesi

Özet

Açık uçlu yazma görevleri söz konusu olduğunda dil sınavlarının puanlama geçerliliğini etkileyen iki önemli unsur, değerlendirme ölçekleri ve değerlendiricilerin ölçekleri ne derece etkili kullandıklarıdır. İngilizcenin değerlendirilmesinde değerlendirme ölçeklerinin geliştirilmesi ve geçerlemesiyle ilgili birçok bilimsel çalışma olmasına rağmen ikinci dil olarak Türkçenin değerlendirilmesinde bu konular üzerinde şimdiye kadar pek çalışılmamıştır. Bu çalışma, ikinci dil olarak Türkçede akademik yazma becerisini ölçmek için kullanılan iki analitik değerlendirme ölçeğinin nasıl geliştirildiğini rapor etmektedir ve bu ölçeklerin geçerliğine nicel kanıt sunmaktadır. Bu amaçla, ikinci dil olarak Türkçe öğrenen 39 öğrencinin yazdığı metinler üç değerlendirici tarafından puanlanmıştır. Analizler çok yönlü Rasch ölçüm modeliyle yapılmıştır. Analiz sonuçları, görgül olarak geliştirilmiş analitik ölçeklerin değerlendiriciler tarafından tutarlı ve uygun bir biçimde kullandığını göstermiştir. Bu da ölçeklerin güvenilirliğine ve etkinliğine kanıt sağlamıştır.

Anahtar kelimeler: Ölçek geliştirme, ikinci dil olarak Türkçede yazma becerisinin değerlendirilmesi, çok yönlü Rasch ölçüm modeli

Appendix

RATING SCALES

YAZMA BECERİSİ DEĞERLENDİRME ÖLÇEĞİ- GÖREV 1

İÇERİK	MÜKEMMEL-ÇOK İYİ	8-7	Görevin bütün gereklerini eksiksiz olarak karşılar: Ana eğilimleri anlaşılır ve etkili bir şekilde tanımlar, önemli bilgileri betimler ve/ veya karşılaştırır.
	İYİ-ORTALAMA	6-5	Görevin gereklerini yeterince karşılar: Ana eğilimleri genel olarak ortaya koyar, fakat bazı önemli bilgileri atlamış ya da birkaç gereksiz ayrıntıya yer vermiş olabilir.
	ORTA-ZAYIF	4-3	Görevin gereklerini karşılamaya çalışsa da hepsini yerine getiremez: Ana eğilimlere dair açık bir genel bakış sunamaz. Gereksiz ayrıntıya ya da yanlış bilgilere yer verir, ya da önemli verileri atlar.
	YETERSİZ	2-1	Görevin hiçbir gereğini karşılayamaz ya da sadece birkaç noktasını ele alır. Cevap, verilen görevle çok az alakalı ya da tamamen alakasızdır.

ORGANİZASYON	MÜKEMMEL-ÇOK İYİ	8-7	Bilgiler iyi düzenlenmiş ve önem derecesine göre mantıklı bir biçimde sıralanmıştır. Bilgiler uygun geçiş sözcükleri ya da sözcük grupları kullanılarak anlamlı bir şekilde sunulmuş, karşılaştırılmış ya da karşıtlanmıştır.
	İYİ-ORTALAMA	6-5	Bilgiler genel olarak düzenlidir ve önem derecesine göre mantıklı bir şekilde sıralanmıştır, fakat bazı noktalarda verilerin önem derecesi karıştırılmış olabilir. Bilgiler genellikle birbiriyle bağlantılıdır; karşılaştırma, bir dizi geçiş sözcüğü ya da sözcük grubu kullanımı sayesinde çoğunlukla anlamlıdır. Fakat bunlar bazen aşırı, bazen yetersiz ya da yanlış kullanılmış olabilir.
	ORTA-ZAYIF	4-3	Bilgiler mantıklı bir sıralama ve ilerleme olmadan sunulmuştur. Bilgiler genellikle birbirinden kopuktur; karşılaştırma, kısıtlı, yanlış ya da uygun olmayan geçiş sözcüğü ya da sözcük grubu kullanımı nedeniyle çoğunlukla başarısızdır.
	YETERSİZ	2-1	Düzen ve sıralama zayıf ya da hiç yoktur; olgu ve bilgiler arasında bağlantı ve tutarlılık yoktur.

DİL KULLANIMI	MUKEMMEL-ÇOK İYİ	8-7	Çok sayıda karmaşık gramer yapısını etkili bir biçimde kullanır. Gramer hataları nadiren görülür.
	İYİ-ORTALAMA	6-5	Etkili fakat basit gramer yapısı kullanır. Karmaşık yapıları kullanmaya çalışır, fakat bu yapıları basit yapılar kadar doğru kullanamaz. Bazı gramer ve noktalama hataları yapar, fakat bu hatalar anlamı nadiren bozar.
	ORTA-ZAYIF	4-3	Kullandığı gramer yapıları sınırlıdır. Hem basit hem karmaşık yapılarda sık sık hata yapar. Noktalama sıklıkla hatalıdır. Hatalar çoğunlukla anlamı bozar.
	YETERSİZ	2-1	Cümle yapısına hâkimiyeti çok azdır ya da hiç yoktur. Anlamı bozan gramer hataları çok fazladır.

SÖZCÜK DAĞARCIĞI	MUKEMMEL-ÇOK İYİ	8-7	Görevle ilgili çok çeşitli sözcük dağarcığı unsurunu, sözcük özelliklerine incelikli biçimde hakim olarak kullanır. Sözcük yazımında ya da yapısında nadiren hatalar olabilir.
	İYİ-ORTALAMA	6-5	Görevi yerine getirmek için yeterli çeşitlilikte sözcük dağarcığı unsuru kullanır, fakat sözcük seçiminde, yazımında ve yapısında bazı hatalar olabilir. Hatalar genellikle anlamı bozmaz.
	ORTA-ZAYIF	4-3	Kullandığı sözcük dağarcığı unsuru çeşidi sınırlıdır ve sözcükler uygun olmayan yerlerde ya da yinelenerek kullanılmış olabilir. Sözcük seçiminde, yazımında ve yapısında sık sık hata yapar. Hatalar anlamı bozabilir.
	YETERSİZ	2-1	Türkçe sözcük dağarcığı ve sözcük yapısı bilgisi çok azdır ya da hiç yoktur. Yalnızca birkaç ilgili sözcük kullanır, fakat sözcüklerin anlamları verilen bağlamda genellikle anlaşılmaz.

0	Görevi hiçbir şekilde yapmaz. Tamamen önceden ezberlenmiş bir metin yazar. Değerlendirilemeyecek kadar az yazar (30 kelimeden az).

YAZMA BECERİSİ DEĞERLENDİRME ÖLÇEĞİ – GÖREV 2

İÇERİK	MÜKEMMEL-ÇOK İYİ	8-7	Görevin ve konunun tamamını eksiksiz olarak ele alır: Verilen soruya ilgili argümanlarla iyi geliştirilmiş bir görüş sunar ve uygun sonuçlara varır. Ana fikirleri, mantıklı açıklama ve örneklerle kapsamlı bir şekilde geliştirir.
	İYİ-ORTALAMA	6-5	Verilen görevin tamamını ele alır fakat bazı kısımları yeterince ayrıntılandıramaz: Soruyla alakalı görüş belirtir ve uygun argümanlar sunar; ama bazı kısımlarda ayrıntılar ve örnekler yetersiz kalabilir.
	ORTA-ZAYIF	4-3	Konuyu sadece kısmen ele alır: Konuyla alakalı açık bir görüş belirtemez. Çoğunlukla iyi geliştirilmemiş, kendini tekrar eden ya da konuyla alakasız birkaç argüman sunar.
	YETERSİZ	2-1	Görevi neredeyse hiç ele almaz; Konuyla ilgili bir görüş belirtmez. Birkaç fikir sunsa da fikirleri geliştirmez ya da çok az geliştirir. Cevabı konuyla çok az alakalı ya da tamamen alakasız olabilir.

ORGANİZASYON	MÜKEMMEL-ÇOK İYİ	8-7	Fikirler iyi düzenlenmiş; ve giriş, gelişme, sonuç şeklinde mantıklı bir biçimde sıralanmıştır. Fikirler tutarlı ve bağlantılıdır. Bağdaşıklık araçlarının etkin kullanımıyla, fikirler arası geçişler oldukça başarılıdır.
	İYİ-ORTALAMA	6-5	Fikirler genellikle düzenlidir ve cevap metninde net bir akış vardır. Fikirler genellikle tutarlı ve bağlantılıdır. Birtakım bağdaşıklık araçları kullanılmıştır, fakat bunlar bazen aşırı, yetersiz ya da yanlış kullanılmış olabilir. Gönderim ögeleri her zaman uygun yerlerde ve anlaşılır biçimde kullanılmayabilir.
	ORTA-ZAYIF	4-3	Fikirler sunulmuş fakat mantıklı bir şekilde düzenlenmemiştir ve metinde anlaşılır bütünsel ilerleyiş görülmemektedir. Birkaç temel bağdaşıklık aracı kullanılmıştır, fakat bunlar genellikle yanlış ya da uygun olmayan durumlarda kullanılmıştır. Fikirler gönderim, değiştirim ve bağlantı unsuru kullanımı eksikliğinden çoğunlukla birbiriyle bağlantılı olmayabilir ve/veya birbirini tekrarlayabilir.
	YETERSİZ	2-1	Düzen ve sıralama zayıftır ya da hiç yoktur. Fikirler bağlantısız ve tutarsızdır.

DİL KULLANIMI	MÜKEMMEL-ÇOK İYİ	8-7	Birçok karmaşık gramer yapısını etkili bir biçimde kullanır. Çok nadir gramer hataları yapar.
	İYİ-ORTALAMA	6-5	Etkili fakat basit gramer yapıları kullanır. Karmaşık yapıları kullanmaya çalışır, fakat bu yapıları basit yapılar kadar doğru kullanamaz. Bazı gramer ve noktalama hataları yapar, fakat bu hatalar anlamı nadiren bozar.
	ORTA-ZAYIF	4-3	Kullandığı gramer yapısı sınırlıdır. Hem basit hem karmaşık yapılarda sık sık hata yapar. Noktalama sıklıkla hatalıdır. Bu hatalar çoğunlukla anlamı bozabilir.
	YETERSİZ	2-1	Cümle yapısına hâkimiyeti çok azdır ya da hiç yoktur Anlamı bozan gramer hataları çok fazladır.

SÖZCÜK DAĞARCIĞI	MÜKEMMEL- ÇOK İYİ	8-7	Çok çeşitli sözcük dağarcığı unsurunu, sözcük özelliklerine incelikli biçimde hakim olarak kullanır. Nadir kullanılan kelimeleri etkin bir biçimde kullanır, fakat kelime seçiminde ya da kalıplaşmış söz öbeklerinin kullanımda ara sıra hatalar olabilir. Sözcük yazımında ya da yapısında hiç hata yapmaz ya da nadiren hata yapar.
	İYİ-ORTALAMA	6-5	Yeterli çeşitlilikte sözcük dağarcığı unsuru kullanır. Daha az kullanılan kelimeleri kullanmaya çalışır, fakat bazı hatalar yapar. Sözcük yazımında ve yapısında bazı hatalar yapar, fakat bu hatalar anlamı nadiren bozar.
	ORTA-ZAYIF	4-3	Kullandığı sözcük dağarcığı unsuru sınırlıdır ve bunlar da uygun olmayan yerlerde ya da yinelenerek kullanılabilir. Sözcük yazımında, yapısında ve seçiminde sık sık hata yapar. Hatalar çoğunlukla anlamı bozabilir.
	YETERSİZ	2-1	Türkçe sözcük dağarcığı ve sözcük yapısı bilgisi çok azdır ya da hiç yoktur. Yalnızca birkaç alakalı sözcük kullanır, fakat anlamları verilen bağlamda genellikle anlaşılmaz.

0	Görevi hiçbir şekilde yapmaz. Tamamen önceden ezberlenmiş bir metin yazar. Değerlendirilemeyecek kadar az yazar (40 kelimeden az).