International Journal of Assessment Tools in Education
2022, Vol. 9, No. 1, 80–97
https://doi.org/10.21449/ijate.1059304
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Standard Setting in Academic Writing Assessment through Objective Standard Setting Method

Fatima Nur Fisne[1,*], **Mehmet Sata**[2], **Ismail Karakaya**[3]

[1]Gazi University, Gazi Faculty of Education, English Language Teaching Program, Turkiye
[2]Agri Ibrahim Cecen University, Faculty of Education, Department of Measurement and Evaluation in Education, Turkiye
[3]Gazi University, Gazi Faculty of Education, Department of Measurement and Evaluation in Education, Turkiye

ARTICLE HISTORY

Received: Mar. 19, 2021
Revised: Nov. 19, 2021
Accepted: Jan. 13, 2022

Keywords:
L2 academic writing assessment,
Many-facet Rasch measurement model,
Standard setting,
OSS

Abstract: Performance standards have important consequences for all the stakeholders in the assessment of L2 academic writing. These standards not only describe the level of writing performance but also provide a basis for making evaluative decisions on the academic writing. Such a high-stakes role of the performance standards requires the enhancement of objectivity in standard setting procedure. Accordingly, this study aims to shed light upon the usefulness of Objective Standard Setting (OSS) method in specifying the levels of proficiency in L2 academic writing. On the basis of the descriptive research design, the sample of this research includes the examinees and raters who were student teachers at the university level. Essay task and analytical writing scoring rubric were employed as the data collection tools. In data analysis, OSS method and two-step cluster analysis were used. The analysis results of OSS method based on many-facet Rasch measurement model (MFRM) outline the distribution of the criteria into the levels of proficiency. Also, the main findings in OSS method were validated with two-step cluster analysis. That is, OSS method may be practically used to help the stakeholders make objective judgments on the examinees' target performance.

1. INTRODUCTION

As a multidimensional field, assessing writing in an academic context has been the focus of attention in second language (L2) assessment in recent years. As an essential component of this field, standard setting serves as a basis for defining the levels of language attainment. Also, it provides evidence for decision makers to make instructional judgments on target performance. For this reason, there has been a growing interest in setting L2 academic writing standards over the years.

In broad terms, standard setting is viewed as "the process of determining cut-scores for a test" (Davies et al., 1999, p. 186). These cut-off scores may be single (e.g. pass/fail) or multiple (e.g. level of achievement) (Khatimin et al., 2013). In other words, "setting standards on educational assessments sometimes requires a single level" or "more than two stages or degrees of performance" (Cizek, 1993, p. 92-93). These single or multi-level standards have important

*CONTACT: Fatima Nur FISNE ✉ fatimanurfisne@gazi.edu.tr Gazi University, Gazi Faculty of Education, English Language Teaching Program, Turkiye

e-ISSN: 2148-7456 **/© IJATE** 2022

consequences on stakeholders such as test-takers, instructors, and policy-makers. For example, standard setting is taken into account in making judgments on the placement of the examinees into the appropriate levels (Shin & Lidster, 2017). Furthermore, standard setting procedure may directly influence the whole decision-making process in an educational system (Sondergeld et al., 2020). This aspect of standard setting is primarily related to the decision validity that represents the quality and consistency of the educational decisions (Erkus et al., 2017). Such a significant role of standard setting requires the use of objective methods in setting cut-off scores because the assessment results are open to discussion when the cut-off scores are not set properly (Bejar, 2008). However, Sireci et al. (1997) state that "the most popular methods for setting passing scores and other standards on educational tests rely heavily on subjective judgment" (p. 3). Likewise, Davis-Becker et al. (2011) reveal that standard setting is generally viewed as "one of the most subjective and judgmental components" in spite of the pivotal importance of standard setting "in the test development and validation process" (p. 25) Accordingly, there is a need for more objective methods to determine more valid standards in the educational measurement. In addition, the performance levels should be objectively defined in L2 writing assessment to help the stakeholders reach a valid decision.

In order to meet the needs of objectivity in standard setting procedures and ensure the decision validity in L2 writing assessment, this research mainly utilizes OSS method in defining the levels of target performance objectively, determining a valid cut-off score, and then making objective decisions about the students' performance in L2 academic writing.

1.1. Review of Literature

Standard setting basically refers to "setting cutscores" in assessment (Sireci et al., 1997, p. 3). More specifically, "performance standards specify what level of performance on a test is required for a test taker to be classified into a given category" and the process of defining these levels is called standard setting (Cizek, 2012, p. 4). As it functions as a benchmark to define target performance levels and provides a basis for performance-related decisions, it is an essential part of the educational assessment and evaluation.

There are various standard setting methods that are used to determine performance standards. These methods are basically grouped within two categories: test-based and examinee-based standard setting (Yudkowsky et al., 2009). In test-based methods like Angoff (1971) and Ebel (1972), judges examine the test itself and test items and predict the level of the target performance. On the other hand, in examinee-based methods like the Borderline Group and Contrasting-Groups, judges mainly focus on test takers' performance, gather evidence on the performance levels and then estimate the standards. Livingston and Zieky (1982) provide a comprehensive overview of the commonly used standard setting methods. To illustrate, Nedelsky method (1954), which is one of the earliest methods, is used to determine the passing score for multiple-choice tests. In this method, judges attempt to define the wrong answers that a borderline test taker would recognize. Calculations are carried out through the elimination of the possible wrong answers. In Angoff method (1971), unlike Nedelsky, judges examine each item holistically without considering the possible wrong answers and make estimations on whether borderline test takers would be able to give a correct answer to each item. Based on the probability of the correct answers, passing score is calculated. In Ebel method (1972), judges make decisions by considering the difficulty and relevance levels of the items. In this method, a matrix including the dimensions of the difficulty (i.e. easy, medium, hard) and relevance (essential, important, acceptable, questionable) is constructed, and test items are placed into the appropriate cells. Following that, judges predict the possible correct answers. Passing score is defined on the basis of the calculations including the percentage of correct answers. As for the examinee-based methods, the Borderline Group method focuses on test takers' performance and requires judges to identify the borderline test takers in terms of target

knowledge and skills. Passing score is set according to the median of the scores that are assigned to the borderline test takers. In the Contrasting-Group method, test takers are divided into two groups in consideration of their level of knowledge and skills, and passing score depends on the degree at which there is almost equal number of the test takers from both groups.

Sondergeld et al. (2020) assert that there are some concerns on the use of traditional standard settings methods, and to tackle these concerns, modern methods mainly based on item response theory (IRT) have been introduced. OSS method is one of these methods that aim to minimize the problems faced in setting standards like subjectivity and rater agreement/disagreement. It is basically established on test content rather than the direct expert opinions (Stone, 2001). Expert judgments are also used in this method, but the goal is not to specify the ratio/number of the correct responses; instead, experts discuss the essential content that might indicate the test takers' achievement (Sondergeld et al., 2020). As one of the modern criterion-based standard setting methods (Bichi et al., 2019), OSS method based on Rasch measurement model and Wright and Grosse's (1993) standard setting principles considers the expert opinions, test takers' performance and test/item difficulty at the same time (Khatimin et al., 2013). Through Rasch model, the measurement outputs are displayed on the logit scale, and the raw score can be analyzed on this scale in regard to the task/content achievement (Sondergeld et al., 2020). There are three important steps in OSS method: "defining the criterion set", "refining the criterion point", and "expressing the error" (Stone et al., 2011, p. 950). Hence, OSS method enables the examiners to analyze the level of performance in consideration of the standard error of measurement.

In the relevant literature, some research studies use and compare the standard setting methods, and examine the effectiveness and utility of these methods. For example, Davis-Becker et al. (2011) examined the Bookmark method in terms of item-ordering. Stone et al. (2011) compared OSS method and the Angoff approach on a longitudinal basis. In another study, Shin and Lidster (2017) discussed the comparative effectiveness of the Bookmark method, the Borderline group method, and cluster analysis in ESL (English as a Second Language) placement context. MacDougall and Stone (2015) emphasized the strengths of OSS method in standard setting procedure. In the research context of L2 writing assessment, some standard setting studies are related to the alignment of examinations to the Common European Framework of References (CEFR) that attempts "to describe the levels of proficiency required by existing standards, tests and examinations" (CoE, 2001, p. 21). In these studies, it is intended to link some language exams to the CEFR levels. For example, Tannenbaum and Wylie (2008) aimed to define the cut scores for two large-scale tests in accordance with the CEFR levels. Green (2018) investigated the English for Academic Purposes (EAP) context in terms of relating EAP testing to the CEFR. Fleckenstein et al. (2020) put emphasis on the writing profiles of students and tried to link EFL writing competences to the CEFR.

As a productive skill, writing encompasses different social, cultural and cognitive dynamics (Weigle, 2002). Owing to its dynamic structure, the assessment of L2 academic writing skills should be constructed on the systematic basis that entails the operationalization of the task characteristics and underlying dimensions. Harsch and Rupp (2011) call attention to the use of open-tasks in writing assessment and its advantages in enabling the examinees to produce a broad variety of written output. Use of these tasks in L2 writing assessment requires the attribution of levels or numerical values to target writing performance by the raters. In this respect, the rater-related issues are scrutinized in L2 writing assessment research. For example, Schaefer (2008) focused on the rater bias patterns in EFL writing assessment. The study findings pointed out that some criteria were severely rated whereas raters were lenient in some other criteria. Also, severity and leniency behaviours changed according to the students' level of ability in writing. Goodwin (2016) analyzed the rater behaviours in an academic language

test aiming at both reading and writing skills and found differences between the attributions of the scores in admission and placement tests. Trace et al. (2017) underlined the importance of rater negotiation and explicated its effect on reducing rater bias in writing performance assessment. Elder et al. (2007) paid attention to the rater subjectivity and bias. Along with the rater behaviours in L2 writing assessment, the presentation of objective and valid performance standards to the raters is another crucial issue. In some settings, assessing writing has large-scale outcomes like "promotion, placement, and admission" (Wind & Engelhard, 2013, p. 297), and therefore objective performance standards should be given to the raters in order to provide absolute and credible evidence for decision-makers. In this regard, the importance of standard setting in L2 writing assessment comes into prominence. From this perspective, this research aims to investigate the usefulness of OSS method in determining objective and valid cut-off scores and performance standards in L2 academic writing assessment. The following research questions guide the researchers to explore the utility of OSS method in setting objective standards in L2 writing assessment:

1. What are the procedures of setting objective standards in L2 academic writing assessment through OSS method?
2. To what extent are OSS method-based decisions validated?

2. METHOD

2.1. Research design and participants

This study adopts a descriptive research design that presents the researchers with opportunities to elaborate on the variables to be examined (Best & Khan, 2006). Within the framework of the descriptive research, this study attempts to explain how useful OSS Method is in defining the cut-off scores and standards of L2 academic writing proficiency. The subject group includes 64 raters and 39 examinees who were the student teachers in the department of English language teaching (ELT) at the tertiary level. The raters were the third graders taking Educational Measurement course, and they were familiar with not only the process of academic writing but also the assessment of writing performance. Since the number of raters plays an essential role in standard setting procedures, the 3rd grade student teachers (n = 64) were selected as the participants with the practical purposes. With respect to the demographics of the raters, the mean age was 21.84. While 12 raters (18.78%) were male, 52 of the raters (81.25%) were female. As for the examinees, they were the first graders attending Advanced Reading and Writing course II at the same department. They completed Advanced Reading and Writing course I in the fall term and reinforced their knowledge and skills on how to develop outlines, specify topic, thesis and supporting sentences, sequence their ideas in a logical way, and ensure task achievement.

2.2. Data Collection

In this research, there are two sequential steps in the data collection. First, a sample writing task of IELTS (The International English Language Testing System)* was used. This task requires the examinees to write an essay by expressing agreement or disagreement on the given topic. This sample task was selected owing to the authenticity of the topic in which the examinees might address their real experiences. After they completed the task, in the second step, the essays were anonymously distributed to the raters, and each rater scored all the essays individually. With the aim of scoring the academic essays, Analytical Scoring Rubric for Academic Essays (ASRAE) was developed by the researchers (see Appendix A). This rubric

* This task was taken from the section of Sample Test Questions/Academic Writing on the official web page of IELTS (https://www.ielts.org/)

was also used in a different study conducted by the researchers (Sata & Karakaya, 2021). As explained in this study, ASRAE includes seven main criteria and 16 sub-criteria that intend to measure the components of academic writing. The main logic behind the development of this rubric is the characteristics of the target participants. Since the examinees were highly proficient in L2 and attained a mastery level in academic writing, the researchers felt the necessity to develop such kind of a rubric. The content validity of ASRAE was ensured through analyzing the essays, reviewing the literature review and calculating content validity index and ratio proposed by Lawshe (1975). For construct validity, exploratory factor analysis (EFA) was conducted, and EFA results indicated that the explained total variance was .73 with one-factor structure. In what follows the validation of the content and construct, the reliability of the rubric was determined. For this calculation, the reliability coefficient (ω) suggested by McDonald (1999) was employed, and the results show that McDonald ω coefficient was .97 (95% reliability interval: .96-.98) as elucidated in detail in Sata and Karakaya (2021). To sum up, validity and reliability results point out that ASRAE can be used as a reliable and valid tool to measure L2 academic writing proficiency.

2.3. Data Analysis

In order to analyze the rater scoring, set objective standards for academic writing proficiency, and validate these standards, OSS method based on MFRM and two-step cluster analysis were used in the current research. There were three main facets used in the Rasch analysis for OSS method: raters, examinees, and criteria. The raters scored each individual essay by considering the criteria given in ASRAE. So, fully crossed design was employed in this analysis. In line with three steps given in OSS method (Stone et al., 2011), four steps were followed in this analysis: (1) ensuring content validity, (2) specification of the performance levels, (3) difficulty level and standard errors, and (4) determination of proficiency levels.

OSS method requires meeting some assumptions of MFRM such as unidimensionality, local dependence, and model-data fit. With the aim of ensuring these assumptions, some analyses were conducted. Firstly, EFA was employed to test the unidimensionality, and the EFA results display that the factor structure is unidimensional. Following that, G2 statistics (Chen & Thissen, 1997) was used to test the local dependence. The results point out that LD $\chi 2$ values estimated for each criterion pairs are below 10. This result could be viewed as the indicator of the local dependence. As for the assumption of the model-data fit, the standardized values were examined. Linacre (2017) states that in order to ensure model-data fit, the number of the standardized values which are not between -2 and +2 should not exceed 5% of all the data. In this research, the number of total data was 37396, and the number of the standardized values that are not between -2 and +2 is 1547 [%4.14]). It is seen that there is a fit between model and data. Besides that, it is also crucial to examine the fit values of the target items (Khatimin et al., 2013). Accordingly, biserial correlation, outfit values, and standards of outfit values were examined to identify the misfit items (see Appendix B). According to the results, biserial correlations(x) are between .17 and .51, outfit values (MNSQ) are between 0.71 and 1.38, and the standards of outfit values (ZSTD) are between -9.00 and 9.00 (fit values: $0.4<x<0.8$, $0.5<MNSQ<1.5$ and $-2.0<z<2.0$). That is to say, there is no misfit in the dataset except for the standards of the outfit values. Holistically speaking, all the assumptions of OSS method were tested and ensured. It is noteworthy to state that the criterion of "Title of Essay" was excluded from the analysis of the OSS method since most examinees did not write a title for their essays unintentionally, and this exceptional case might cause an invalid standard setting. On the other hand, "Title of Essay" is still the component of the rubric (see ASRAE in Appendix A).

As the second data analysis method, two-step cluster analysis was conducted to provide evidence on the validity of the performance standards to be set through OSS method because Khalid (2011) explains that clustering analysis is based on less subjective process. The

avoidance of subjectivity is the primary rationale to choose cluster analysis as the validation tool of OSS-method results. Cokluk et al. (2012) explain the main function of cluster analyzing as the identification of the similarities among the items/examinees and classification of them according to these similarities. To be more specific, cluster analysis can categorize target groups according to "distance" and "similarity" (Violato et al., 2003, p. 62). That is why this technique was selected to clarify and confirm OSS method results. The important assumptions to be met in the cluster analysis are the representativeness of the universe and avoidance of multicollinearity problem and outliers (Kayri, 2007). In this study, there are not any multicollinearity problems between variables and outliers in the datasets. However, larger samples may be required to offer more representativeness for the universe. So, the analysis results will be discussed in the target sample of the participants in this research.

3. RESULTS

This section elaborates on the standard setting procedures in L2 writing assessment through OSS method, identification of the cut-off score, and then presents the consistency between the results of OSS Method and two-step cluster analysis.

3.1. Standard Setting through OSS Method

In line with the first step given in OSS method, the content validity of ASRAE or definition of the criteria/content was ensured through expert opinions. 11 experts evaluated the appropriateness of the criteria on the basis of the rubric construct and components. They were asked to decide whether or not the criteria are essential. According to the expert judgments, content validity ratio (CVR) and index were calculated (Lawshe, 1975). CVR was reported as .75 and this value is above .59 that indicates the evidence for content validity with 11 experts (Wilson et al., 2012). The results show that the rubric has the content validity at the expected level.

Table 1. *Distribution of the criteria to the specified criterion points*

Criterion Points	Criteria	Logit Value	Standard Error
Criterion Point 1	Syntactic Complexity	0.43	0.02
	Idea Development	0.35	0.02
	Topic Sentence	0.34	0.02
	Lexical Range	0.33	0.02
Criterion Point 2	Thesis Statement	0.30	0.02
	Supporting Sentence	0.28	0.02
	Linking	0.24	0.02
Criterion Point 3	Accuracy of Grammatical Forms	0.01	0.03
	Coherence	-0.01	0.03
	Introduction-Body-Conclusion	-0.07	0.03
	Word Choice	-0.07	0.03
Criterion Point 4	Topic Relevance	-0.34	0.03
	Appropriate Length	-0.49	0.03
	Punctuation	-0.59	0.03
	Spelling	-0.70	0.03

The second step guides the specification of the performance levels in L2 academic writing. In this respect, the field experts suggested five levels of academic writing in consideration of the CEFR as A2, B1, B2, C1, and C2. The reason why the level of A1 is not included in this specification is that the examinees, who were student teachers in ELT department, had L2 writing experiences and had been practicing English language writing for a long time. In accordance with these five proficiency levels, four criterion points were defined for the analysis through OSS Method. When the number of the criteria was divided by the number of the criterion points (15/4 = 3.75), the number of the criteria to be assigned to each level was found as 3.75. That is, each level requires the proficiency almost in four criteria. Table 1 illustrates the distribution of the rubric criteria to the criterion points that are specified above. In the third step of OSS method, the mean difficulty levels and mean standard errors were calculated for each criterion point. These difficulty levels and standard errors are given in Table 2. In this table, negative logit values represent the criteria that are relatively easy for the examinees to achieve. On the other hand, positive logit values indicate relatively more difficult criteria in L2 academic writing assessment.

Table 2. *Mean difficulty and mean standard errors of the criterion points.*

Criterion Point	Mean Logit Value	Mean Standard Error
Criterion Point 1	+0.36	0.02
Criterion Point 2	+0.27	0.02
Criterion Point 3	-0.04	0.03
Criterion Point 4	-0.53	0.03

After the calculation of difficulty and standard errors, the cut-off score was estimated in the relevant data set. Khatimin et al. (2013) put forward that the examinees will be accepted as successful if they complete at least 60% of the task. The value of sixty-percent means the achievement of 9 criteria in ASRAE (15 x (60 / 100) = 9). In ASREA, when all the criteria are successively ordered in terms of the difficulty level, the logit value of the ninth criterion corresponds this value; in other words, the logit value .24 is accepted as the cut-off score (see Table 1).

When Table 3 is examined, it is seen that cut-off score (+0.24 logit) and criterion point 2 are in the same order. To find out the confidence interval of the cut-off score, standard error (0.03) was multiplied by ±1.96. It is seen that the confidence interval with 95% is between +0.18 and +0.30. This proves that calculated cut-off score is at confidence interval. Also, as given in Table 3, there is no more option for the cut-off score apart from +0.24 logit value because there is no logit value at confidence interval except for +0.24 logit. Table 3 provides information about the cut-off score, criterion points and the levels of the academic writing proficiency. Accordingly, out of 39 examinees, 15 examinees had high proficiency in L2 academic writing whereas 24 examinees had low proficiency in the same skill. In the final step, the examinees' proficiency levels were determined in consonance with the performance standards. Table 4 illustrates these levels and descriptive statistics. It can be seen that Level 5 and Level 2 have high frequencies. That is to say, the examinees can be holistically divided into two main groups in academic writing proficiency.

Table 3. *Estimation of the cut-off score and levels of academic writing proficiency.*

	Examinee	Observed Average	Fair-M Average	Logit Measure	Standard Error	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	
	S16	3.58	3.61	1.45	0.06	1.02	0.40	1.05	0.90	
	S21	3.47	3.51	1.15	0.05	1.19	3.50	1.26	4.60	
	S24	3.44	3.48	1.07	0.05	1.01	0.10	1.09	1.70	
	S28	3.39	3.43	0.95	0.05	0.88	-2.60	0.88	-2.50	
	S29	3.36	3.40	0.89	0.05	0.88	-2.50	0.87	-2.70	
	S15	3.34	3.38	0.85	0.05	0.90	-2.20	0.98	-0.40	
	S17	3.32	3.36	0.81	0.05	1.06	1.20	1.12	2.40	
	S37	3.29	3.33	0.74	0.05	1.01	0.20	1.04	0.90	
	S19	3.29	3.33	0.74	0.05	0.96	-0.80	1.01	0.10	
	S03	3.21	3.25	0.58	0.04	1.24	4.80	1.27	5.30	
	S34	3.19	3.23	0.56	0.04	0.94	-1.20	0.97	-0.50	
	S26	3.17	3.21	0.51	0.04	0.95	-1.00	0.97	-0.70	
	S01	3.11	3.15	0.41	0.04	0.99	-0.20	1.01	0.30	
CP-1	S13	3.11	3.15	0.40	0.04	0.86	-3.10	0.86	-3.20	
CP-2	S12	3.05	3.09	0.30	0.04	1.09	1.90	1.13	2.70	CS
	S31	2.88	2.92	0.03	0.04	0.68	-8.00	0.69	-7.70	
	S07	2.88	2.91	0.02	0.04	1.34	6.90	1.40	8.10	
	S25	2.87	2.90	0.00	0.04	0.76	-5.70	0.78	-5.40	
	S23	2.84	2.88	-0.03	0.04	0.74	-6.50	0.74	-6.30	
CP-3	S02	2.84	2.87	-0.04	0.04	0.93	-1.60	0.92	-1.90	
	S22	2.83	2.86	-0.05	0.04	0.84	-3.80	0.86	-3.20	
	S32	2.83	2.86	-0.05	0.04	0.97	-0.60	0.96	-0.80	
	S27	2.81	2.84	-0.08	0.04	0.72	-6.80	0.73	-6.80	
	S14	2.80	2.84	-0.09	0.04	1.01	0.20	1.04	1.00	
	S11	2.73	2.76	-0.21	0.04	0.99	-0.20	0.98	-0.40	
	S35	2.68	2.71	-0.27	0.04	1.10	2.20	1.10	2.30	
	S20	2.67	2.70	-0.29	0.04	0.89	-2.50	0.92	-1.90	
	S36	2.53	2.56	-0.48	0.04	0.75	-6.20	0.77	-5.60	
	S30	2.52	2.55	-0.49	0.04	0.93	-1.50	0.96	-1.00	
	S18	2.51	2.54	-0.50	0.04	0.89	-2.50	0.89	-2.50	
	S33	2.51	2.54	-0.51	0.04	0.97	-0.60	0.99	-0.30	
CP-4	S05	2.49	2.52	-0.53	0.04	1.08	1.70	1.13	2.90	
	S09	2.33	2.36	-0.74	0.04	0.95	-1.00	0.95	-1.00	
	S10	2.23	2.25	-0.87	0.04	1.47	9.00	1.50	9.00	
	S08	2.07	2.08	-1.07	0.04	0.78	-5.60	0.78	-5.40	
	S06	2.02	2.03	-1.14	0.04	1.01	0.10	1.01	0.20	
	S38	1.89	1.89	-1.30	0.04	1.38	8.10	1.39	8.40	
	S39	1.86	1.86	-1.34	0.04	1.37	7.90	1.38	8.10	
	S04	1.83	1.83	-1.37	0.04	1.16	3.70	1.17	3.80	

CP-1 (Criterion Point 1); CP-2 (Criterion Point 2); CP-3 (Criterion Point 3); CP-3 (Criterion Point 3), CS (Cut Score) Cut score and Criterion Point 2 are at the same line.

Table 4. *The levels of proficiency and descriptive statistics.*

Achievement Levels	Frequency	Percentage	Mean	Std. Deviation
Level 5 (0.36 - the highest logit)	14	35.90	0.79	0.30
Level 4 (0.27 and 0.35 logit)	1	2.56	0.30	--
Level 3 (-0.04 and 0.26 logit)	5	12.82	-0.01	0.03
Level 2 (-0.53 and -0.05 logit)	12	30.77	-0.30	0.20
Level 1 (-0.54 the lowest logit)	7	17.95	-1.12	0.24

3.2. Two-step Cluster Analysis Results

The mean of the scores that the raters assigned for each examinee was used in two-step cluster analysis. The analysis results highlight the existence of two clusters (the quality of clustering: 0.67, and Silhoutte coefficient: 0.58). With respect to the placement of the examinees to these clusters, it can be understood that two clusters show consistency with two groups divided by the cut-off score in OSS method. Put it another way, the first cluster includes the examinees with high proficiency in academic writing, and the second cluster includes the examinees with low proficiency. Therefore, two-step cluster analysis confirms and validates the findings in OSS method. Table 5 shows the comparative results of two-step cluster analysis and OSS Method.

Table 5. *Comparison of OSS method and two-step cluster analysis.*

Two-step Cluster Analysis			OSS Method		
Silhoutte Coefficient	Rank of Cluster Analysis	Cluster	Rank in OSS Method	Logit Value	Proficiency
0.723	S16	1	S16	1.45	High
0.788	S21	1	S21	1.15	High
0.807	S24	1	S24	1.07	High
0.833	S28	1	S28	0.95	High
0.841	S29	1	S29	0.89	High
0.845	S15	1	S15	0.85	High
0.846	S17	1	S17	0.81	High
0.843	S19	1	S37	0.74	High
0.824	S37	1	S19	0.74	High
0.814	S03	1	S03	0.58	High
0.802	S34	1	S34	0.56	High
0.776	S26	1	S26	0.51	High
0.698	S01	1	S01	0.41	High
0.696	S13	1	S13	0.40	High
0.566	S12	1	S12	0.30	High
0.070	S31	2	S31	0.03	Low
0.108	S07	2	S07	0.02	Low
0.147	S25	2	S25	0.00	Low
0.227	S23	2	S23	-0.03	Low
0.257	S02	2	S02	-0.04	Low
0.276	S22	2	S22	-0.05	Low
0.278	S32	2	S32	-0.05	Low
0.321	S27	2	S27	-0.08	Low
0.337	S14	2	S14	-0.09	Low
0.465	S11	2	S11	-0.21	Low

Table 5. *Continues*

0.528	S20	2	S35	-0.27	Low
0.562	S35	2	S20	-0.29	Low
0.623	S30	2	S36	-0.48	Low
0.628	S18	2	S30	-0.49	Low
0.629	S33	2	S18	-0.50	Low
0.631	S05	2	S33	-0.51	Low
0.631	S36	2	S05	-0.53	Low
0.624	S09	2	S09	-0.74	Low
0.612	S10	2	S10	-0.87	Low
0.586	S08	2	S08	-1.07	Low
0.575	S06	2	S06	-1.14	Low
0.545	S38	2	S38	-1.30	Low
0.538	S39	2	S39	-1.34	Low
0.529	S04	2	S04	-1.37	Low

4. DISCUSSION and CONCLUSION

"Standard setting involves judgments about the ideal performance standard and test score that reflect this standard" (Hsieh, 2013). It has important consequences for the stakeholders such as students, teachers, and policy-makers in different areas (Fulcher, 2013; Shin & Lidster, 2017; Sondergeld et al., 2020; Stone et al., 2011). However, standard setting methods may include subjective evaluation and judgments (Davis-Becker et al., 2011). Considering this perspective, the current research study employed OSS method in order to provide objective and valid cut-off scores and performance standards for the stakeholders. In this way, it was attempted to establish a basis for making credible and valid decisions on L2 academic writing.

OSS method is based on item response theory and Rasch model and analyzes the data at item/rater/difficulty level. Stone et al. (2011) put emphasis on three important points in OSS method: "defining criterion set", "refining criterion point", and "expressing error" (p. 950). This study adopted these perspectives and set performance standards in four steps. Firstly, the content of criterion set was defined in line with the review of literature and validated in light of the expert opinions. 7 main criteria and 16 sub-criteria were specified in accordance with the feedback received from the experts. In the second step, the performance levels were determined with reference to the CEFR, and criterion points were defined in consonance with these levels. In the following step, mean difficulty levels and standard errors were calculated. Then the cut-off score was estimated by considering the task achievement level accentuated in Khatimin et al. (2013). It was found that the cut-off score (+0.24 logit) and the criterion point 2 were at the same line. Besides that, the confidence interval at which the cut-off score could be placed was found. The cut-off score divided the examinees into two groups with high proficiency (n = 15) and low proficiency (n = 24) in L2 academic writing. Finally, L2 academic writing proficiency, levels of the examinees and basic descriptive statistics were presented. The validity of OSS method results, especially the cut-off score, was confirmed by two-step cluster analysis which is based on less subjectivity (Khalid, 2011). Two-step cluster analysis results pointed out the emergence of two clusters, and the same examinees at high proficiency level and low proficiency level were successively placed into these two clusters. Therefore, it can be concluded that OSS method facilitates the specification of objective performance standards and valid cut-off scores in L2 academic writing assessment. MacDougall and Stone (2015) and Stone et al. (2011) found that OSS method was effective in the construct development in the target area when compared to other standard setting methods. In this research, it can be seen that OSS method serves as an objective basis for setting performance standards in L2 academic

writing, specifying the valid cut-off score, and making reliable and objective decisions about L2 writing performance.

With regard to limitations of the study, the results may appear to be lack of generalizability in L2 writing assessment context due to the fact that this research was carried out with a specific subject group. Further research may focus on larger samples including more raters in different educational settings. Thus, performance standards and cut-off scores may be validated in various contexts. Another limitation of the study is related to the rater training. The raters in this study did not have any professional training about how to rate language skills. So, this standard setting study may be replicable by proving the raters with sufficient training about how to rate academic writing performance. As further research studies, different standard setting methods may be compared in terms of the objectivity in L2 academic writing assessment. This comparison may give more concrete evidence on the utility of performance standards. This study also suggests the use of OSS method in standard setting studies for different language skills.

Acknowledgments

The preliminary findings of this study were presented at an international conference titled "6th International Congress on Measurement and Evaluation in Education and Psychology" in September, 2018. The extended abstract was published in the abstract booklet

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research and publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s). Ethics Committee Approval and its number should be given by stating the institution name which gave the ethical approval. Ethics Committee Number: Gazi University, 80287700-302.08.01-54466.

Authorship Contribution Statement

Fatima Nur Fisne: Introduction, Review of Literature, Methodology (Data Collection, Instrument Development), Discussion and Conclusion. **Mehmet Sata**: Methodology (Instrument Development, Data Collection, Data Analysis), Results. **Ismail Karakaya:** Supervision.

Orcid

Fatima Nur FISNE https://orcid.org/0000-0001-9224-2485
Mehmet SATA https://orcid.org/0000-0003-2683-4997
Ismail KARAKAYA https://orcid.org/0000-0003-4308-6919

REFERENCES

Bejar, I.I. (2008). Standard setting: What is it? Why is it important*? R&D Connections, 7*, 1-6.

Best, J.W., & Khan, J.V. (2006). *Research in Education (10th Edition).* Pearson.

Bichi, A.A., Talib, R., Embong, R., Mohamed, H. B., Ismail, M. S., & Ibrahim, A. (2019). Rasch-based objective standard setting for university placement test. *Eurasian Journal of Educational Research*, *19*(84), 57-70. https://doi.org/10.14689/ejer.2019.84.3

Chen, W.H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*(3), 265-289. https://doi.org/10.3102/10769986022003265

Cizek, G.J. (1993). Reconsidering standards and criteria. *Journal of Educational Measurement, 30*(2), 93-106. https://doi.org/10.1111/j.1745-3984.1993.tb01068.x

Cizek, G.J. (Ed.). (2012). An introduction to contemporary standard setting: concepts, characteristics, and concepts. *In Setting performance standards: Concepts, methods, and perspectives.* Mahwah, NJ: Lawrence Erlbaum Associates.

Council of Europe [CoE]. (2001). *Common European framework of reference for languages: Learning, teaching, assessment.* Cambridge, England: Cambridge University Press.

Cokluk, O., Sekercioglu, G., & Buyukozturk, S. (2012). *Sosyal bilimler icin cok degiskenli istatistik: SPSS ve LISREL uygulamalari* (2nd edition) [Multivariate statistics for social sciences: SPSS and LISREL applications], Pegem Akademi.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Studies in language testing 7: Dictionary of language testing.* Cambridge University Press.

Davis-Becker, S.L., Buckendahl, C.W., & Gerrow, J. (2011). Evaluating the bookmark standard setting method: The impact of random item ordering. *International Journal of Testing, 11*(1), 24-37. https://doi.org/10.1080/15305058.2010.501536

Elder, C., Barkhuizen, G., Knoch, U., & Von Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing, 24*(1), 37-64. https://doi.org/10.1177/0265532207071511

Erkus, A., Sunbul, O., Omur-Sunbul, S., Yormaz, S., & Asiret, S. (2017). *Psikolojide olcme ve olcek gelistirme-II* (1st edition) [Measurement in psychology and scale development-II], Pegem Akademi.

Fleckenstein, J., Keller, S., Krüger, M., Tannenbaum, R.J., & Köller, O. (2020). Linking TOEFL iBT® writing rubrics to CEFR levels: Cut scores and validity evidence from a standard setting study. *Assessing Writing, 43,* 1-15. https://doi.org/10.1016/j.asw.2019.100420

Fulcher, G. (2013). *Practical language testing*. Routledge. https://doi.org/10.4324/9802037673 99

Goodwin, S. (2016). A Many-Facet Rasch analysis comparing essay rater behavior on an academic English reading/writing test used for two purposes. *Assessing Writing, 30*, 21-31. https://doi.org/10.1016/j.asw.2016.07.004

Green, A. (2018). Linking tests of English for academic purposes to the CEFR: The score user's perspective. *Language Assessment Quarterly, 15*(1), 59-74. https://doi.org/10.1080/15434303.2017.1350685

Harsch, C., & Rupp, A.A. (2011). Designing and scaling level-specific writing tasks in alignment with the CEFR: A test-centered approach. *Language Assessment Quarterly, 8*(1), 1-33. https://doi.org/10.1080/15434303.2010.535575

Hsieh, M. (2013). An application of multifaceted Rasch measurement in the Yes/No Angoff standard setting procedure. *Language Testing, 30*(4), 491-512. https://doi.org/10.1177/0265532213476259

IELTS (The Internatinal English Language Testing System). https://www.ielts.org/

Kayri, M. (2007). Two-step clustering analysis in researches: A case study. *Eurasian Journal of Educational Research (EJER), 28*, 89-99.

Khalid, M. N. (2011). Cluster analysis-a standard setting technique in measurement and testing. *Journal of Applied Quantitative Methods, 6*(2), 46-58.

Khatimin, N., Aziz, A.A., Zaharim, A., & Yasin, S.H.M. (2013). Development of objective standard setting using Rasch measurement model in Malaysian institution of higher learning. *International Education Studies, 6(*6), 151-160. https://doi.org/10.5539/ies.v6n6p151

Lawshe, C.H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*(4), 563-575. https://doi.org/10.1111/j.1744-6570.1975.tb01393.x

Linacre, J.M. (2017). *A user's guide to FACETS: Rasch-model computer programs.* Chicago: MESA Press.

Livingston, S.A., & Zieky, M.J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests.* Educational Testing Service: New Jersey.

McDonald, R.P. (1999). *Test theory: A unified approach.* Mahwah, NJ: Erlbaum.

MacDougall, M., & Stone, G.E. (2015). Fortune-tellers or content specialists: Challenging the standard setting paradigm in medical education programmes. *Journal of Contemporary Medical Education, 3*(3), 135. https://doi.org/10.5455/jcme.20151019104847

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing, 25*(4), 465-493. https://doi.org/10.1177/0265532208094273

Shin, S.Y., & Lidster, R. (2017). Evaluating different standard-setting methods in an ESL placement testing context. *Language Testing, 34*(3), 357-381. https://doi.org/10.1177/0265532216646605

Sireci, S.G., Robin, F., & Patelis, T. (1997). Using cluster analysis to facilitate standard setting. *Applied Measurement in Education, 12*(3), 301-325. https://doi.org/10.1207/S15324818AME1203_5

Sondergeld, T.A., Stone, G.E., & Kruse, L.M. (2020). Objective standard setting in educational assessment and decision making. *Educational Policy, 34*(5), 735-759. https://doi.org/10.1177/0895904818802115

Stone, G.E. (2001). Objective standard setting (or truth in advertising). *Journal of Applied Measurement, 2*(2), 187-201.

Stone, G.E., Koskey, K.L., & Sondergeld, T.A. (2011). Comparing construct definition in the Angoff and Objective Standard Setting models: Playing in a house of cards without a full deck. *Educational and Psychological Measurement, 71*(6), 942-962. https://doi.org/10.1177/0013164410394338

Sata, M. & Karakaya, I. (2021). Investigating the effect of rater training on differential rater function in assessing academic writing skills of higher education students. *Journal of Measurement and Evaluation in Education and Psychology, 12*(2), 163-181. https://doi.org/10.21031/epod.842094

Tannenbaum, R.J., & Wylie, E.C. (2008). Linking English-language test scores onto the common European framework of reference: An application of standard-setting methodology. *ETS Research Report Series, 2008*(1), i-75. https://doi.org/10.1002/j.2333-8504.2008.tb02120.x

Trace, J., Janssen, G., & Meier, V. (2017). Measuring the impact of rater negotiation in writing performance assessment. *Language Testing, 34*(1), 3-22. https://doi.org/10.1177/0265532215594830

Violato, C., Marini, A., & Lee, C. (2003). A validity study of expert judgment procedures for setting cutoff scores on high-stakes credentialing examinations using cluster analysis. *Evaluation & The Health Professions, 26*(1), 59-72. https://doi.org/10.1177/0163278702250082

Weigle, S.C. (2002). *Assessing writing.* Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511732997

Wilson, F.R., Pan, W., & Schumsky, D.A. (2012). Recalculation of the critical values for Lawshe's content validity ratio. *Measurement and Evaluation in Counseling and Development, 45*(3), 197-210. https://doi.org/10.1177/0748175612440286

Wind, S.A., & Engelhard Jr, G. (2013). How invariant and accurate are domain ratings in writing assessment? *Assessing Writing, 18(*4), 278-299.

Wright, B.D., & Grosse M. (1993). How to set standards. *Rasch Measurement Transactions, 7*(3), 315-316.

Yudkowsky, R., Downing, S. M., & Tekian, A. (2009). Standard setting. In R. Yudkowsky & S. Downing (Ed.), *Assessment in health professions education* (pp. 86-105). Routledge. https://doi.org/10.4324/9781315166902-6

APPENDIX

Appendix A

ANALYTIC WRITING SCORING RUBRIC FOR ACADEMIC ESSAYS

	ORGANIZATION						CONTENT	
Point	**Title of Essay**	**Introduction-Body-Conclusion**	**Thesis Statement**	**Topic Sentence**	**Supporting Sentences**	**Appropriate Length**	**Topic Relevance**	**Idea Development**
4	Title of essay *comprehensively* represents the focus of the written text. It is *highly* relevant to the task.	The organization of introduction, body, and conclusion paragraphs is *highly* appropriate to written genre.	Thesis statement is *noticeably* given in introduction paragraph. It *comprehensively* includes the specific idea(s) to be elaborated in the written text.	Topic sentence *comprehensively* addresses and supports the specific idea(s) given in thesis statement. It *extensively* demonstrates the main idea of the paragraph.	Supporting sentences *comprehensively* illustrate the main idea given in topic sentence.	There are *at least 250 words* in written text. It is constructed with *appropriate length*.	Written text is *highly* relevant to assigned topic in task. It *comprehensively* addresses all parts of the task.	*Extensive* details are provided to develop, support and illustrate information or ideas presented in written text.
3	Title of essay *adequately* represents the focus of the written text. It is relevant to the task.	The organization of introduction, body, and conclusion paragraphs is *largely* appropriate to written genre.	Thesis statement is *evidently* given in introduction paragraph. It *mostly* includes the specific idea(s) to be elaborated in the written text.	Topic sentence *mostly* addresses and supports the specific idea(s) given in thesis statement. It *largely* demonstrates the main idea of the paragraph.	Supporting sentences *adequately* illustrate the main idea given in topic sentence.	Text length is between *200 and 249 words*. It is *slightly* shorter than required length.	Written text is *mostly* relevant to assigned topic in task. It *adequately* addresses the basic parts of the task.	*Adequate* details are provided to develop, support and illustrate information or ideas presented in written text.

2	Title of essay *moderately* represents the focus of the written text. It is relevant to the task in *some respects.*	The organization of introduction, body, and conclusion paragraphs is *moderately* appropriate to written genre.	Thesis statement is *less explicitly* given in introduction paragraph. It *moderately* includes the specific idea(s) to be elaborated in the written text.	Topic sentence *moderately* addresses and supports the specific idea(s) given in thesis statement. It demonstrates the main idea of the paragraph in *some respects.*	Supporting sentences *moderately* illustrate the main idea given in topic sentence.	Text length is between 150 *and 199 words*. It is *seemingly* shorter than required length.	Written text is *moderately* relevant to assigned topic in task. It *partially* addresses the basic parts of task.	*Basic* details are provided to develop, support and illustrate information or ideas presented in written text.
1	Title of essay *slightly* represents the focus of the written text. It is *partially* relevant to the task.	There is *inadequate* organization of introduction, body, and conclusion paragraphs in the written text.	Thesis statement is *vaguely* given in introduction paragraph. It *slightly* includes the specific idea(s) to be elaborated in the written text.	Topic sentence *partially* addresses and supports the specific idea(s) given in thesis statement. It *slightly* demonstrates the main idea of the paragraph.	Supporting sentences *partially* illustrate the main idea given in topic sentence.	Text length is between 100 *and 149 words*. It is *considerably* shorter than required length.	Written text is *slightly* relevant to assigned topic in task. It lacks addressing the basic parts of the task.	*Some details are* provided but they are not enough to develop, support and illustrate information or ideas presented in written text.
0	Written text does not include a title or title of essay is *completely* irrelevant.	Written text lacks organization of introduction, body and conclusion paragraphs.	Thesis statement is not given in introduction paragraph or it does not include any specific idea(s) to be elaborated in the written text.	Topic sentence is not included in written text, or it does not address the thesis statement or demonstrate the main idea of the paragraph.	Written text does not include supporting sentences or they do not illustrate the main idea given in topic sentence.	Text length is *below 99 words.* It does not meet the requirement of appropriate length.	Written text is irrelevant to assigned topic in task. It fails to address the task adequately.	Information or ideas are not *thoroughly* developed, supported or illustrated in written text.

	COHERENCE	COHESION	GRAMMAR		VOCABULARY		MECHANICS	
Point	**Coherence**	**Linking**	**Accuracy of Grammatical Forms**	**Syntactic Complexity**	**Word Choice**	**Lexical Range**	**Spelling**	**Punctuation**
4	Information or ideas sequenced in paragraphs are *highly* consistent. There is a *considerably* logical progression between sentences in written text.	A *wide* range of cohesive devices used to connect ideas in written text provides a smooth transition between sentences.	All grammatical forms are *accurately* used in written text. The communication is *successfully* established.	Complex and sophisticated sentences are *extensively* used in written text in which syntactic structures are *highly* diverse.	All the words and phrases are *appropriately* used. The intended meaning is *clearly* conveyed in written text.	There is a *wide range* of vocabulary used in written text which includes *highly* sophisticated words and phrases.	All the needed spelling rules are *accurately* used in written text.	All the needed punctuation rules are *accurately* used in written text.
3	Information or ideas sequenced in paragraphs are *mostly* consistent. There is an *adequately* logical progression between sentences in written text.	An *adequate* range of cohesive devices used to connect ideas in written text provides an easy transition between sentences.	The use of the grammatical forms is *mostly accurate* in the written text. There are *few grammatical errors* which do not impede communication.	Complex and sophisticated sentences are *widely* used in written text in which syntactic structures are *adequately* diverse.	The use of words and phrases is *mostly appropriate*. There are *few* misused words or phrases which cannot obscure the intended meaning.	There is an *adequate range* of vocabulary used in written text which includes *largely* sophisticated words and phrases.	All the needed spelling rules are *mostly accurate* in written text but there are *few errors* which violate these rules.	All the needed punctuation rules are *mostly accurate* in written text but there are few errors which violate these rules.
2	Information or ideas sequenced in paragraphs are *moderately* consistent but there are some inconsistencies which *partially* interrupt logical progression between sentences.	The use of cohesive devices *at basic level* to connect ideas in written text provides a complete transition between sentences.	It is attempted to use the grammatical forms accurately in written text but there are *occasional grammatical errors* which slightly impede communication.	Complex and sophisticated sentences are *moderately* used in written text in which syntactic structures are *partially* diverse.	It is attempted to use the words and phrases appropriately but there are *occasionally* misused words or phrases which *slightly* obscure the intended meaning.	The *basic* vocabulary is used in written text which includes *moderately* sophisticated words and phrases.	It is intended to use the needed spelling rules *accurately* in written text but there are *occasional errors* which violate these rules.	It is intended to use the needed punctuation rules *accurately* in written text but there are *occasional errors* which violate these rules.

1	Paragraphs are constructed with *slightly* consistent information or ideas which interrupt logical progression and sequence between sentences.	A *limited* range of cohesive devices used to connect ideas in written text makes transition between sentences fragmentary.	The use of the grammatical forms is *generally inaccurate* in written text. There are *frequent grammatical errors* which largely impede communication.	Complex and sophisticated sentences are *slightly* used in written text in which syntactic structures are diverse to some extent.	The use of words and phrases is *generally inappropriate*. There are *frequentl*y misused words or phrases which *largely* obscure the intended meaning.	There is a *limited range* of vocabulary used in written text which includes *slightly* sophisticated words and phrases.	The use of the needed spelling rules is *largely* inaccurate. There are *frequent errors* which violate these rules.	The use of the needed punctuation rules is *largely* inaccurate. There are *frequent errors* which violate these rules.
0	Written text lacks consistency and logical progression between sentences.	There is an *inadequate* use of cohesive devices in written text which lacks transition between sentences.	The use of grammatical forms is *completely inaccurate* in the written text. This causes a breakdown in communication.	Written text lacks sentential complexity, sophistication and syntactic variety.	The use of vocabulary is completely inappropriate in written text. The intended message is obscured.	A repetitive vocabulary is largely used in written text which lacks sophistication.	All the needed spelling rules are *inaccurately* used in written text.	All the needed punctuation rules are *inaccurately* used in written text.

Appendix B

Fit values of the rubric criteria

Criteria	Logit value	Standard Error	Infit	ZStd	Outfit	ZStd	Correlation
Syntactic Complexity	0.43	0.02	0.72	-9.00	0.73	-9.00	0.40
Idea Development	0.35	0.02	0.71	-9.00	0.72	-9.00	0.48
Topic Sentence	0.34	0.02	1.16	5.60	1.13	4.70	0.47
Lexical Range	0.33	0.02	0.70	-9.00	0.71	-9.00	0.41
Thesis Statement	0.30	0.02	1.40	9.00	1.37	9.00	0.43
Supporting Sentence	0.28	0.02	0.80	-7.60	0.81	-7.30	0.48
Linking	0.24	0.02	0.86	-5.50	0.87	-4.90	0.45
Accuracy of Grammatical Forms	0.01	0.03	0.97	-1.00	1.00	-0.10	0.27
Coherence	-0.01	0.03	0.81	-7.30	0.83	-6.60	0.44
Introduction-Body-Conclusion	-0.07	0.03	1.38	9.00	1.26	8.40	0.51
Word Choice	-0.07	0.03	0.81	-7.00	0.84	-6.10	0.35
Topic Relevance	-0.34	0.03	1.10	3.40	1.12	4.00	0.39
Appropriate Length	-0.49	0.03	1.28	8.70	1.21	6.40	0.45
Punctuation	-0.59	0.03	1.18	5.70	1.24	7.40	0.25
Spelling	-0.70	0.03	1.30	9.00	1.38	9.00	0.17
Mean	0.00	0.03	1.01	-0.3	1.01	-0.2	0.40
S (Universe)	0.36	0.00	0.25	7.4	0.23	7.2	0.09
S (Sample)	0.37	0.00	0.26	7.7	0.24	7.4	0.10

Model, Universe: RMSE = 0.03 Adjusted S = 0.36 Separation Ratio = 14.18
Separation Index = 19.25 Reliability = 1.00
Model, Sample: RMSE = 0.03 Adjusted S = 0.37 Separation Ratio= 14.69
Separation Index = 19.91 Reliability = 1.00

Model, Chi-square (Fixed Effect) : 2818.10 $sd = 14$ $p = .00$

Model, Chi-square (Normal) : 13.90 $sd = 13$ $p = .38$

Biserial correlations(x) are between .17 and .51, outfit values (MNSQ) are between 0.71 and 1.38, and the standards of outfit values (ZSTD) are between -9.00 and 9.00 (fit values: $0.4<x<0.8$, $0.5<MNSQ<1.5$ and $-2.0<z<2.0$). As a consequence, it is understood that there is no misfit in the dataset except for the standards of the outfit values.