**AN ALTERNATIVE MULTI-TRAIT RUBRIC FOR THE PERFORMANCE-BASED ASSESSMENT OF EFL WRITING PROFICIENCY AT BURSA ULUDAĞ UNIVERSITY, SCHOOL OF FOREIGN LANGUAGES**

**Doktora Tezi**

**Aliye Evin YÖRÜDÜ**

**Eskişehir 2021**

# AN ALTERNATIVE MULTI-TRAIT RUBRIC FOR THE PERFORMANCE-BASED ASSESSMENT OF EFL WRITING PROFICIENCY AT BURSA ULUDAĞ UNIVERSITY, SCHOOL OF FOREIGN LANGUAGES

**Aliye Evin YÖRÜDÜ**

**DOKTORA TEZİ**

**Yabancı Diller Eğitimi Anabilim Dalı, İngilizce Öğretmenliği Programı**

**Danışman: Prof. Dr. Fatma Hülya ÖZCAN-ÖNDER**

**Eskişehir**

**Anadolu Üniversitesi**

**Eğitim Bilimleri Enstitüsü**

**Ağustos 2021**

## JÜRİ VE ENSTİTÜ ONAYI

Aliye Evin YÖRÜDÜ'nün "An Alternative Multi-trait Rubric for the Performance-Based Assessment of Writing at Uludağ University, School of Foreign Languages" başlıklı tezi 01.07.2021 tarihinde aşağıdaki jüri tarafından değerlendirilerek Anadolu Üniversitesi Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliği'nin ilgili maddeleri uyarınca Yabancı Diller Eğitimi Anabilim Dalı İngilizce Öğretmenliği Programı'nda, Doktora tezi olarak kabul edilmiştir.

| | Unvanı-Adı Soyadı | İmza |
|---|---|---|
| Üye (Tez danışmanı) | : Prof. Dr. F. Hülya ÖZCAN ÖNDER | ……. |
| Üye | : Prof. Dr. Handan YAVUZ | ……. |
| Üye | : Prof. Dr. Demet YAYLI | ……. |
| Üye | : Prof. Dr. Çiler HATİPOĞLU | ……. |
| Üye | : Dr. Öğr. Üyesi Aysel KILIÇ | ……. |

Prof. Dr. Bahadır ERİŞTİ
Anadolu Üniversitesi
Eğitim Bilimleri Enstitüsü
Müdürü

# ÖZET

BURSA ULUDAĞ ÜNİVERSİTESİ YABANCI DİLLER YÜKSEKOKULU İNGİLİZCE HAZIRLIK BÖLÜMÜ'NDE İNGİLİZCE YAZILI ANLATIM BECERİSİNİ ÖLÇMEK ÜZERE GELİŞTİRİLEN ÇOK BOYUTLU NOTLANDIRMA ÖLÇEĞİNİN GÜVENİRLİK VE GEÇERLİLİĞİ

Aliye Evin YÖRÜDÜ

Yabancı Diller Eğitimi Anabilim Dalı

İngilizce Öğretmenliği Programı

Anadolu Üniversitesi, Eğitim Bilimleri Enstitüsü, Ağustos 2021

Danışman: Prof. Dr. Fatma Hülya ÖZCAN-ÖNDER

Bu çalışma, Bursa Uludağ Üniversitesi, Yabancı Diller Yüksekokulu, İngilizce Hazırlık Programı'nın (BUÜ-YDYO-İHP) yeterlilik sınavındaki yazılı anlatım bölümünün değerlendirilmesinde kullanılan ölçeğe yönelik güvenirlik ve geçerlik kaygıları nedeniyle, kurama dayanan, güvenirlik ve geçerliliği nicel olarak kanıtlanmış ve yazılı anlatım becerisini gereksinimlere göre ölçebilecek bir ölçek geliştirmeyi amaçlamaktadır. Ölçek kullanılarak elde edilen notların güvenirliğinin kanıtlanabilmesi, ölçeğin nicel ve nitel araştırma yöntemleri ile değerlendirilmesini gerektirmektedir. Bu nedenle, çalışmada karma yöntem benimsemiştir. Çalışmanın nicel evresinde, öğretmen-puanlayıcılar öğrenci metinlerini yeni ölçeği kullanarak notlandırmış, verinin Çok-yüzeyli Rasch analizi, ölçeğin güvenilir ve geçerli olduğunu göstermiştir. Nitel veri de sonuçları desteklemiştir. Çalışma, Türkiye'de İngilizce öğreten kurumların uyguladığı sınavlarda yazılı anlatım becerisinin değerlendirilmesi için güvenilir ve geçerli bir ölçeğin geliştirilmesi konusunda yol göstermektedir.

**Anahtar Sözcükler:** Yazılı anlatım becerisinin değerlendirilmesi, Ölçek geçerlik ve güvenirliği, Çok-yüzeyli Rasch Ölçme Modeli.

## ABSTRACT

AN ALTERNATIVE MULTI-TRAIT RUBRIC FOR THE PERFORMANCE-BASED ASSESSMENT OF EFL WRITING PROFICIENCY AT BURSA ULUDAĞ UNIVERSITY, SCHOOL OF FOREIGN LANGUAGES

Aliye Evin YÖRÜDÜ

Department of Foreign Language Education

Program in English Language Teaching

Anadolu University, Graduate School of Educational Sciences, August 2021

Supervisor: Prof. Dr. Fatma Hülya ÖZCAN-ÖNDER

Considering the contextual and other validity-related concerns related to the analytic rubric in use for the performance assessment of EFL writing proficiency at Bursa Uludağ University, School of Foreign Languages, Intensive English Program (BUU-SFL-IEP), the purpose of this research is to develop a theoretically-based and an empirically-validated multi-trait writing rubric which may serve to measure writing proficiency more validly and reliably. Validation requires the evaluation of an instrument through a variety of quantitative and qualitative forms of evidence to support inferences from test scores. Hence, pure mixed-methods research approach is utilized in this study through the application of both qualitative and quantitative methods. In the quantitative phase, teacher-raters rated student essays using the new rubric, and Many Faceted Rasch Measurement (MFRM) analysis revealed that it was statistically reliable and valid. The qualitative phase explored the teacher-raters' perspectives on the efficacy of the new rubric. The findings support the quantitative findings indicating that all participants were satisfied with the new rubric. Finally, the study offers suggestions and guidance for other EFL contexts in Turkey that use high-stakes performance assessment.



**Keywords:** Performance-based assessment of writing proficiency, Rubric validation, MFRM.

## ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisor, Prof. Dr. Fatma Hülya ÖZCAN-ÖNDER, for her continuous support of my Ph.D research and guidance and understanding that she has provided throughout the writing of this dissertation. I was extremely lucky to have the opportunity to work with her.

Besides, I am grateful to my dissertation committee members Prof. Dr. Handan KOPKALLI-YAVUZ and Dr. Aysel KILIÇ for their constant encouragement and constructive criticism and, other members of my dissertation committee, Prof. Dr. Çiler HATİPOĞLU and Prof. Dr. Demet YAYLI for their insightful comments. With their extensive knowledge, expertise in the field, constructive feedback, and positive attitudes, this dissertation improved significantly.

Very special thanks go to the three experts, Prof. Dr. Hüsnü ENGİNARLAR, Dr. William Snyder, and Tony Gurr, who provided their invaluable opinions during the process of rubric design.

I would like to extend my gratitude to Dr. Kübra Karakaya-ÖZYER for her invaluable assistance in the statistical analyses of my dissertation. She was always there whenever I needed help or had a question.

My deepest appreciation goes to all of my colleagues at BUU-SFL-IEP who were participants in the different phases of this study. I would also like to thank to our students who gave consent for the use of their essays for the purposes of this research.

I would also like to say a heartfelt thank you to my dearest friends and my precious family for their unwavering support, belief in my abilities, and for their love and support in each and every challenge that I dare to confront.

And my late brother-in-law Akın DÜNDAR… Words just stick in my throat. I am having great difficulty finding the appropriate words to describe him and his support to me at each and every important stage of my life. I miss you every day.

**28/07/2021**

## ETİK İLKE VE KURALLARA UYGUNLUK BEYANNAMESİ

Bu tezin bana ait. özgün bir çalışma olduğunu: çalışmamın hazırlık. veri toplama. analiz ve bilgilerin sunumu olmak üzere tüm aşamalarında bilimsel etik ilke ve kurallara uygun davrandığımı; bu çalışma kapsamında elde edilen tüm veri ve bilgiler için kaynak gösterdiğimi ve bu kaynaklara kaynakçada yer verdiğimi; bu çalışmanm Anadolu Üniversitesi tarafından kullanılan "bilimsel intihal tespit programı"yla tarandığını ve hiçbir şekilde "intihal içermediğini" beyan ederim. Herhangi bir zamanda, çalışmamla ilgili yaptığım bu beyana aykırı bir durumun saptanması durumunda, ortaya çıkacak tüm ahlaki ve hukuki sonuçları kabul etliğimi bildiririm.

Aliye Evin Yörüdü

**TABLE OF CONTENTS**

**Page**

**Page**

**Page**

**Page**

**Page**

**Page**

## LIST OF TABLES

**Page**

**Page**

## LIST OF FIGURES

**Page**

## LIST OF ABBREVIATIONS

BUU-SFL-IEP : Bursa Uludağ University, School of Foreign Languages, Intensive English Program

MFRM : Many Faceted Rasch Measuremen

# 1. INTRODUCTION

## 1.1. Background to the Study

Performance-based assessment of writing proficiency has become the norm since the early part of the 1990's together with the impact of the communicative era in language teaching on language testing (Brown, 2004; McNamara, 1996, 2002; Shohamy, 1995; Weigle, 2002). According to Weigle (2002, p. 46), "any assessment procedure that involves either the observation of behavior in the real world or a simulation of a real-life activity" could be viewed as performance-based assessment because the written product is a form of a performance of writing. However, not only is writing a complex and multifaceted activity, but the performance-based assessment of writing is also complex and multifaceted (Hamp-Lyons, 1995, 2016a, 2016b); that is, there is a variety of factors, variables, or components of the measurement situation that is assumed to affect test scores in a systematic way (Eckes, 2011).

From a socio-cognitive perspective, Shaw and Weir (2007, p. x) group these facets into three general categories as follows:

- the test-taker's cognitive abilities,
- the context in which the task is performed, and
- the scoring process.

Shaw and Weir refer to these three internal dimensions of performance-based assessment successively as *cognitive validity*, *context validity*, and *scoring validity* and emphasize that the three constitute "an innovative conceptualization of *construct validity*". In the Dictionary of Language Testing (Davies et al., 1999, p. 33), construct validity of a language test is defined as:

> "An indication of how representative a test is of an underlying theory for writing proficiency… Recent views of construct validity consider broader range of factors such as performance differences across different groups, times, and settings and test taker and rater behavior".

In this view, not only is language use but also language assessment is a cognitively processed phenomenon and socially situated. According to Shaw and Weir (2007, p. x), such an understanding of validity has sound theoretical and direct practical relevance for the assessors of performance-based writing proficiency. Crusan (2010) adds another type of validity to this list: *consequential validity*, which she believes is "the most important" type of validity (p. 41) and emphasizes that when a test is administered, the consequences

of this test need to be contemplated by the test makers. Altogether, different types of validity mentioned so far comprises test validity.

In relation to the validity argument, one facet is central in the performance-based assessment of writing: the rubric (Knoch, 2007, p. 4; Weigle, 2002, p. 108) since "the writing rubric and the way raters interpret it represents the defacto test construct" (Knoch, 2011, p. 81), in particular in high-stakes tests (Eckes, 2011, p. 3). Because rubrics are the foundation of a rater's scoring process, principled rubric use requires systematic review as rubrics are developed, adopted, or adapted into different local contexts (Crusan, 2010; Janssen, Meier, & Trace, 2015). According to Janssen, Meier, and Trace (2015), an ongoing rubric analysis to validate the rubric is a necessity in contexts that use high-stakes performance assessment of writing proficiency.

In the process of developing a valid writing rubric for high-stakes performance assessment of writing proficiency, one of the major decisions to be made is what type of rubric will be used (Barkaoui, 2007; Weigle, 2002). This decision is critical because "the score is ultimately what will be used in making decisions and inferences about writers" (Weigle, 2002, p. 108). The type of rubric to be used in the scoring process is determined according to two general scoring approaches categorized as *norm-referenced* and *criterion-referenced* (Hyland, 2003). As expressed by Hyland (2003), judging a student's writing performance in comparison with the performance of other students is defined as *norm-referenced* method. However, it has largely given way to *criterion-referenced* practices where the quality of each writing performance is judged against its own right according to some external criteria, such as coherence, grammatical accuracy, and so on (Hyland, 2003). Criterion referenced procedures take a variety of forms and fall into three main categories:

- *holistic*,
- *analytic*, and
- *trait-based* (Weigle, 2002, pp. 108-39).

To summarize the in-depth information on criterion-referenced approaches provided by Weigle, holistic rubrics offer a general impression of a piece of writing while analytic rubrics are based on separate scales of overall writing features. Trait-based rubrics differ from holistic and analytic methods in that they are context-sensitive; that is, they judge performance traits relative to a particular task. They are further categorized into two as *primary-trait* and *multiple-trait*. Primary-trait scoring involves the scoring of

a piece of writing in relation to one principal trait specific to that task. Multiple-trait scoring, on the other hand, requires raters to provide separate scores for different writing features as in analytic scoring.

Hamp-Lyons explains her disfavor for the term "analytic" in her following words:

> "I don't like the term 'analytic' because it takes us back to the time when, in the US particularly, the direct assessment of writing had fallen into disfavour and educational measurement gurus argued that indirect measures of writing were as good as the direct scoring methods of the time, and more reliable. …The term *analytic* is best reserved for the attempts to capture (usullay merely hypothesized) characteristics and skills of writing through the use of multiple-choice and other indirect or semi-direct test item types" (Hamp-Lyons, 2016a, pp. A1-A2).

Some researchers consider multi-trait and analytic rubrics synonymous (Davies et al., 1999, p. 126; Lee, Gentile, & Kantor, 2010, p. 391; Weigle, 2002, p. 109) because many of the characteristics ascribed to multiple-trait rubrics have to do more with procedures for developing and using rubrics, rather than with the descriptions of rubrics themselves (Weigle, 2002, p. 109). However, there is one important difference between the two as stated by Hyland (2003, p. 230): the emphasis put on the *context*. Multiple-trait rubrics treat writing as a multifaceted construct which is situated in particular contexts and purposes, so they can address traits that do not occur in general analytic rubrics such as "the ability to summarize a course text, consider both sides of an argument, or develop move structure of an abstract" (Hyland, 2003, p. 230). Other than that, multiple-trait and analytic rubrics have similar features. However, there certainly are significant differences between multi-trait and holistic rubrics.

A multi-trait rubric rather than a holistic one has been designed for the purposes of this research due to two reasons. The first reason is that in multi-trait scoring raters are required to provide separate scores for each of several facets or traits of the performance as opposed to holistic scoring where raters judge a performance impressionistically according to its overall properties (Davies et al., 1999). The second reason is that according to Hamp-Lyons (1995, 2016a, 2016b), while holistic scoring is appropriate for scoring first-language (L1) essays, multiple-trait scoring has higher validity and reliability when rating second- or foreign-language (L2) essays because different learners have different levels of proficiency in different aspects of L2 writing.

In light of the brief literature review provided above, it can be concluded that a context-sensitive rubric validation where the type of rubric to be used needs to be

judiciously decided is a requirement for institutions that use high-stakes performance assessment of writing proficiency. Chalhoub-Deville claims that:

> "… in high stakes testing where critical decisions are made (e.g., certification, fulfilling a degree requirement, admission into a programme, progressing into a higher grade, securing a job, etc.), it is imperative that resources be allocated for assessment frameworks to be validated in their context of use. In high-stakes testing, the deficiency of evidence to support an assessment framework in a given context of application weakens the validity of test interpretation and use, which has grave ramifications" (Chalhoub-Deville, 1997, p. 17).

As Akşit (2018) put forward, admission decisions need to be backed up with research as warrants of their validity. Otherwise, decisions of admission or refusal based on test scores are not meaningful, fair, or justifiable. It would not be wrong to state that the intensive English programs of the universities in Turkey need such validation processes for their proficiency examinations as these tests are examples of high stakes testing based on which students are admitted to study in their departments or not. Validation is particularly important for the performance-based assessment components of these examinations due to their subjective nature. The context of the current research, Bursa Uludağ University, School of Foreign Languages, Intensive English Program (hereafter BUU-SFL-IEP), was also in need for such a validation process for the assessment of writing performance carried out in the proficiency examination that takes place at the end of each academic year as an exit examination.

### 1.2. Context of the study

BUU-SFL-IEP pursues skills-based language instruction in which each of the four language skills (i.e., reading, writing, listening, and speaking) and language systems (i.e., grammar and vocabulary) are taught and assessed separately throughout the academic year. At the end of the academic year, a large-scale in-house English language proficiency examination at the B1+ proficiency level is administered for over 1.500 students to measure whether they are proficient enough to carry on studying in their departments where 30% of their content courses (100% for a few departments) are conducted in English. The minimum score that test takers need to be able to move to any of the undergraduate programs at BUU is 60 out of 100, except for the English language teaching department, which requires 75. Alternatively, equivalent scores from language examinations given by one of the two external organizations; namely, the TOEFL IBT (72 – 102 points) by Educational Testing Service (ETS) (www.ets.org) and IELTS (5.0 –

6.0), which is jointly owned by British Council, IDP: IELTS Australia, and Cambridge English Language Assessment (www.ielts.org), are also accepted as valid proof of English language proficiency. The students who obtain the required scores from one of these examinations start their subject studies whereas those who fail to receive the required minimum score, study at the BUU-SFL-IEP for one year before taking the examination again at the end of the instructional period in June. Considering the consequences of the BUU-SFL-IEP proficiency examination for students, it might then be concluded that it is a high-stakes test, and consequential validity is surely an issue to be taken into consideration seriously.

The English language proficiency examination comprises four sections: language use, listening, reading, and writing. Language use, listening, and reading sections are 100 minutes long and consists of 80 multiple-choice questions. These sections comprise 80% of the test-takers' total score. The writing section is 50 minutes long and offers test takers a choice of three argumentative essay prompts. They choose one and are expected to write between 200-250 words. The prompts require test takers to give their opinion on a statement and justify their opinion using supporting details. The writing section of the examination which is an example of performance-based assessment writing proficiency comprises 20% of the test-takers' total score. Thus, the writing section also has a high-stakes use. Due to the high number of students who take the examination, all instructors take part in the different phases of the process as proctors and/or raters.

The scoring process of the performance writing assessment at BUU-SFL-IEP proceeds as follows. The process begins a day after the proficiency examination is administered with the training and norming session that is moderated by the writing course coordinator in collaboration with the continuous professional development unit head. In this meeting, BUU-SFL-IEP instructors are asked to go over the analytic rubric that is used for scoring and independently rate 5 sample essays selected from the pool of the essays written by the students in the proficiency examination in 30 minutes. At the end of the 30-minute period, raters negotiate on the scores that they assign for each of the 5 essays with the guidance of the moderators. After the raters reach an agreement on the scores and ask any questions that they may have regarding the scoring process and the use of the rubric, the actual scoring process begins. Each rater is expected to score 50 essays in a total of five hours. After each rater scores 25 essays in two and a half hours, s/he receives the second batch that consists of another set of 25 essays, which means each

essay is scored independently by two trained raters to give each performance a score out of 20 by using the analytic rubric. The two raters' scores are summed by the testing unit to arrive at the final score for each performance. Rater agreement is monitored throughout the scoring process. If raters give non-adjacent scores, the performance is re-evaluated by a third examiner. All instructors with minimum two years of experience in scoring EFL writing performance take part in the scoring process as raters due to the large number of students taking the examination.

### 1.3. Statement of the problem

For the assessment of performance writing component of the proficiency examination at BUU-SFL-IEP, an adapted version of the analytic rubric, ESL Composition Profile (See Appendix 1), which was developed by Jacobs, Zinkgraf, Wormuth, Hartfield, and Hugley (1981, p. 30) and regarded as "one of the best known and most widely used analytic scales in ESL", is used because of the supposedly strong construct validity it has in terms of proposed writing course goals (Weigle, 2002, p. 115). However, several concerns arise concerning the use of this rubric in terms of its context validity, scoring validity, construct validity, and in turn its consequential validity.

The first problem is related to the context validity because the adapted version of the rubric is "intuition-based" as adapted or adopted rubrics defined in the literature rather than a locally designed and locally controlled one (Crusan, 2010, p. 72; Janssen, Meier, & Trace, 2015, p. 53; Knoch, 2007, p. 5) . They are called intuition-based since they are based on other rubric samples, reflecting the experience of other rubric developers. Taking an intact rubric and modifying it for the local assessment context is a strategy likely to be adequate for most classroom assessment, yet it simply does not suffice in tests with more high-stakes uses (Janssen, Meier, & Trace, 2015). According to Broad (2003), this lack of contextual relevance and failure to grow organically from contexts and purposes make many traditional rubrics problematic. Thus, good writing assessment should be contextualized and locally developed (Hamp-Lyons, 1995; Crusan, 2010, 2015).

The second problem is related to the scoring validity of the rubric. In the recent years, it has been observed that the number of essays re-evaluated by the third examiner increased drastically. In order to find out where the problem arose from, raters were requested to provide feedback during the staff meetings that were held to evaluate the

program at the end of the 2016-2017 academic year. These meetings revealed concerns, particularly in relation to the rubric's difficulty of use when scoring. For instance, the majority of the raters have criticized the rubric's use of impressionistic terminology or relativistic wording in descriptors, which is open to subjective interpretations and make it difficult to differentiate between score or band levels. Raters stated to have relied on their "gut feeling" of the level of a performance, which is not a legitimate way to assess test takers' writing proficiency, particularly for such a high-stakes test. Although teacher-raters have voiced their concerns periodically in different platforms since 2010, the year when the writing rubric used currently was initiated, no action was taken until 2016, the year when the curriculum renewal process was started.

Another problem pointed by the raters is related to the construct validity of the test: the number and weighting of the different categories (i.e., content, organization, grammar, and vocabulary) in the rubric. Most raters indicate that there seems to be "a category or more is missing" in the rubric because increasing scores are not consistently representative of increased examinee ability. The majority of the raters indicated that there needs to be a category to measure the logical flow of ideas, i.e., fluency or coherence which they reckon is not measured in the rubric in use. Regarding the weighting of each category, all raters agreed that they need to be reevaluated being dependent on what is valued more in the performance assessment of writing proficiency in our context.

Taken together, these issues negatively affect the test validity of the performance-based assessment of writing proficiency at BUU-SFL-IEP. Hitherto test validity has been emphasized more than reliability, which is defined as "the extent to which results can be considered consistent" (Bachman & Palmer, 1996, p. 16). However, it is not because reliability is not an important issue, but because test validity is essential to test reliability (Crusan, 2010, p. 42). In other words, if a test is not valid, there is no point in discussing reliability because test validity is required before reliability can be considered in any meaningful way.

### 1.4. Statement of the purpose

Considering the contextual and other validity-related concerns in relation to the analytic rubric in use at present for the performance-based assessment of EFL writing proficiency at BUU-SFL-IEP, the purpose of this research study was to develop an

alternative theoretically-based and an empirically-validated multi-trait writing rubric which may serve to measure writing proficiency more validly and reliably.

The three research questions guiding the study are presented below:

1. What are the teacher-raters' perspectives on the rubric that is currently used for the performance-based assessment of writing proficiency?
2. To what extent is an alternative theoretically-based and empirically-developed multi-trait rubric of academic writing (that has been newly designed) valid and reliable for the measurement of performance-based assessment of EFL writing proficiency at BUU-SFL-IEP?
3. What are the teacher-raters' perspectives on the use of this alternative multi-trait writing rubric that has been newly designed?

### 1.5. Significance of the Study

English is the medium of instruction in many public and foundation Turkish universities fully or partially. Hence, it is a general practice in these institutions to use a variety of assessment instruments and procedures to make admission decisions concerning language proficiency. Some intensive English programs administer tests that are prepared by international organizations such as the Educational Testing Service's TOEFL or the IELTS that is a product of a partnership between British Council and IDP Australia. Examples of such institutions that use these tests are Koç University in Istanbul and TOBB ETÜ University in Ankara (Akşit, 2018). Others design their own English language tests. Some of the state (S) and private (P) universities that prepare their own proficiency tests are: Anadolu University (S), Ankara Yıldırım Beyazıt University (S), Atılım University (P), Bahçeşehir University (P), Başkent University (P), Bilgi University (P), Bilkent University (P), Boğaziçi University (S), Bursa Technical University (S), Çağ University (P), Çankaya University (P), Erciyes University (S), Eskişehir Osmangazi University (S), Gazi University (S), Hacettepe University (S), İstanbul University (S), İstanbul Technical University (S), Konya Selçuk University (S), Middle East Technical University (S), Sabancı University (P), and TED University (P) (Akşit, 2018). In language programs where in-house tests are used, test development and administration processes are usually kept confidential. Thus, the amount of information from within the language programs of those institutions is limited. This is also the case in the amount of research into the validity of English proficiency tests used in the intensive

English programs of Turkish state and private universities. To the best knowledge of the researcher of this study, there are only six studies focusing on the validity of some aspects of English proficiency tests used in these programs (Akşit, 2018; Ataman, 1999; Gürsoy, 2013; Kutevu, 2001; Yapar, 2003; Yeğin, 2003).

At the local level, it is expected then that the current research may improve practice in the field of language teaching and testing in other English as a Foreign Language (hereafter EFL) contexts in Turkey by modeling how an assessment and validation framework is carried out in an EFL context to theoretically design and empirically validate a multi-trait rubric to be used for the writing section of a high-stakes language test. Test designers, raters, instructors, academic coordinators, administrators, and other policy makers might be informed of the procedures and processes implemented in rubric design and validation for the performance-based assessment of writing proficiency. It may also provide guidance for the design of a valid and reliable writing rubric because the development and modification of writing rubrics is rarely discussed in the language assessment literature in general (Banarjee et al., 2015; Knoch, 2009, 2011; Lallamamode, Daud, & Abu Kassim, 2016) and Turkish EFL context in particular (Hatipoğlu, 2015). The last but not the least, the current study may contribute to produce accurate, consistent, and fair results for the performance-based assessment of EFL writing proficiency at BUU-SFL-IEP.

At the global level, this research may contribute to the wider knowledge base of the application of a framework for validation purposes for a rubric to be used for the performance-based assessment of writing. Using an assessment/test design framework provides a sound basis on which to build an assessment tool. Moreover, as the framework is used in a wider variety of contexts and in different backgrounds, it is possible to obtain more information on the different facets of writing performance and whether all aspects presented in the framework are applicable in contexts other than it was created.

## 2. LITERATURE REVIEW

### 2.1. Performance Assessment of Writing

Language testing and linguistic theories of the time have always followed a parallel path (Shohamy, 1995). A movement of criticism against the traditional non-communicative tests aroused with the advent of the communicative era in language teaching in the 1970s. This has led a radical move in language testing towards the development and use of performance tests in the last three decades based on the prospect that "such tests would assess a more valid construct of what it really means to know a language" (Shohamy, 1995, p. 188). Thus, the development and use of tests that are similar to characteristics of real language use and necessitate test takers to perform language that is "authentic, direct, communicative, and performance-based" have become the norm (Shohamy, 1995, p. 189).

Performance tests which are also known as authentic tests or direct tests are defined as "any tests that are designed to elicit performances of the specific language behaviors that the testers wish to assess" (Brown, 2004, p. 92). They are designed to elicit students' abilities to write or speak, and they are generally scored in terms of the linguistic features of the writing or speaking performance that the test designer deems to be important for theoretical and/or pedagogical reasons (Brown, 2004).

McNamara (1996) differentiates between a *strong sense* and a *weak sense* of performance assessment. In the strong sense of performance assessment, the focus is on the successful realization of the task rather than the successful use of language in performing the task. For instance, if the writing task is about making an official request, the task will be evaluated primarily on real-world criteria. Thus, the performance of the task itself is the focus of the assessment. At the center of the weak sense of performance assessment, on the other hand, lies the language use. The target is to elicit a demonstration of writing ability even though the task used to obtain a writing sample might be similar to real-world tasks. The differentiation between the strong and weak forms of performance assessment is of great importance in terms of L2 writing assessment because the rubric to be used to evaluate the performance needs to express the descriptions of the construct of L2 writing performance as clearly as possible (McNamara, 1996). If not, test scores may reflect *construct irrelevant variance*, which is defined as "a type of systematic measurement error where there is some variance in the test scores that is due to factors other than the construct in question" (Davies et al., 1999, p. 32).

According to Hamp-Lyons (1991), a performance test of writing needs to possess minimally the following features:

- The sample of written performance produced by the test taker should be composed of a minimum of 100 words of continuous text.
- The test taker is provided with some scope and expected to construct a response to the task prompt, but s/he still has freedom to express himself/herself.
- Each written response is evaluated by at least one, and more often two, human raters (with a third if there is a discrepancy between the two), who have been trained to equip them with the necessary writing evaluation skills.
- Raters' judgments should be in compliance or associated with a common standard which may include a set of exemplar performances or a clear description of expected performance at particular proficiency levels.
- Raters' judgments are stated explicitly in numerical terms, and a permanent record of test scores is kept and made available when required.

When brought in at an early stage of instruction, performance tests might be useful for providing learners with information about the importance of language learning outcomes, instructors' expectations, and criteria for assessing performances (Shohamy, 1995). Shohamy (1995, p. 190) goes on to support that "texts and tasks used in performance testing make very effective instructional tasks, and ratings gained from performance tests could be transformed into diagnostic feedback in the form of a profile score". Hence, they might be used for different test purposes such as proficiency, placement, formative diagnosis, or achievement (Shohamy, 1995).

### 2.1.1. Historical Development of Writing Assessment

An understanding of the past is vital to an understanding of the present (Crusan, 2010). As Matsuda (2003b, p. 15) claims, "Without knowing the context in which certain theories or pedagogical strategies developed, we will not be able to apply them or modify them in other contexts or in light of new theoretical insights". This also applies to writing assessment because the theoretical stance that an assessor has is exemplified in his/her practice. Thus, teachers and/or assessors need to be aware of assessment theory and history and the ways they are represented in their pedagogy (Crusan, 2010).

According to Crusan (2010), the history of writing assessment is inevitably related to the history of composition, and L2 histories of writing and writing assessment are

intertwined with histories of their L1 counterpart. Further, the way(s) that writing assessments are today are considered as productions of the interplay over time between writing theorists, test makers, teachers, and administrators, as emphasized by Behizadeh and Engelhard (2011).

The history of written tests dates back to the 19th century when different institutions of education used writing as a tool for assessing knowledge. Harvard University started administering a written entrance exam instead of an oral one in the late 1800's (Knoch, 2007). In the meanwhile, performance assessment of writing where writing was tested by sampling actual examples of writing was started to be used in Europe when colonial powers needed literate administrators in different countries around the world. After these developments in Europe and the United States, an increased level of standardization was required, which brought about an interest in measurement theory (Knoch, 2007). Since then measurement theory has had a strong effect on writing assessments rather than writing theory according to Behizadeh and Engelhard's (2011) historical analysis that covers selected time periods in the 20th century up to now in the United States.

Until the 1950's, writing assessment was chiefly carried out by individual teachers in the context of their classes; nonetheless, the increasing number of university enrolments brought a greater demand for reliability (Knoch, 2007). Responding to this demand, indirect assessments; that is, discrete-item tests were developed by psychometricians during the first half of the 1950's. The Test of Standard Written English (TSWE) which was developed by Educational Testing Service (ETS) was a very powerful test that had multiple-choice items to measure writing ability for English L1 writers. Reliability was considered to be more important than validity in these discrete-item tests.

The birth of modern writing assessment in the United States and Britain was in the late 1950's and early 1960's (Hamp-Lyons, 2017). During this time period, Carroll and others in the US army and air force developed an aptitude test (the Foreign Language Aptitude Battery) that aimed to make reliable predictions of how well an individual would be able to master a language (Carroll, 1962). The prominent work of Robert Lado at the University of Michigan on the Michigan tests, the Certificate of Proficiency in English (ECPE) in particular, and David Harris' work at the American University Language Center, which led to the Test of English as Foreign Language (TOEFL), were again developments of this era (Hamp-Lyons, 2017). Change also began in Britain mainly due to the upsurge in the number of foreign candidates who applied to UK universities in the

mid-1960's. The English Proficiency Test Battery (EPTB: Davies, 2008), which would be replaced by another world-famous test, the English Language Testing System (ELTS) and which eventually would become the International English Language Testing System (IELTS), appeared in this era. According to Broad (2003), modern writing assessment was precisely born in 1961 when Diederich, French, and Carlton of the Educational Testing Service (ETS) published *Factors in Judgments of Writing Ability*, which was based on a decade of research done at ETS and elsewhere on writing assessment and inter-rater reliability. Broad (2003, p. 6) asserts that "And thus was born what became the standard, traditional, five-point rubric, by some version of which nearly every large-scale assessment of writing since 1961 has been strictly guided". The emphasis was more on reliability rather than validity once again.

The next major change in approaches in language testing was stimulated by the influential work on communicative competence of Hymes (1972), Widdowson (1978), and Canale and Swain (1980). The emphasis on communicative language teaching brought about communicative language testing, which led to performance-based assessment of speaking and writing. Thus, in the late 1970's and early 80's, performance assessment of writing became commonplace in English L1 contexts and also in its L2 counterpart (Hamp-Lyons, 2017). Since then the testing of writing has generally followed the following procedure: test-takers write a brief essay within a 30- or 40-minute period of time, and then, it is rated either holistically or analytically by trained raters using a writing rubric.

The publication of Lyle Bachman's (1990) *Fundamental Considerations in Language Testing* (1990) and the presentation of his model of communicative language ability was another hallmark in the sophistication of concepts of test purpose in language testing (Hamp-Lyons, 2017). Around the same time, a leading undertaking was commencing in Europe which was funded and supported by the Council of Europe to develop a common European framework for languages; that is, the *Common European Framework for Reference* (CEFR). This was also a time of increasing awareness of the construct in test design. The meaning of proficiency was given more thought, and it was recognized that if a test aims to assess a learner's language proficiency, it needs to assess that proficiency in all four skills so that the clearest and fullest possible picture of what the learner *can do* could be obtained. In the 1990's language testers also started becoming more aware of the studies on validity in psychological and educational measurement,

particularly that of Messick (1989) and Kane (1992), the influence of whom still continue in research in language testing today. According to Hamp-Lyons (2017), this increasing understanding of the significance of validity required language testers to become more self-aware of their work, and it also put more emphasis on liability in the profession.

### 2.1.2. Current practices

Today there is a strong argument for making tests as direct as possible (Shaw & Weir, 2007). At this point it is important to look into the practices of some major standardized assessments around the world because not only can we understand current practices better but also an inevitable effect (sometimes an intended outcome) of the use of standard language tests is to influence what is taught and how it is taught (Hamp-Lyons, 2017).

One broadly administered writing test around the world is the writing section of the IELTS, which includes two tasks. The first task requires the test taker to describe information given in a graph or table, and the second task to produce a slightly longer argumentative essay. Using an analytical writing rubric, the tasks are evaluated by one trained rater, which might lower its reliability.

Another large-scale performance assessment of writing is administered by the ETS as a component of the TOEFL iBT (Internet-based test). Test takers are required to produce two pieces of writing based on two tasks: one integrated task where the test taker reads a text, listens to a 2- or 3-minute lecture that challenges the argument in the reading, and writes an essay comparing the reading and the lecture in 20 minutes; one independent task where test takers write an argumentative essay in 30 minutes. Using a holistic writing rubric, both tasks are evaluated by two trained raters in addition to the ETS e-rater software which mostly focuses on language use and mechanics.

Both tests use performance assessment of writing including various tasks with timed writing, raters, and writing rubrics and are considered to be proficiency tests as they are designed to assess general writing ability. Most of the intensive English programs in Turkey that are part of state or private universities have performance assessment of writing evaluated by teacher raters using a rubric as part of the proficiency examination that serves as an exit test, regardless of the language used as a medium of instruction.

While it is widely acknowledged that performance-based assessment of writing should be the norm to assess writing proficiency, it is not without its problems because

of the subjectivity that is involved in the rating process. Various models of performance assessment of writing were developed in order to minimize the undesired effects of the process, which is the topic of the following section.

### 2.1.3. Models of performance assessment of writing

Due to its subjective nature, there is more undesired variance in the test score obtained through performance-based assessment. Since this kind of unwanted variability hinders the construct being measured, it is also called *construct-irrelevant variance* and threatens validity and fairness of assessment outcomes (Eckes, 2011; Weir, 2005). Task characteristics, such as task difficulty, and rater characteristics, such as rater background, severity, bias, and decision making processes could be listed as sources of such undesired variance.

A variety of performance assessment models have been designed initially for the context of proficiency testing of oral language performance (e.g., Fulcher, 2003; McNamara, 1996; Skehan, 1998), which serve two purposes: to organize language testing research and account for the abovementioned factors that cause the systematic variance of performance test score (Knoch, 2007). Each model builds on each other and is just valid for written test performance, as stressed by Knoch. The current research, on the other hand, is guided by Shaw and Weir's (2007) framework due to its particularity and multidimensional approach to the assessment of writing performance and its unified approach to establishing the overall validity of a test. See Figure 2.1. below.

Shaw and Weir (2007) adopt a socio-cognitive perspective for conceptualizing writing test performance. In their own words, their approach is "effectively an *interactionalist* position which sees the construct as residing in the interactions between the cognitive ability and the context of use – hence the socio-cognitive model" (Shaw & Weir, 2007, p. 3). In addition, the validation process is conceptualized by identifying various types of validity evidence that need to be collected at each stage in the test development, monitoring, and evaluation cycle.

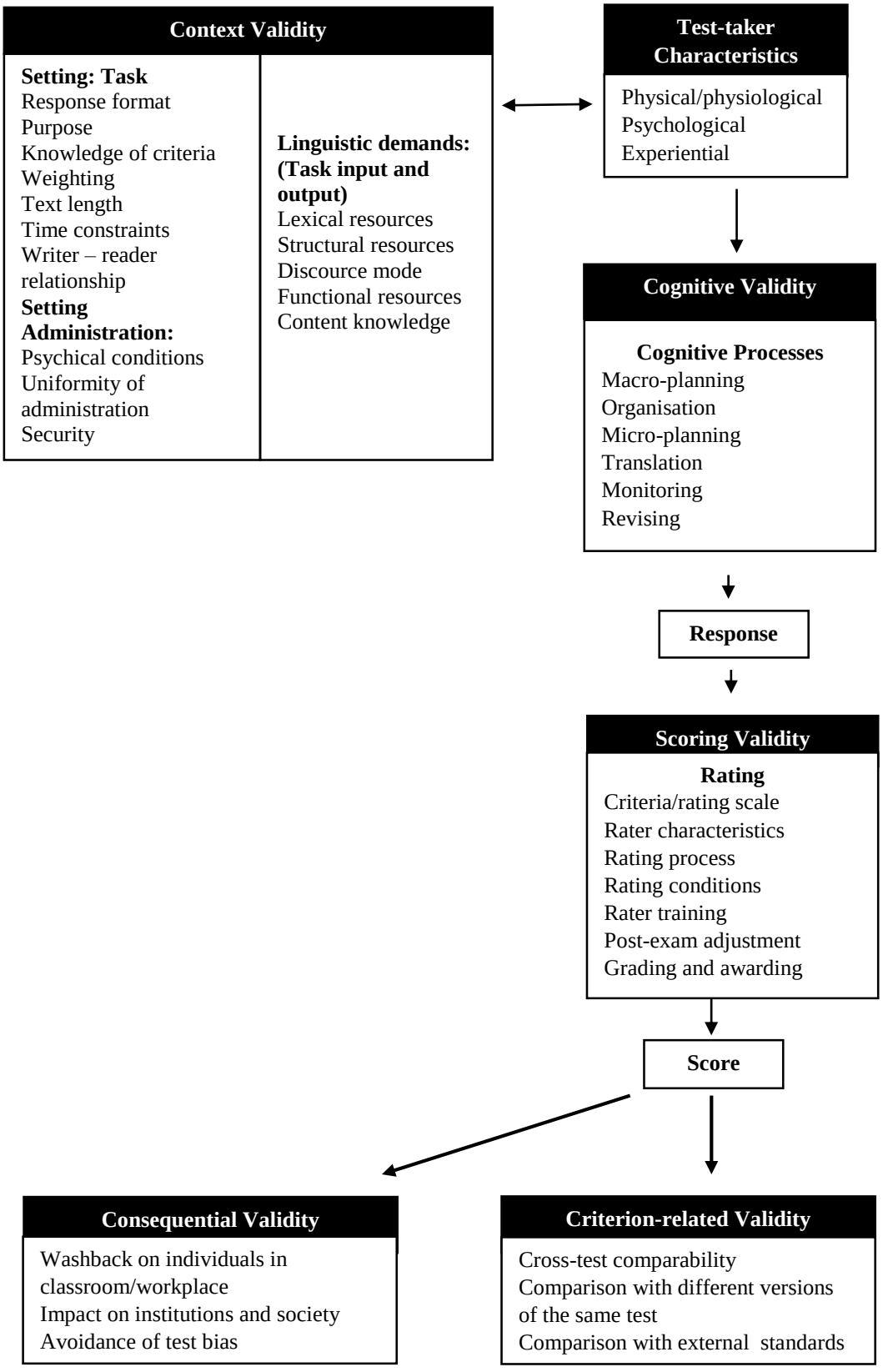


**Figure 2.1.** *Framework for conceptualizing writing test performance (Shaw & Weir, 2007, p. 4)*

According to Shaw and Weir (2007), there surely is a symbiotic relationship that exists between context, cognitive, and scoring validity. Considering the ultimate aim of the present study which was to develop an alternative theoretically-based and an empirically-validated multi-trait writing rubric that could serve to measure writing proficiency more validly and reliably at BUU-SFL-IEP, scoring validity and the writing rubric was dealt with more elaborately within the limited scope of this research. Scoring validity deals with all the facets of the testing process that can have an impact on the reliability of test scores (Shaw & Weir, 2007). Specifically, it explains "the extent to which scores are grounded on appropriate criteria, demonstrates consensual agreement in marking, are free as possible from measurement error, stable over time, consistent in terms of content sampling, and bring about confidence as reliable decision-making indicators" (p. 143).

According to Weigle (2002), two of these facets are of great significance for reliable scoring: defining the criteria in a writing rubric and ensuring that raters use the rubric properly and consistently.

### 2.2. Rubrics

Rubrics are an essential tool for all language teachers in this age of communicative and task-based language teaching and assessment (Brown, 2012). They can do for a curriculum what objectives do – they can assist in explaining terms and clarifying expectations (Crusan, 2010). They enable teachers and testers to efficiently communicate to learners or test takers what is expected from them in the productive language abilities of speaking and writing and then effectively assess those abilities for a variety of purposes, such as general proficiency, placement, achievement, and so forth. If used properly, they may assist in increasing the reliability of performance assessment and setting a common standard and meaning for the scoring process (Alderson et al., 1995). According to Brown (2012), rubrics are not only important to teachers in classrooms and administrators in language programs, but also to the development of the entire language teaching profession due to the relationships between large-scale proficiency frameworks and rubrics.

### 2.2.1. Definition of a Rubric

A rubric – referred to in some contexts as a rating scale or scoring guide – is defined as an instrument that lists specific criteria for scoring speaking or writing performance, and a guide that describes a particular level of performance within a scale (Crusan, 2015). To the best knowledge of the researcher of this study, the most comprehensive definition of this essential tool is given by Davies et al:

> "A scale for the description of language proficiency consisting of a series of constructed levels against which a language learner's performance is judged. Like a test, a rubric provides an operational definition of a linguistic construct such as proficiency. Typically such scales range from zero mastery through to an end-point representing the well-educated native speaker. The levels or bands are commonly characterized in terms of what subjects can do with the language (tasks and functions which can be performed) and their mastery of linguistic features (such as vocabulary, syntax, fluency and cohesion). Rubrics are descriptions of groups of typically occurring behaviors; they are not in themselves test instruments and need to be used in conjunction with tests appropriate to the population and test purpose. Raters or judges are normally trained in the use of proficiency rubrics so as to ensure the measure's reliability" (Davies et al, 1999, p. 153).

Teachers/raters face with many issues when scoring writing performance, such as what to weigh in making judgments, equity and fairness, and comparability of assessment; that is, will one teacher's evaluation of a student's work match another's evaluation (Crusan, 2010). Due to issues like this, assessors on all scales, large and small, turn to rubrics as they may increase objectivity in addition to reliability and validity in scoring. Both Crusan (2015) and Davies et al. (1999) emphasize that reliability and validity will not improve unless there is training on effective rubric creation and use.

Brown defines rubrics in line with the two broad types of rubrics; that successively are, *holistic* and *analytic*:

> "A *rubric in language teaching* is typically a grid set up in one of two ways (a) with scores along one axis of the grid and language behavior descriptors inside the grid for what each score means in terms of language performance or (b) with language categories along one axis and scores along the other axis and language behavior descriptors inside grid for what each score within each category means in terms of language performance" (Brown, 2012, p. 1).

In the following section, where the process of rubric design is explained, an overview of both types of rubrics are provided in addition to the third type, *trait-based*, as described in the pertinent literature.

### 2.2.2. The rubric design process

The CEFR conceptualizes and classifies the rubric development into three methodologies: *intuitive*, *qualitative*, and *quantitative* (Hawkey & Barker, 2004, pp. 128-129). Intuitive methodologies rely on other rubric samples. Qualitative methodologies depend on focus groups to collect information about the distinguishing features of different levels of writing proficency and how to describe them in the rubric. Quantitative methodologies are based on empirical methodologies, such as Many Faceted Rasch Measurement (MFRM), to associate test takers' proficiency levels with rubric descriptors on an integer scale (pp. 128-129). The rubric development methodology adopted in this study was both qualitative and quantitative. While the researcher analyzed rubric samples that are exemplified in the relevant research and used in global large-scale examinations, it would not be right to state that the process was based on intuitive methodologies.

For the specifics of the rubric design, two resources in particular formed the backbone of the process of writing rubric development which was in consistence with the contextual needs and the perceptions of the teacher-raters at BUU-SFL-IEP: Brown (2012) and Knoch (2011). While Knoch (2011) provides the reader with a more general framework, Brown (2012) guides the reader through the practical aspects of the process by providing a comprehensive list that is broken up into major stages and the minor steps that are part of each stage.

Knoch (2011) lists the elements that need to be considered while designing a rubric for writing assessment in pursuance of Weigle:

1. What type of a rubric is to be used?
2. Who is going to use the rubric?
3. What aspects of writing are most important and how will they be divided up?
4. How many scoring levels (bands) will be used?
5. How will scores be reported? (Weigle, 2002, pp. 122-125)

According to Knoch (2011), each of the elements listed above needs to be weighed judiciously in order for a rubric to be valid. Each element is explained in detail in the following five subsections.

#### 2.2.2.1. *What type of a rubric is to be used?*

Brown (2012) emphasizes that the choice between designing a holistic or analytic/multi-trait rubric impacts all the other elements in the scoring process; therefore, the decision should be made early and seriously.

Two general scoring approaches determine the type of rubric that is to be utilized during the scoring process: *norm-referenced* and *criterion-referenced* (Hyland, 2003). Criterion referenced procedures may take different forms and are categorized into three types:

- *holistic*,
- *analytic*, and
- *trait-based* (Weigle, 2002, pp. 108-39).

See Section 1.1. for an explanation of the two types of general scoring approaches and three types of rubrics, the last of which is further categorized into two as *primary-trait* and *multiple-trait*.

As recommended by Brown (2012), the type of rubric to be used should be decided on the basis of the purposes of the writing assessment in the first place. The opinions of the experienced teacher-raters in an institution should also be consulted because involving teacher-raters in the process will not only create teacher buy-in to the rubric but also add ideas on the issues of what categories to include, ways to describe the behavior in a particular category at a particular band level and so forth (See Methodology, Phase 1 for the exploration of the perspectives of the teacher-raters at BUU-SFL-IEP on an effective writing rubric). Crusan (2010) adds that rubrics can be even more powerful when they are created specifically for each asssignment and when created with students.

Hamp-Lyons (1991) who has put forward the multiple-trait scoring for the first time lists six assets of multiple-trait instruments as follows:

- *Salience*: features to be assessed can be determined by different writing contexts depending on the writing qualities deemed to be important.
- *Reality and community*: the scoring is based on the raters' agreement on the construct of what writing is.
- *Reliability*: multiple-trait scoring enhances the reliability of single composite number scores built from its components.
- *Validity*: multiple-trait scoring satisfies the construct and content validity since it reflects the accurate measurement of the behavior which defines the construct, and also the traits in the multiple-trait scoring derive from concrete expectations in the specific writing context.
- *Increased information*: performance on different components of writing is assessed and reported.

- *Washback*: the increased accuracy and the details of the information provided by the multiple-trait scoring can bring about the positive effect on teaching.

In her argument for multiple trait scoring, Hamp-Lyons (2016b) makes an anology between holistic scoring and building a house with only one brick and adds that multiple-trait scoring needs to be considered as an option for carrying out the important work of making fair decisions about the quality of written work, particularly in high-stakes contexts. Two advantages of analytic/multi-trait rubrics are listed in the pertinent literature. The first advantage is that in multi-trait scoring raters are required to provide separate scores for each of several facets or traits of the performance as opposed to holistic scoring where raters judge a performance impressionistically according to its overall properties (Davies et al., 1999). The second advantage is that according to Hamp-Lyons (1995, 2016a, 2016b), while holistic scoring is appropriate for scoring first-language (L1) essays, multiple-trait scoring has higher validity and reliability when rating second or foreign language (L2) essays because different learners have different levels of proficiency in different aspects of L2 writing.

#### 2.2.2.2. *Who is going to use the rubric?*

Rubrics have three purposes in measurement: describing the level of performance, guiding assessors how to rate performance, and providing test designers with information on test specifications (Bukta, 2014, p. 53). Alderson (1991b, pp. 72-74) lists three functions of the rubric depending on who uses them:

1. *Constructor-oriented* rubrics are meant to guide the tester in the creation of tests at appropriate levels and include reference to the kinds of writing tasks that examinees would be expected to encounter.
2. *Assessor-oriented* rubrics are meant to guide the rating process and focus on comparing the written text with the descriptors on the rubric.
3. *User-oriented* rubrics are written with a focus on providing useful information to help test users understand test scores.

A rubric needs to be assessor-oriented in the first place as it is designed for assessors' use; however, a "parallel rubric" that is user-oriented may be required to guide learners to interpret test scores (Knoch, 2011).

#### 2.2.2.3. *What aspects of writing are most important and how will they be divided up?*

As expressed by Lantolf and Frawley (1985), the validity of a rubric will be limited unless the underlying framework of the rubric takes account of linguistic theory and research in the definition of proficiency. Thus, the rubric that will be used to assess writing performance is an implicit or explicit reflection of the theoretical framework the test is based on (Weigle, 2002). As McNamara points out "the communicative movement has found performance assessment to be its natural accompaniment" (McNamara, 2002, p. 221); hence, researchers make reference to communicative competence modeling for the textual features which a rubric for the performance-based assessment of writing must have (See Chiang, 1999, 2003; East, 2009; Knoch, 2007, 2011).

Canale and Swain (1980) elucidate the theoretical framework of communicative competence, which consists of *grammatical*, *sociolinguistic*, and *strategic competences,* in their groundbreaking work titled *Theoretical Bases of Communicative Approaches to Second Language Teaching and Testing*. Grammatical competence is defined as "knowledge of lexical items and of rules of morphology, syntax, sentence-grammar semantics, and phonology" (p. 29). Sociolinguistic competence refers to "knowledge of sociocultural appropriateness of meanings" (p. 30). Lastly, strategic competence comprises "verbal and non-verbal communication strategies that may be used to compensate for communication breakdowns for some reason" (p. 30).

Canale (1983) expanded the model to incorporate *discourse competence*. His definition of discourse competence comprised "*cohesion* (the use of grammatical links) and *coherence* (the use of appropriate combination of communicative functions) of groups of utterances" (p. 338). The construct of discourse competence has usually been referred to as cohesion and coherence in ESL/EFL writing pedagogy, as well (Chiang, 2003). Medve and Takac (2013) reached the conclusion that the mastery of cohesion and coherence is of great significance for the definition of discourse competence in their review of the concept and role of discourse competence, which is based on the prominent communicative competence models, such as Canale (1983), Bachman (1990) and CEFR (2001). Tanskanen (2006, p. 7) defines cohesion as follows: "Cohesion, an important linguistic resource in the expression of coherent meaning, consists of grammatical and lexical elements on the surface of a text that can form connections between parts of the text (Tanskanen, 2006, p. 7)", whereas coherence is referred to as "the consistency of the elements as a mental representation of the text that is created by the reader" (McNamara

et al., 2010, p. 60). It can be concluded that cohesion and coherence are of great importance for ESL/EFL writing pedagogy because a written text will be meaningful only if the writer is able to construct sentences accurately and use cohesive devices properly to form a coherent text (Modhish, 2012; Kalajahi & Abdullah, 2015).

Weigle (2002, p. 122) supports that "More detailed information about various aspects of language use would be particularly appropriate when the focus of the assessment is on the acquisition of specific language skills, such as in low-proficiency non-acdemic classes or general foreign-language instruction". Therefore, *cohesion* and *coherence* are included in Knoch's (2011) taxonomy together with *accuracy*, *fluency*, *complexity*, *mechanics*, and *content* just like Chiang's (1999, 2003) and East's (2009) taxonomies.

In addition to cohesion and coherence, another important construct that is emphasized in the extant literature is *argumentation*. Based on the writing components of standardized tests such as the IELTS, the TOEFL, and the GRE, Hirvela concludes that "The capacity to write effective argumentative essays is an important marker of second language writing ability as argumentation is at the heart of second language writing assessment" (2017, p. 69), which is also pointed out by different theorists and researchers (e.g., Bacha, 2010; Liu & Stapleton, 2014; Qin & Karabacak, 2010; Wingate, 2012). Although it is seen as critical in academic environments and relied so heavily on to tell how well students write academically, there is a lack of consensus on how argumentation is defined and what must be included in an argumentative essay (Hirvela, 2017, p. 69; Plakans & Gebril, 2017, p. 85).

As for the conceptualization of argumentation, Hirvela (2017, p. 71) proposes two "interesting and important options" that could help second language writing teachers based on a distinction in the L1 argument scholarship: *learning to argue* and *arguing to learn*. This dichotomy looks at argumentation through different lenses. According to the first paradigm, which is based on the groundbreaking work of Toulmin and known as "Toulmin model" (1958, 2001), argument is a form of *reasoning*, and the emphasis is on argument as a *product*. The second paradigm by Kuhn (1991, 2005) sees argument as a form of *inquiry*, a tool or means toward a larger end, not a product. Hirvela (2017) supports that the Toulmin model, which sees argumentation as a combination of reasoning/product, deserves discussion in the second language writing field for two reasons: it helps students understand key aspects of logic and ways of using them to build

convincing arguments, and it stresses the uses of reasoning or logic to produce a well-structured argument essay, with an emphasis on the final result. The Toulmin model has six elements of a persuasive argument: *ground (data), claim, warrant, backing, qualifier,* and *rebuttal* (1958, 2001). The *claim* refers to the initially stated conclusion, which is controversial and subjective. The *ground (data)* are the facts supporting the claim. The *warrants* establish connections between the data and the claim. *Backings* state the assumptions on which the warrants rest. *Qualifiers* set limits on the strength of the claim, and *rebuttals* are arguments that refute or exceptions to the elements of the argument. A number of studies on students' argumentative writing have adapted or simplified the Toulmin framework because some of the elements in the model have been found to be overlapping or classified under more than one element in students' argumentative writing (Stapleton & Wu, 2015). (See Bacha, 2010; Liu & Stapleton, 2014; Qin & Karabacak, 2010; Stapleton & Wu, 2015 for the examples of the adapted Toulmin model in action in the L2 writing context). Among these studies, Qin and Karabacak's (2010) stands out because of the paucity of research on L2 university argumentative writing and the sound argument framework they proposed based on the Toulmin model (Hirvela, 2017; Liu & Stapleton, 2014; Stapleton & Wu, 2015).

In their model six Toulmin elements are reshaped as follows:

> *Claim*: An assertion in response to a controversial topic or a problem.
>
> *Data*: Evidence to support a claim, such as facts, statistics, anecdotes, research studies, expert opinions, definitions, analogies, and logical explanations.
>
> *Counterargument claim*: The possible opposing views that can challenge the validity of a writer's claim; these opposing views can also be supported by data.
>
> *Counterargument data*: Evidence to support a counterargument claim.
>
> *Rebuttal claim*: Statements in which the writer responds to a counter-argument by pointing out the possible weakness in the claim, data, or warrant, such as logical fallacies, insufficient support, invalid assumptions, and immoral values.
>
> *Rebuttal data*: Evidence to support a rebuttal claim (Qin & Karabacak, 2010, p. 449).

The latest CEFR Companion Volume with new descriptors (2018, p. 142) highlights the significance of counterargument starting from the level of B1: *Can introduce a counterargument in a discursive text*. In an article devoted to counterargumentation in argumentative writing in a high-stakes test, Lui and Stapleton (2014) also emphasize that counterargumentation is a key factor contributing to the persuasiveness of argumentative essays and propose that counterargumentation be

considered in the writing prompts and scoring rubrics of high-stakes English tests in addition to classroom instruction on argumentative writing.

Major standardized assessments also include these constructs – more or less – in the rubrics for the performance-based assessment of writing. See Table 2.1 below for the main traits of scoring rubrics for six tests of ESL writing (Haswell, 2007, p. 111).

**Table 2.1.** *Main Traits of Scoring Rubrics for Six Tests of ESL Writing (Haswell, 2007, p. 111)*

| **Test** | **Trait** |
|---|---|
| Test in English Educational Purposes (Associated Examining Board) | Content |
| | Organization |
| | Cohesion |
| | Vocabulary |
| | Grammar |
| | Punctuation |
| | Spelling |
| Certificate in Communicative Skills in English (Royal Society of Arts/University of Cambridge Local Examinations Syndicate) | Accuracy [of mechanics] |
| | Appropriacy |
| | Range [of expression] |
| | Complexity [organization and coherence] |
| Test of Written English (Educational Testing Service) | Length |
| | Organization |
| | Evidence |
| | Style |
| | Grammar |
| | Sentences |
| Michigan English Language Battery | Topic development |
| | Sentences |
| | Organization/coherence |
| | Vocabulary |
| | Mechanics |
| Canadian Test of English for Scholars and Trainees | Content |
| | Organization |
| | Language use |
| International English Language Testing System | Register |
| | Rhetorical organization |
| | Style |
| | Content |

While some theorists, such as Broad (2003) and Kohn (2006), argue that rubrics used in high-stakes testing reduce writing to a formula that moulds instruction and perception of writing, others support that rubrics are relevant to help learners prepare to move past gatekeepers, and they could be "starting points from which we make our own rubrics" (Crusan, 2010, p. 44).

After the categories of the rubric are finalized, the tester/the rubric designer should make a decision on how to divide them up (Weigle, 2002). This process of dividing the categories up is labelled as *weighting*, that is "the awarding of value to certain items, assessment criteria, tasks, or sub-tests" (Davies et al, 1999, p. 225). As Davies et al. (1999) highlights, the test designer might want to give some components more weight in the total score, which demonstrates a perception of the relative prominence of the various test components.

#### 2.2.2.4. *How many scoring levels (bands) will be used?*

*Level* or *band* is defined as "a measure (e.g., 1 to 9 or A to E) or description of the proficiency or ability of a test taker, normally as described on some kind of scale and determined on the basis of test performance" (Davies et al., 1999, p. 107). Knoch (2011) supports that the context where the rubric will be used needs to be the determining factor in the number of levels to be included in a rubric. Having said that, Knoch also suggests the ideal number of levels on a rubric on the basis of the purpose:

> "If a scale for a writing test is administered to a very varied ability group of test takers, the seven (plus or minus two) rule is applicable. However, if the scale is to be used at certain proficiency level, three to four categories may be sufficient. The guiding principle here should be the usefulness of the feedback provided to test takers/users" (Knoch, 2011, p. 92).

Once the number of levels is decided on, the following task of the rubric designer is to form the descriptors for each level (Weigle, 2002). Davies et al. (1999, p. 43) defines *descriptor* as "a statement which describes the level of performance required of candidates at each point on a proficiency scale". According to Knoch (2011, p. 94), "a concrete and objective formulation style should be used" if higher rater reliability is desired. Descriptors in the form of a checklist may be a way to achieve this, or raters may be provided with an assessor guide that further explains the phrases in the descriptors, as it was done in Phase 3 and Phase 4 of the current research.

#### 2.2.2.5. *How will the scores be reported?*

According to Weigle (2002), if analytic scoring is utilized, scores in each category can be reported in a total score, or scale scores can be provided separately for diagnostic purposes. Weigle (2002) supports that "reporting separate scores provides more useful diagnostic information and generally provides more accurate picture of test takers' abilities in writing" (p. 124). Additionally, Knoch (2011) recommends the provision of an in-depth description of a test taker's writing behavior in the different categories of the rubric.

All in all, the five elements that a rubric designer is recommended to consider could be summoned as follows:

- The type of the rubric to be used should be determined depending on the purpose of assessment, contextual needs, and teacher-rater expectations. If the purpose is to inform learners about their strengths and weaknesses in different aspects of writing and to provide them with useful feedback, an analytic/multi-trait rubric could be more appropriate.
- The rubric should be both assessor- and user-oriented.
- The rubric should be based on a theory or model of language development.
- There should be adequate number of levels depending on the context where the rubric is to be used, and level descriptors should be concise and clear for not only the assessor but also the user.
- The way scores are reported should provide the test takers with as much feedback as possible.

As can be seen, Knoch (2011) provides a general framework that consists of 5 vital questions to guide the rubric designer. Once these questions are answered, the process of rubric development could be started. According to Brown (2012), a comprehensive list of the steps involved in rubric design will be easier to understand if it is divided into major stages and minor steps that are part of each stage. See Table 2.2 below for Brown's suggested stages and steps in the rubric development process.

**Table 2.2.** *Suggested stages and steps in rubric development process (Brown, 2012, p. 18)*

| stage | step |
|---|---|
| 1: Planning | 1.1. Define the goal. |
| | 1.2. Go to the source material. |
| | 1.3. Brainstorm. |
| | 1.4. Analytic or holistic? |
| | 1.5. Decide the categories. |
| | 1.6. Decide the range of scores to be used. |
| 2: Designing the rubric | 2.1. Put scores on one axis. |
| | 2.2. Put the categories on the other axis. |
| | 2.3. Fill in the rubric descriptors for each score/band level. |
| 3: Planning the assessment procedures and using the rubric | 3.1. Decide on the stimulus formats. |
| | 3.2. Decide on the response formats. |
| | 3.3. Write clear instructions. |
| | 3.4. Make sure the instructions and stimulus materials are ready. |
| | 3.5. Arrange for the mechanics of assessment. |
| | 3.6. Actually do the assessment. |
| | 3.7. Train raters to use the rubric. |
| 4: Evaluating the reliability/ fairness of the rubric | |
| 5: Evaluating the quality of the rubric | 5.1. Evaluate the validity of the rubric. |
| | 5.2. Evaluate the usability of the rubric. |
| 6: Planning feedback and revise for pedagogically useful ratings | 6.1. Plan for student and teacher feedback. |
| | 6.2. Set up a cycle of revision and improvement. |

Both Brown (2012) and Knoch (2011) guided the present research in the different phases of the study. Knoch (2011) focuses more on the rubric design process; that is, she poses and answers the general questions that a rubric designer needs to consider before the design process begins. Brown (2012), on the other hand, refers also to stages that need to be followed to evaluate the quality of the rubric after it is used, which brings the rubric designer to the stage of rubric validation.

### 2.2.3. Rubric validation

Validation is the process of confirming the validity of a test, which is one of the basic concerns of language testing (Davies et al., 1999). Validity is defined as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical

rationales support the adequacy and appropriateness of inferences and actions based on test scores" by Messick (1989, p. 13), a prominent figure in the validity argument, who proposed a *unified concept of validity* for the first time. The unified concept of validity entails taking into consideration all aspects of the testing situation that have an impact on test performance, such as the specification of the construct domain with regard to topical knowledge, test specifications, and criteria for scoring (Tsai, 2015). Thus, an ongoing rubric anlaysis to validate the rubric is a necessity in contexts that use high-stakes performance-based assessment of writing proficiency (Janssen, Meier, & Trace, 2015).

The increasing use of performance-based, constructed-response tasks in language tests in the 1970's and 1980's brought in applications of new methods for investigating the quality of assessments (Xi & Davis, 2017). Cronbach (1971, 1984) and Messick (1975, 1980) were the pioneers of this movement. Kane's (1992, 2006, 2013) *argument-based approach to test validation*, Bachman's (2005) and Bachman and Palmer's (2010) the concept of an *assessment use argument* that was built on Kane's theoretical work, and Weir's (2005) and Shaw and Weir's (2007) *socio-cognitive framework* followed the movement. The current research is guided by Shaw and Weir's (2007) framework due to its particularity and multidimensional approach to the assessment of writing performance and its unified approach to establishing the overall validity of a test. See figure 2.2. below.

Shaw and Weir (2007, p. 2) build their framework on Cambridge ESOL's traditional approach to validating tests, namely "the VRIP approach" in which the concern is with *Validity* (the conventional sources of validity evidence: construct, content, criterion-related), *Reliability*, *Impact*, and *Practicality*. The construct validity of a language test is a manifestation of how representative it is of a basic theory of language learning, which includes analyzing how tests scores might be interpreted with regard to the theoretical framework underlying the construct the test is designed to measure (Davies et al., 1999). Construct validity might also be said to involve content validity, which is defined as a conceptual or non-statistical validity based on systematic review of the test content to find out whether it consists of a sufficient sample of the target domain to be measured (Davies et al., 1999). Criterion-related validity refers to the statistical establishment of a new test with regard to the closeness of a test to an external measure, such as a well-known test within the same domain (concurrent validity) or a future test (predictive validity) (Davies et al., 1999).

What differentiates Shaw and Weir's (2007) socio-cognitive framework from traditional approaches is its endeavor to redesign validity to demonstrate how its components (context, cognitive processing, and scoring) interact with each other. Moreover, it conceptualizes the validation process in a *temporal frame* which classifies the various types of validity evidence to be collected at each phase of the test design, monitoring, and evaluation cycle. The last but not the least, the framework defines the construct more specifically than tradional approaches do (p. 3).

The socio-cognitive framework consists of both *a priori* (before-the-test event) validation constituents of context and cognitive validity and *a posteriori* (after-the-test event) validation constituents of scoring validity, consequential validity, and criterion-related validity (Shaw & Weir, 2007; Weir, 2005). According to Weir:

> "The more comprehensive the approach to validation, the more evidence collected on each of the components of this framework, the more secure we can be in our claims for the validity of a test. The higher the stakes of the test the stricter the demands we might make in response of all of these" (Weir, 2005, p. 47).

Shaw and Weir list the critical questions that need to be addressed in the application of the socio-cognitive framework as follows:

- How are the physical/physiological, psychological, and experiential characteristics of the test-takers maintained by this test?
- Are the cognitive processes needed to complete the test tasks suitable? (cognitive validity)
- Are the characteristics of the test tasks and their administration is sufficient and fair to the test-takers? (context validity)
- How far are the test scores from the test reliable? (scoring validity)
- How do the test and test score affect various stakeholders? (consequential validity)
- What external evidence is there outside of the test scores themselves that the test is objective? (criterion-related validity) (Shaw and Weir, 2007, p. 4)

In the literature of language testing, there are several studies utilizing Weir's (2005) socio-cognitive framework as a basis for test validation most of which focus on the testing of reading or listening (e.g., Bannur, Abidin, & Jamil, 2015; Geranpayeh & Taylor, 2013; He & Jiang, 2020; Khalifa & Weir, 2009; Weir, Hawkey, Green, & Devi, 2009; Akşit, 2018). Akşit (2018) is special in this respect because it is one of the very few attempts to generate validity evidence using Weir's (2005) framework for the reading section of a high-stakes test in the context of an intensive English program of an English medium Turkish state university. Regarding the validation of performance-based assessment of speaking and writing, there is a paucity of research at both the local and

global level. Thus, studies using Weir's (2005) framework is also quite limited (e.g., Chan, 2011; Ghanbari, Barati, & Moinzadeh, 2012; Nakatsuhara, 2013; Shaw & Falvey, 2008; Shaw & Weir, 2007; Taylor & Galaczi, 2011; Zainal, 2012). To the best knowledge of the researcher of the current study, there does not exist any research into the validation of performance-based assessment of writing in Turkish EFL context where Weir's (2005) socio-cognitive framework is used.

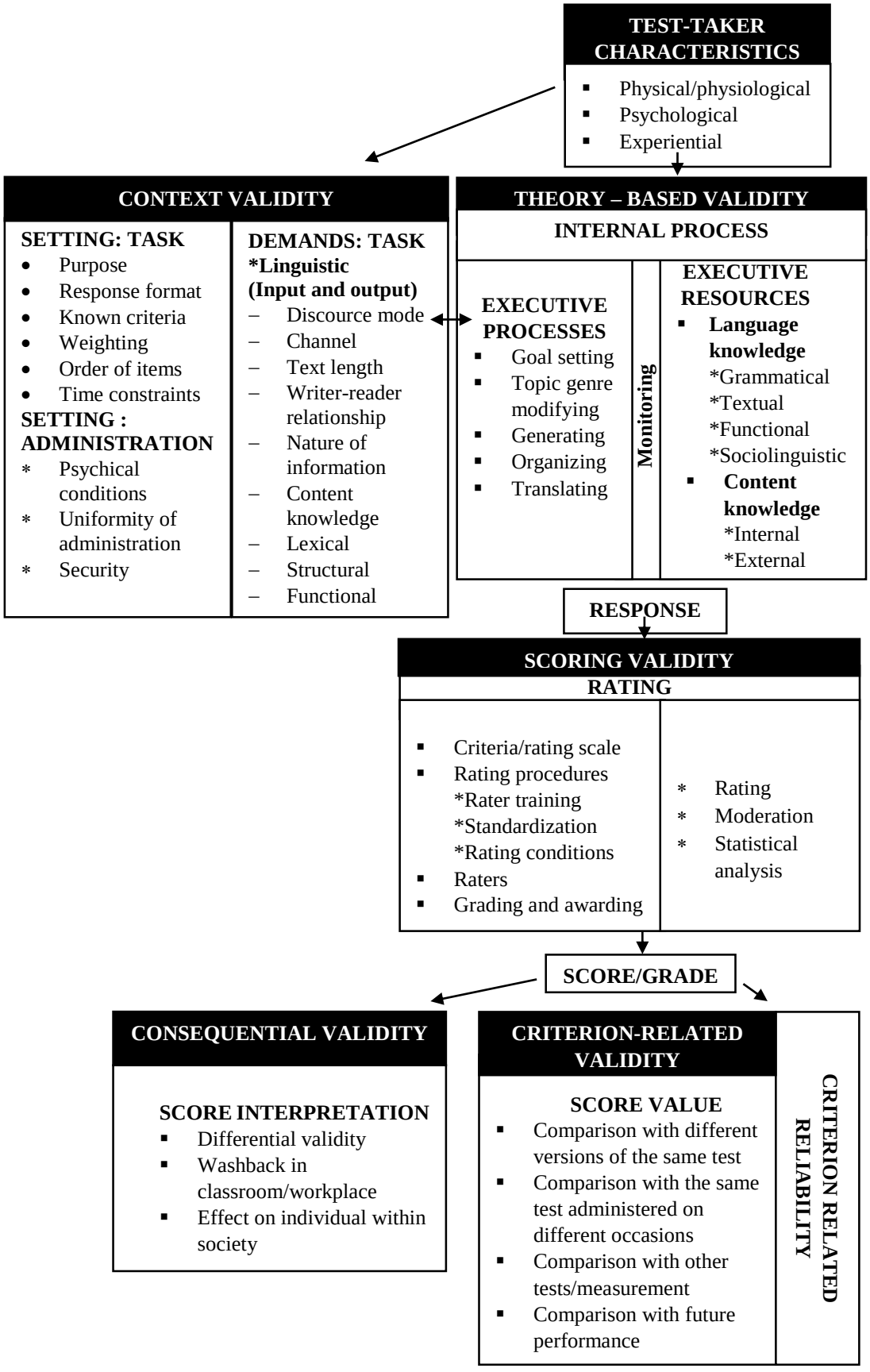


**Figure 2.1.** *A socio-cognitive framework for validating writing tests (Weir, 2005, p. 19)*

With respect to the relatively recent research specifically into writing rubric validation, the pertinent literature indicates that different validation frameworks for evaluating the validity of performance-based assessments have been utilized. For instance, while Hattingh (2009) predominately adopted Weir's (2005) socio-cognitive framework, Becker (2018) and Janssen, Meier, and Trace (2015) followed Kane's (2006) argument approach to validation. However, at an empirical level, all three studies and many others have used Many Faceted Rasch Measurement (MFRM) to statistically verify the reliability and validity of the rubrics that were utilized to assess writing performance in various contetxs (e.g., Eckes, 2015, Edmond, 2012; Knoch, 2007; Küçük, 2017; Wind & Peterson, 2018). Küçük (2017) used Weir's (2005) socio-cognitive framework and MFRM to investigate and provide theoretical and empirical evidence for validity and reliability of the two writing tasks to assess academic writing proficiency of learners of Turkish as a foreign language. To the best knowledge of the reseacher, there does not exist a context-specific rubric design and validation process in Turkish EFL context that used a theoretical framework and MFRM.

#### 2.2.3.1. *Many Faceted Rasch Measurement (MFRM)*

Many Faceted Rasch Measurement (MFRM) is a general psychometric modelling approach that is described "particularly well-suited to dealing with many-facet data typically generated in rater-mediated assessments" (Eckes, 2015, p. 19). McNamara (1996) calls such influences in performance assessment as *facets,* i.e., the features of the assessment setting. He supports that the interactions of these facets may determine the probability of specific test scores. In this respect, using a resourceful approach like MFRM which allows the researcher to closely examine each of these facets and their interrelationships might prove to be very useful to evaluate the psychometric quality of many-facet data. See Figure 2.3. by Eckes (2015, p. 49) below.

As displayed in the figure, there are *distal facets* which might influence the ratings in an indirect way and *proximal facets* which have immediate effect on the scores awarded to test takers (Eckes, 2015). According to Eckes (2015, p. 52), "MFRM modeling generally provides a well-structured and detailed account of the role played by each facet (proximal and/or distal) that is deemed relevant in a given assessment context".

Furthermore, a MFRM analysis facilitates a number of useful indices for investigating the functioning of rubrics, which is the ultimate aim of the current research.

Brown and Edmonds encapsulates this information that MFRM provides in a set of questions. According to these researchers, a MFRM analysis can be very beneficial for analyzing the results of rubric-based assessments since the analysis answers the following questions:

1. How are the performances of examinees, raters, and categories related when they are placed on an equal interval scale?
2. How able are the examinees relative to the rubric categories, and vice versa?
3. To what degree are the different raters scoring in the same ways?
4. How severe or lenient are the raters compared to each other?
5. To what degree are the different categories producing similar results?
6. How difficult or easy are the categories compared to each other?
7. How well are the different scores on the rubric distinguished from each other? (Brown and Edmonds, 2012, p. 81)

Knoch (2007) explains why classical test theory and generalisability theory (G-theory), another statistical approach that can be used with rater-mediated assessment, have limitations when compared to MFRM. According to Knoch, MFRM is superior to ANOVA-based and regression approaches because possible interaction effects in ANOVA can contaminate main effects and make the interpretation of the main effects more difficult. MFRM, on the other hand, can go beyond the main effects and interaction effects since it makes the detection of individual level effects possible.

As it is with MFRM, G-theory can identify sources of varience ascribed to each facet and its interactions. However, the effect of such differences on the test takers' scores during a specific examination is not adjusted unlike it is in MFRM. That is, the test takers receive the raw scores from the raters that they encounter in G-theory, but they get the corrected raw scores in MFRM.

For the above-mentioned reasons, MFRM is adopted in the current research.

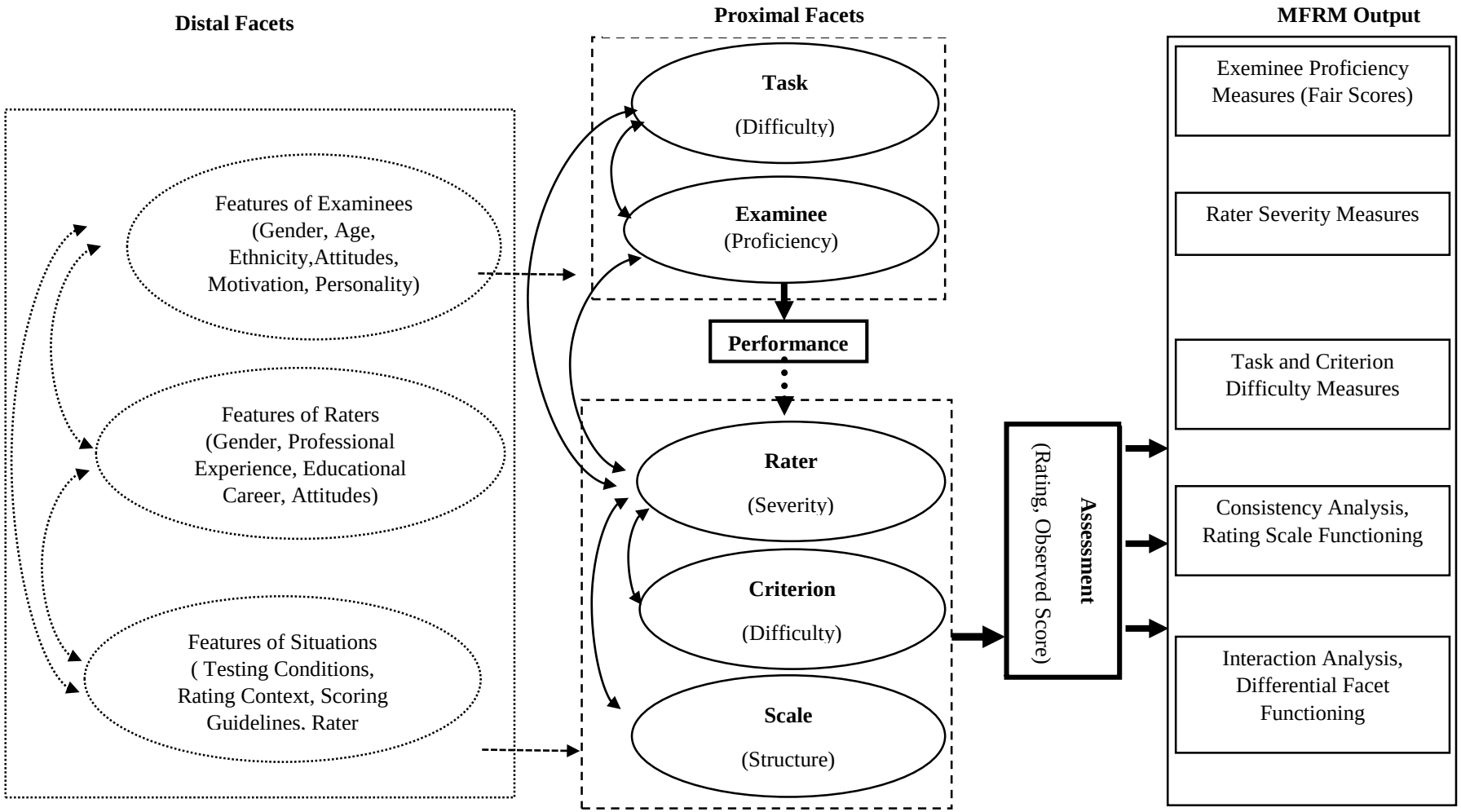


**Figure 2.2.** *A conceptual-psychometric framework for rater-mediated assessment (Eckes, 2015, p. 49)*

## 3. METHODOLOGY

### 3.1. General aim of the study

A context-sensitive rubric validation where the type of rubric to be used needs to be judiciously decided is a requirement for institutions that use high-stakes performance assessment of writing proficiency. The overarching aim of the current research was to develop an alternative theoretically-based and empirically-validated multi-trait rubric for such a high-stakes test, the performance-based assessment of EFL writing proficiency at BUU-SFL-IEP administered at the end of each academic year as part of the English language proficiency examination at the B1+ proficiency level. Validation requires the evaluation of an instrument based on a variety of quantitative and qualitative forms of evidence to support inferences from test scores (Weir, 2005, p. 15). For this reason, pure mixed-methods research (hereafter MMR) approach to gathering and reporting data was utilized for the purposes of the study through the complementary application of both qualitative and quantitative research methods in an equal and a balanced way.

Before explaining different phases in detail, we explained the context of performance-based assessment of writing proficiency at BUU-SFL-IEP so that the rationale behind the overarching aim of the study could be elucidated.

Since 2010, an adapted version of the analytic rubric, ESL Composition Profile (See Appendix 1), which was developed by Jacobs, Zinkgraf, Wormuth, Hartfield, and Hugley (1981, p. 30), has been used for the assessment of performance writing component of the proficiency examination at BUU-SFL-IEP. In this section of the exam, students are required to write an argumentative essay. Adaptations were made in the original rubric by the writing course coordinator and the testing unit head for the sake of practicality, which can be listed as follows.

To start with, in the original rubric the total score is calculated out of 100, but in the adapted version the total score is 20 since the writing section of the proficiency test comprises 20% of the total score. Another difference between the original and adapted versions is the number of categories. While there are five categories in the original rubric (i.e., content, organization, vocabulary, language use, and mechanics), there are four categories in the adapted one. The category of mechanics where spelling, punctuation, and capitalization are assessed is not included in the adapted rubric. Thirdly, the number of band levels differ in two rubrics. In the original rubric, there are four levels ranging from excellent to poor (i.e., excellent to very good, good to average, fair to poor, and very

poor) whereas in the adapted version there are only three (i.e., good to average, fair to poor, and very poor). Thus, excellent to very good level is excluded in the adapted version. Fourthly, the weightings of each category are different in the two rubrics. In the original rubric, the categories of content and language use have the highest weightings, yet in the adapted one only content has the highest weighting. The categories of organization, vocabulary, and language use are weighed equally. Finally, in the adapted version of the rubric, there is an additional section for student essays that are irrelevant with the topic(s) in the prompt, which does not exist in the original rubric. If a student essay is evaluated by the rater to be irrelevant with the topic in the prompt, it is scored 1 out of 20. To sum up, it can be concluded that the original and the adapted versions of the ESL Composition Profile (Jacobs et al.,1981) differ in terms of the total score they assign for a performance, the number of categories, the number of levels, and the weightings that they assign for each category in addition to an extra section in the adapted version for student essays that are not relevant with the topic in the prompt.

As also mentioned in the introduction chapter of the current research, several concerns arose concerning the use of this adapted version of the ESL Composition Profile (Jacobs et al., 1981) in terms of its context validity, scoring validity, construct validity, and hence its consequential validity. The first problem was related to the context validity because the adapted version of the rubric was "intuition-based", that is it was not a locally designed and locally controlled one but an adapted version of a rubric created to be used in another context (Crusan, 2010, p. 72; Janssen, Meier, & Trace, 2015, p. 53; Knoch, 2007, p. 5). The second problem was related to the scoring validity of the rubric, one of the most important indicators of which was the lack of consensual agreement between raters (Weir, 2005). In the recent years, it has been observed that the number of essays re-evaluated by the third examiner increased drastically, which indicated a lack of agreement between raters. Raters pointed out two problems in relation to this issue; one was the difficulty of scoring, and the other one was the number and weighting of the different categories, which is related to the construct validity.

All these inevitably affected the consequential validity of the scoring process in a negative way. As a result, the necessity to design a locally-controlled rubric that could cater to the needs of the specific context of this study arose.

An overview of the process, which consisted of five phases, is presented below:

Phase 1: Exploration of raters' perspectives on the current writing rubric, i.e., the adapted version of ESL Composition Profile, and their expectations from an alternative writing rubric,

Phase 2: Development of a new rubric,

Phase 3: Trial and refinement of draft rubric and open-ended questionnaire,

Phase 4: Psychometric analysis of the new rubric through Many Faceted Rasch Measurement (MFRM),

Phase 5: Exploration of raters' perspectives on the new rubric.

While the first two phases were the a-priori constituents of the validation process, the other three were the a-posterioi constituents. For each of the five phases, the sections of aim(s), participants, instrument(s), procedure and data collection, and data analysis were explained individually since each of these phases had a different purpose to serve for the general aim of the study.

### 3.2. Phase 1: Exploration of raters' perspectives on the adapted version of ESL Composition Profile

#### 3.2.1. Aim

As mentioned above in the general aim of the study, raters referred to some problems in relation to the use of the adapted version of the analytic rubric ESL Composition Profile by Jacobs et al. (1981, p. 30). In this regard, there were two aims:

- To receive more detailed information from the raters on the problems that they voiced in different platforms i.e., the perceived strengths and weaknesses of the writing rubric used currently in the proficiency examination and the qualities of an effective writing rubric, as perceived by the raters at BUU-SFL-IEP and
- To contribute to the assessment literacy of the participants, which is defined by Hamp-Lyons (2018) as the "ability and willingness to ask and answer important questions about fundamental issues in assessment", such as what and who are assessed, what tool or tools would be suitable for a specific context and so on.

To these ends, the first phase of the study adopted qualitative methodology and employed a qualitative research instrument which yielded qualitative data.

### 3.2.2. Participants

Participants in the study were 24 university instructors teaching EFL at BUU-SFL-IEP who took part in the study on voluntary basis. See Appendix 2 for an example consent form that was signed by each participant. All 24 participants of the first phase of the current research had ELT experience ranging from 5 to 33 years. Thus, they could all be classified as experienced based on the extant literature (Tsui, 2003, 2005; Richards, Li, & Tang, 1998). Specifically, one participant had 5 years, one participant had 7 years, one participant had 9 years, and twenty-one participants had 10 or more years of experience in the ELT field. Following Lim (2011), experience was used in this paper to refer the length of time a rater had been rating or to the amount of rating a rater had done. Therefore, raters with at least five years of EFL teaching and rating experience at BUU-SFL-IEP were considered "experienced raters". This means the participants in the current research were not only experienced teachers but also experienced raters. Regarding gender, only four participants out of 24 were male. Their age range was mostly from 30 to 39 years except for one participant who was between 20-29 age ranges, 6 participants who were between 40-49 age ranges, and one participant who was over 50. Out of twenty-four participants, 14 of them had MA degrees. While 11 of them had MA in ELT, one of them had an MA degree in Translation Studies, one had an MA degree in Educational Management, and one in Women Studies. All participants were non-native speakers of English (For more detailed information see Appendix 3).

### 3.2.3. Instrument

The data was collected through an open-ended questionnaire in the spring semester of the 2017-2018 academic year. An open-ended questionnaire adapted from Knoch (2007) was used in order to explore the strengths and weaknesses of the writing rubric used currently in the proficiency examination, the efficacy of the rubric in fair assessment of students' written work, the raters' confidence level in using the rubric, and their expectations from an alternative rubric, as perceived by the experienced EFL instructors at BUU-SFL-IEP. Knoch (2007) designed a questionnaire in order to elicit raters' perceptions of the measurement efficacy and usability of a rubric that she designed for diagnostic assessment of L2 writing proficiency at a university in New Zealand. Because the present study and Knoch (2007) had similar purposes, the content of Knoch's questionnaire was mostly retained except for a few modifications that resulted from the

piloting and expert consultation processes explained below. In essence the items in Knoch's instrument were chiefly yes-no questions which did not require in-depth explanations in response. For the purposes of this phase of the study, those questions were modified into wh- questions where participants were asked to provide detailed information for their answers. Unlike it was in Knoch's (2007) instrument, the participants of this phase of the study were also requested to indicate the categories that should be involved in a writing rubric from a list of writing categories that represent a variety of writing constructs that are stated to be pivotal in assessing writing in the related literature (Brown, 2012; Knoch, 2007; 2009; 2011).

Based on McKay (2006), the open-ended questionnaire was piloted with 6 instructors who had 5 to 16 years of experience in the field of ELT so that the potential problems, such as the clarity of items, could be discovered. Once the piloted questionnaire was received, an item analysis was carried out evaluate the effectiveness of each item in gathering the required data, as suggested by Zacharias (2012). This analysis revealed only few minor changes in the wordings of some items. Following Huck (2004), three experts, two professors and one associate professor in an ELT department of a Turkish state university, were consulted before the distribution of the instrument for data collection so that the reliability and validity of the instrument could be ensured.

The open-ended questionnaire took its final shape based on the feedback gained through the piloting process and the expert consultation (See Appendix 4). The questionnaire consisted of two parts. Part 1 elicited demographic information on the participants. Part 2 comprised questions addressing the perceived strengths and weaknesses of the writing rubric, the efficacy of the rubric in fair assessment of students' written work, the raters' confidence level in using the rubric, and the categories that should be involved in an effective writing rubric according to the participants. There were 8 questions in total. In each part, there were fill-in questions followed by short-answer questions in order to gather in-depth information on the aforementioned topics.

### 3.2.4. Procedure and data collection

After the consent necessary to conduct the study was taken from the presidency of BUU and the administration of BUU-SFL-IEP, the research process began (See Appendix 5). The researcher who was also an instructor at BUU-SFL-IEP distributed the questionnaires to the instructors who taught there and volunteered to participate in the

study. These instructors were given brief information on the research process. They were requested to hand in the filled-in questionnaires in a ten-day period. 40 questionnaires were distributed in total, and 24 filled-in questionnaires were received at the end of the ten-day period. Each filled-in questionnaire was numbered in order to organize the data for easy retrieval and identification and anonymize it.

### 3.2.5. Data analysis

The data that was gathered through the open-ended questionnaire included fill-in questions and short-answer questions. Before the data analysis process began, the researcher first familiarized herself with the data by reading the whole data a few of times, as Creswell (2012) and Lacey and Luff (2007) recommended. In order to analyze the fill-in questions, the frequency distribution of the responses for each question was counted, and the data was presented in tables in the results and discussion section. For the analysis of the short-answer questions, as Knoch (2009) suggests, the broad, overarching coding categories or themes were devised a priori based on the content of the short-answer questions (such as the strengths and weaknesses of the writing rubric used currently in the proficiency examination). The responses of the 24 participants for each short-answer question were read thoroughly and subcategories under the broad categories were identified and highlighted with the help of color-coding as follows: Each subcategory was given a different color, and participants' responses were looked through by color-coding them accordingly. The data analysis process ended with the finalization of subcategories that the color-coding revealed and the exploration of the relationships between the broad categories and subcategories. Due to the manageable amount of data obtained from a relatively small number of participants, the data was analyzed by hand rather than a software program like Maxqda or NVivo. Another reason why such a program was not utilized was the determination of overarching categories and themes a priori.

Following Barber and Walczak (2009), peer debriefing was performed with 20% of the data so that a check against biases within the analysis could be made. Another reason of peer debriefing was to aid with consistency, credibility, and reliability throughout the coding process in order to strengthen the trustworthiness of the work. The peer debriefer in the present study was a colleague of the researcher, who had a PhD degree from an ELT program of an English-medium university. She helped the researcher by reviewing the coding of the 20% of the open-ended questionnaires (which was equal to 6 coded

questionnaires in total). As a result of the peer debriefing process, the selection of the categories was verified. As many quotes as possible by the participants were provided in the results and discussion section so that each category could be exemplified.

### 3.3. Phase 2: Development of a new rubric

#### 3.3.1. Aim

As mentioned above in Phase 1, goals and objectives and source materials were visited before the rubric design process commenced. Taking into consideration these contextual requirements of BUU-SFL-IEP and the expectations of the participants from a writing rubric based on the results of Phase 1 of the study, the aim of the second phase was to design a theoretically-based rubric with the guidance of experts in performance-based assessment of writing proficiency and the relevant literature (Banarjee et al., 2015; Becker, 2011, 2018; Brown, 2012; Chiang, 1999, 2003; Crusan, 2010; East, 2009; Hamp-Lyons, 1991, 1995, 2016a, 2016b, 2018; Hattingh, 2009; Janssen, Meier, &Trace, 2015; Knoch, 2007, 2011; Weigle, 2002; Weir, 2005; Weir & Shaw, 2007).

#### 3.3.2. Participants

In addition to the researcher of the current study, following Banarjee et al. (2015), three experts from different institutions and five experienced-raters at BUU-SFL-IEP were also asked to contribute to the process of writing rubric development by giving their opinions on the draft rubric designed by the researcher of this study. Their participation in the process was on voluntary-basis. In the selection of the participants, experience and expertise in EFL writing, performance-based assessment of EFL writing, and writing rubric development were taken into consideration. Therefore, raters with at least five years of EFL teaching and rating experience at BUU-SFL-IEP were considered "experienced raters". Expertise, on the other hand, refers here to people with particular competence in a subject area, i.e., EFL writing, performance-based assessment of EFL writing, and writing rubric development for the purposes of the current research. Two professors (one from an ELT department of an English- medium university in Turkey and the other from a TEFL department of an English-medium university in Japan) and a freelance teacher trainer specialized in performance-based assessment of EFL proficiency and educational assessment and evaluation contributed to this phase of the study with their invaluable opinions. One of the professors was a non-native speaker of English while

the other professor and the freelance teacher-trainer were native speakers of English. All three experts had more than thirty years of experience in the field of ELT. As stated by Huck (2004: 89), expert opinion is valuable because it aids in ensuring the reliability and validity of an instrument. In-depth personal information about the instructors participating in this phase of the study is presented in the table provided in Appendix 6.
In addition to experts and experienced-raters, the writing course coordinator and the head of the testing unit at BUU-SFL-IEP were also consulted during the process of rubric development because of their expertise in the specific context of this research and key positions in the assessment practices in the institution.

### 3.3.3. Instruments

Two instruments were used in the current phase of the study: a general information document given to the participants and the draft rubric designed by the researcher of the study based on the results of Phase 1 which explored the perceptions of the teacher-raters on the writing rubric that is currently in use and their expectations from a well-functioning writing rubric in addition to the extant literature on the development of a rubric for performance-based assessment of writing.

#### 3.3.3.1. *General information document*

The five experienced teacher-raters and the three experts who were the participants of this phase of the study were provided with a general information document so as to familiarize them with the context, aim, and phases of the current research (See Appendix 7 for the document). A colleague of the researcher, who worked as an instructor in an IEP of an English-medium state university and held a PhD degree from the same university, assisted the researcher by reading the document and giving her opinions about the clarity and scope of it, as an outsider. After few revisions with wording, the document took its final shape.

#### 3.3.3.2. *Draft rubric*

The researcher designed the first draft of the new rubric which was an assessor-oriented analytic/multi-trait rubric with five categories (Content, Organization, Grammar, Vocabulary, and Mechanics), five-band levels (0, 1, 2, 3, 4) and concrete descriptors (See Appendix 8 for the first draft of the new writing rubric).

#### 3.3.3.3. *Procedure*

Before sharing the draft rubric with the three experts, the writing course coordinator, the head of the testing unit, and the five experienced teacher-raters were requested to state their opinions in a brief meeting held by the researcher on the following aspects of the draft rubric:

- The categories in the rubric,
- The weighting of these categories,
- The number of bands levels in each category, and
- The descriptors written for each band level.

Based on their feedback, no revisions were made on the draft rubric as they stated that they found the rubric satisfactory apart from the modification of the font size. Thus, the second draft of the new rubric was completed (See Appendix 9). Together with the second draft of the new rubric, general information on the context, aim, and phases of the study were shared with the three experts from different institutions to ask for their opinions on and modifications of the new rubric. After a three-month period, the experts sent their opinions via electronic mail. The second-draft of the new rubric was modified based on the recommendations made by the three experts (See Appendices 10 for the refinements by the first expert and 11 for the refinements by the second expert on the draft rubric). Thus, the third draft of the new rubric which would be piloted in Phase 3 of the current research was designed.

#### 3.3.3.4. *Reporting of expert opinion*

Each of the three experts sent their opinions on the various aspects of the draft rubric mentioned above and in-depth explanations of their opinions via e-mail. The comments, suggestions, and recommendations of each expert were reported in section 4.2. Development of a new rubric as part of the results and discussion section.

In conclusion, Phase 2 resulted in the draft writing rubric to be used in Phase 3: Trial and refinement of draft rubric, the topic of the following section.

### 3.4. Phase 3: Trial and refinement of draft rubric

#### 3.4.1. Aim

The aim of this phase of the study was two folded:

1. To trial the draft rubric through an initial pilot implementation, identify possible weaknesses in the use of it, and refine the rubric based on the feedback that would be received from the teacher-raters as carried out in Knoch (2007) and Hattingh (2009) and
2. To pilot the open-ended questionnaire that was used in Phase 5 of the study to explore the perceptions of the teacher-raters on the efficacy of the draft rubric and discover the potential problems that may exist in the open-ended questionnaire, such as the clarity of the items. (Zacharias, 2012, p. 71).

Because the raters who would trial the draft rubric were going to be asked their perceptions on the rubric, piloting the open-ended questionnaire with these five teacher-raters was found to be useful before using the instrument with the thirteen raters in Phase 5 of the study.

#### 3.4.2. Participants

Five experienced teacher-raters who worked at BUU-SFL-IEP and who were not participants in the previous phases of the study participated in the third phase on voluntary basis (See Appendix 2 for an example consent form that was signed by each participant). Following Lim (2011), the definition of experience was considered as the length of time a rater had been rating or to the amount of rating a rater had done, as it was in the first of phase of the study, which explored the raters' perspectives on the current writing rubric and their expectations from an alternative rubric. Thus all five participants had experience in EFL teaching and rating ranging from 11 to 28 years, and they could all be classified as experienced. Specifically, one participant had 11 years, one participant had 12 years, one participant had 17 years, one participant had 18 years, and one participant had 28 years of experience in the ELT field. All participants were non-native speakers of English. Regarding gender, all participants were female. The ages of three participants ranged from 30 to 39 years while the age of one participant ranged from 40-49 years, and one participant was +50. Out of five participants, 4 of them had MA degrees. While 3 of them had MA degrees in ELT, one of them had an MA degree in Gender Studies (See Appendix

12 for the in-depth personal information about the instructors participating in this third phase of the study).

### 3.4.3. Instruments

The instruments used for this phase of the study were as follows: an assessor guide for the assessment process, student essays, rating sheets, scores assigned by each rater to each essay, and an open-ended questionnaire to discover the perceptions of the raters on the use of the draft rubric.

#### 3.4.3.1. *Assessor guide for the draft rubric*

The five participants were provided with an assessor guide on how to use the rubric including the definitions of constructs represented in each category and explanations for phrases in the descriptors for clarification in a brief meeting held by the researcher. During the meeting, the participants were given the opportunity to ask questions and make comments on the assessment process to pilot the draft rubric. The meeting was not in the form of a training and norming session following Chiang's (2003, p. 475) recommendation, which emphasizes that "elaborate training would be inappropriate, as it might introduce an undesirable effect in the raters' decision-making processes". Another reason for not having a training and norming session with the participants was to discover the extent to which the draft rubric could function effectively on its own (See Appendix 13 for the assessor guide for the draft rubric).

#### 3.4.3.2. *Student essays*

Over 1000 students took the 2017-2018 academic year English proficiency exam at BUU-SFL-IEP. The researcher made a visit to nine classes from three different language proficiency levels one week before the proficiency examination was administered. There was a total of 225 students in the nine classes. 182 of the students agreed to participate in the study voluntarily with the essays they were going to write in the exam. The consent of all 182 students were received by the researcher before the exam took place. See Appendix 14 for an example consent form that was signed by each student. Out of 182 student essays written by the students who sat the exam, 10 student essays were selected by the researcher in a way that could represent each score level from 0 to 4 since this was the piloting phase of the draft rubric. Therefore, essays which could be considered

irrelevant or which lacked a paragraph or more were also included in the batch of the 10 essays unlike the essays selected for the fourth phase of the study, Phase 4: Psychometric analysis of the new rubric through Many Faceted Rasch Measurement (MFRM). Each of the ten essays was numbered to anonymize the data and also for easy retrieval and identification.

#### 3.4.3.3. *Rating sheets*

Based on the categories in the draft rubric, a rating sheet was prepared for scoring. The five raters were requested to note down on this sheet both the scores that they assigned for each of the categories for each essay and the total score that they assigned for each essay. See Appendix 15 for a sample rating sheet.

#### 3.4.3.4. *Scores assigned by each rater to each essay*

Each rater rated each of the 10 student essays using the assessor guide for the draft rubric, the draft rubric, and the rating sheet. Both the scores that the raters assigned for each of the categories for each essay and the total score that they assigned for each essay generated the quantitative data that was used for the psychometric analysis of the draft rubric.

#### 3.4.3.5. *Open-ended questionnaire*

An open-ended questionnaire adapted from Knoch (2007) was used for the collection of the qualitative data in the spring semester of the 2018-2019 academic year. Knoch (2007) designed a questionnaire to elicit raters' perceptions of the measurement efficacy and usability of a rubric that she designed for diagnostic assessment of L2 writing proficiency at a university in New Zealand. Since the present study and Knoch (2007) had similar aims, the content of Knoch's instrument was maintained by and large except for a few minor modifications that resulted from the expert consultation process. The items in Knoch's instrument were chiefly yes-no questions which did not require in-depth explanations in response. For the purposes of this phase of the study, those questions were modified into wh- questions where participants were asked to provide detailed information for their answers.

In order to ensure the reliability and validity of the instrument, an expert, who was an assistant professor in an ELT department of a Turkish state university and whose area

of special interest was qualitative research, was consulted before the distribution of the instrument for data collection (Huck, 2004, p. 89). One item was excluded and two items were included in the open-ended questionnaire used in this phase of the research following the feedback gathered from the expert.

The open-ended questionnaire took its final shape based on the feedback gained through the expert consultation. The questionnaire consisted of two parts. Part 1 elicited demographic information on the participants. Part 2 comprised questions addressing the perceived strengths and weaknesses of the draft writing rubric, the efficacy of the rubric in fair assessment of students' written work, and the raters' confidence level in using the rubric. There were 7 questions in total. In each part, there were fill-in questions followed by short-answer questions in order to gather in-depth information on the aforementioned topics (See Appendix 16 for the open-ended questionnaire).

### 3.4.4. Procedure and data collection

In order to start the third phase of the study, the teacher-raters who volunteered to participate in the third phase of the study were visited in their offices by the researcher, given brief information on the research process and a file including the assessor guide for the draft rubric, the draft rubric, 10 student essays, rating sheets, and the open-ended questionnaire and requested to go over the documents in the file before their next meeting. Then, the researcher and the five teacher-raters determined a meeting day together according to their teaching schedules. This meeting was held in the form of a panel discussion, where the participating teacher-raters were able to ask any questions and evaluate assessment process in any way they would like. After making sure that they did not have any concerns with regard to the assessment process, they were given a week to complete the scoring process and another week to fill in the open-ended questionnaire. At the end of the two-week period, raters were requested to hand in the files together with the filled-in rating sheets and the open-ended questionnaire.

### 3.4.5. Data analysis

Two types of data were collected in this phase of the study: the quantitative data gathered through the scoring process of the essays and the qualitative data gathered through the open-ended questionnaire. For the analysis of the quantitative data, a computer program called FACETS (Version 3.80.3; Linacre, 2017) was used. The

program admits data files, with data suitably formatted, coming from various sources, such as Excel, R, SAS, SPSS, or STATA (Eckes, 2015). In order to prepare the input data consisting of ratings which 5 raters awarded to essays written by 10 examinees in a live examination, the researcher entered the scores that raters assigned for each of the categories for each essay and the total score that they assigned for each essay into an Excel file. See Appendix 17 for a sample of the file.

Following Knoch (2007, p. 203) and Eckes (2015, p. 152), a fully crossed design was selected to make the MFRM analysis used in the study more powerful. That is, all 5 raters rated the same 10 student essays. Such a design is not a requirement of FACETS to run the analysis; however, it enhances the stability of analysis and the quality of the conclusions that can be drawn from the results (Eckes, 2015; Myford & Wolf, 2003).

FACETS used the scores that raters awarded to examinees on each of the five categories (i.e., Content, Organization, Vocabulary, Grammar, and Punctuation, Spelling, and Mechanics) so as to measure individual examinee proficiencies, rater severities, category difficulties, and band level difficulties. The program calibrated the examinees, raters, rubric categories, and band levels onto the same equal-interval scale, which is called a logit scale, a variable map, or the Wright map and which enables direct comparisons between, and within, the facets under consideration. In addition, a MFRM analysis provides fit statistics for each facet to identify misfitting persons, raters, or categories. The term *misfit* describes examines, raters, or categories that do not fit the predicted pattern of responses by the model. Each of the tables demonstrates its own list of elements and make estimations for each arrangement of each facet (Edmonds, 2012). A higher score equals a positive logit, which is a higher measure, and a lower score equals a negative logit, which is a lower measure. The extreme score measurement was set at 0.3 for MFRM. Convergence was set at 0.1 score points, the smallest observable difference between raw scores and 0.01 logits, the smallest useful difference. The analysis achieved convergence after fifty-nine iterations, which enabled the configuration of an accurate measurement system.

The analysis of the qualitative data gathered through the open-ended questionnaire was made by following the procedure used in Phase 1 (Exploration of raters' perspectives on the current writing rubric). See Section 3.2.5. The analysis of the qualitative data in Phase 1 was more demanding due to the higher number of participants in that phase (24), while there were five participants in this phase. Broadly, there were two types of questions

in the open-ended questionnaire: fill-in questions and short answer questions. In order to analyze the data gathered through the open-ended questionnaire including fill-in questions and short-answer questions, the researcher first familiarized herself with the data by reading the whole data a couple of times, as suggested by Creswell (2012) and Lacey and Luff (2007). For the analysis of the fill-in questions, the frequency distribution of the responses for a particular question was counted, and the data was presented in tables in the results and discussion section. For the analysis of the short-answer questions, as Knoch (2009) suggests, the broad, overarching coding categories or themes were devised a priori based on the content of the short-answer questions (such as the strengths and weaknesses of the draft rubric). As it was in Phase 1, the responses of the 5 participants for each short-answer question were read thoroughly and subcategories under the broad categories were identified and highlighted with the help of color-coding as follows: Each subcategory was assigned a particular color, and participants' responses were gone through by color-coding them accordingly. At the final stage of the data analysis process, emerging subcategories highlighted through color-coding were finalized, and the relationships between the finalized categories and subcategories were explored. Because of the feasible amount of data gathered from a small number of participants, the data was analyzed by hand instead a software program like Maxqda or NVivo. Another reason why such a program was not used was the determination of overarching categories and themes a priori.

### 3.5. Phase 4: Psychometric analysis of the new rubric through Many Faceted Rasch Measurement (MFRM)

#### 3.5.1. Aim

The aim of the fourth phase, which yielded quantitative data, was to empirically validate the alternative multi-trait rubric designed for the performance-based assessment of writing proficiency at BUU-SFL-IEP through statistical analyses, specifically MFRM, a general psychometric modelling approach that is described "particularly well-suited to dealing with many-facet data typically generated in rater-mediated assessments" (Eckes, 2015, p. 19). The writing section of the English language proficiency exam at BUU-SFL-IEP was also an example of performance-based assessment where a variety of factors, variables, or components of the measurement situation might affect test scores in a systematic way. Overall, the aim of Phase 4 of this study was to investigate the effects of

these facets so that the reliability and the validity of the new rubric could be verified quantitatively.

### 3.5.2. Participants

Thirteen experienced teacher-raters who worked at BUU-SFL-IEP and who did not participate in the previous phases of the current research took part in the fourth phase of the study on voluntary basis (See Appendix 2 for an example consent form that was signed by each participant). As it was in the prior phases, instructors who had at least five years of EFL teaching and rating experience were considered "experienced raters". All 13 participants had experience in EFL teaching and rating ranging from 5 to 20 years. Therefore, they could all be classified as experienced. Specifically, one participant had 5 years, one participant had 7 years, one participant had 8 years, and ten participants had 10 or more years of experience in the ELT field. All participants except one were non-native speakers of English. Regarding gender, only three participants out of 13 were male. The ages of six participants ranged from 30 to 39 years while 5 participants were between 40-49 age ranges. The remaining three participants were between 20-29 age ranges. Out of thirteen participants, 8 of them had MA degrees. While 7 of them had MA degrees in ELT, one of them had an MA degree in Linguistics. In-depth personal information about the instructors participating in this phase of the study is presented in the table provided in Appendix 18.

### 3.5.3. Instruments

Following instruments were used for the empirical validation of the rubric: an assessor guide for the new rubric, student essays, rating sheets, and scores assigned by each rater for each category in each essay and total scores assigned for each essay.

#### 3.5.3.1. *Assessor guide for the new rubric*

For the purposes of the current research, a training and norming session was not carried out with the participants, i.e., 13 the teacher-raters, in this phase of the study following Chiang's (2003, p. 475) advice, which emphasizes that "elaborate training would be inappropriate, as it might introduce an undesirable effect in the raters' decision-making process". Another reason for not having a training and norming session with the participants was to discover the extent to which the new rubric could function effectively on its own. Instead, the raters were provided with an assessor guide for the new rubric

that gave the definitions of the constructs represented in each category and the explanations of the phrases in the descriptors for clarification. Based on the feedback received from the participants in Phase 3 (Trial and refinement of the draft rubric) and Jacobs et al. (1981), the assessor guide used in this fourth phase of the study was more detailed than the one utilized in Phase 3 (See Appendix 13 for the assessor guide used in Phase 3 and Appendix 19 for the one utilized in this phase of the study).

#### 3.5.3.2. *Student essays*

Over 1000 students took the 2017-2018 academic year English proficiency examination at BUU-SFL-IEP. The researcher visited nine classes from three different language proficiency levels one week before the proficiency examination. The total number of the students in the nine classes were 225. 182 of the students accepted that the researcher would include these students' essays in the analysis for research purposes. The consents of all 182 students were received by the researcher before the examination took place. See Appendix 14 for an example consent form that was signed by each student. The 10 essays which were utilized in the third phase of the study where the draft rubric was piloted were excluded from the batch of 182. Out of 172 student essays written by the students who took the exam, fifty student essays which had five paragraphs and were not considered irrelevant were selected by the researcher. Each of the fifty essays was numbered to anonymize the data and also for easy retrieval and identification.

#### 3.5.3.3. *Rating sheets*

Based on the categories in the new rubric, a rating sheet was prepared for marking. Raters were asked to note down on this sheet both the scores that they assigned for each of the categories for each essay and the total score that they assigned for each essay. See Appendix 15 for a sample rating sheet.

#### 3.5.3.4. *Scores assigned by each rater to each essay*

Each rater rated each of the 50 student essays using the assessor guide for the new rubric, the new rubric, and the rating sheet. Both the scores that the raters assigned for each of the categories for each essay and the total score that they assigned for each essay generated the quantitative data that was used for the psychometric analysis of the new rubric.

### 3.5.4. Procedure and data collection

Primarily, the researcher who was also an instructor at BUU-SFL-IEP visited nine classes (three classes from each language proficiency level – A2, B1, and B1+) to give students brief information on the research process and asked them whether they would like to participate in the study by giving the researcher the consent necessary to use the essays which they were going to write in the forthcoming 2017-2018 academic year English proficiency examination. Out of 225 students, 182 of them accepted that the researcher would include these students' essays in the analysis for research purposes. After the exam was administered, a copy of each of the 182 student essays was made by the researcher. Fifty student essays which were relevant with the task and had five paragraphs were selected by the researcher from the pool of 182 student essays. Each of the fifty essays was numbered to anonymize the data and also for easy retrieval and identification. Thirteen experienced teacher-raters who worked at BUU-SFL-IEP and volunteered to participate in the study were visited by the researcher in their offices and given a file including the assessor guide for the new rubric, the new rubric (See Appendix 20), student essays, and rating sheets. Each participant was asked to go through the instruments in the file together with the researcher during her visit, and then, each of the participants was given a few days to examine the file, read the assessor guide carefully, and grade a few essays if they would like to and consult the researcher if they had any questions or queries. They were recommended to start rating after making sure that they were comfortable with the new rubric. They were given one month to score the fifty student essays. At the end of the one-month period, raters were requested to hand in the files together with the filled-in rating sheets.

### 3.5.5. Data analysis

The quantitative data gathered through the rating process explained above was analyzed by means of the computer program FACETS (Version 3.80.3; Linacre, 2017). The program accepts data files, with data correctly formatted, coming from a variety of sources, such as Excel, R, SAS, SPSS, or STATA (Eckes, 2015). In order to prepare the input data consisting of ratings which 13 raters awarded to essays written by 50 examinees in a live examination, the researcher entered the scores that the raters assigned for each of the categories for each essay and the total score that they assigned for each essay into an Excel file (See Appendix 17 for a sample of the file).

Following Knoch (2007: 203) and Eckes (2015: 152), a fully crossed design was selected to make the MFRM analysis used in the study more powerful. That is, all thirteen raters rated the same fifty student essays. Such a design is not a requirement of FACETS to run the analysis; however, it enhances the stability of analysis and the quality of the conclusions that can be drawn from the results (Eckes, 2015; Myford & Wolf, 2003).

FACETS used the scores that raters awarded to examinees on each of the five categories (i.e., Content, Organization, Vocabulary, Grammar, and Mechanics) in order to make an estimation of individual examinee proficiencies, rater severities, category difficulties, and band level difficulties. The program adjusted the examinees, raters, rubric categories, and band levels onto the same equal-interval scale, which is called a logit scale, a variable map, or the Wright map and enables direct comparisons between, and within, the facets under consideration. In addition, a MFRM analysis provides fit statistics for each facet to identify misfitting persons, raters, or categories. The term *misfit* describes examines, raters, or categories that do not fit the predicted pattern of responses by the model. Each of the tables demonstrates its own list of elements and make estimations for each arrangement of each facet (Edmonds, 2012).

A higher score equals a positive logit, which is a higher measure, and a lower score equals a negative logit, which is a lower measure. The extreme score measurement was set at 0.3 for MFRM. Convergence was set at 0.1 score points, the smallest observable difference between raw scores and 0.01 logits, the smallest useful difference. The analysis achieved convergence after fifty iterations, which enabled the configuration of an accurate measurement system.

## 3.6. Phase 5: Exploration of raters' perspectives on the new rubric

### 3.6.1. Aim

The aim of the final phase of the study, which yielded qualitative data, was to explore the raters' perspectives on the efficacy of the alternative multi-trait writing rubric designed for the performance-based assessment of writing proficency in the proficiency examination. The purpose of this phase was to not only elicit the opinions of the raters on the new rubric but also support or interrogate the quantitative findings gathered to validate the new rubric, as suggested in the relevant literature (Knoch, 2007; McNamara & Knoch, 2012).

### 3.6.2. Participants

Thirteen experienced raters who worked at BUU-SFL-IEP and who participated in the fourth phase of the study (Psychometric analysis of the new rubric through MFRM) took part in the fifth phase of the study on voluntary basis (Section 3.5.2). See Appendix 18 for in-depth information on the participants. As it was in the previous phases of the current research, instructors who had at least five years of EFL teaching and rating experience were considered "experienced raters" in pursuant of Lim (2011), Tsui (2003, 2005) and Richards, Li, and Tang (1998).

### 3.6.3. Instrument

The open-ended questionnaire used in Phase 3 of the study (Section 3.4.3.5) was retained as it proved to be effective in collecting the qualitative data required for the purposes of this research. Following McKay (2006, p. 44), the open-ended questionnaire was piloted in the Phase 3 of the current study (Trial and refinement of draft rubric) with 5 instructors who had 11 to 28 years of experience in the field of ELT in order to discover the potential problems that may exist, such as the clarity of the items. Upon receiving the piloted questionnaire, an item analysis was conducted to evaluate the effectiveness of each item in gathering the required data, as recommended by Zacharias (2012, p. 71). Based on this analysis, the open-ended questionnaire remained intact. In order to ensure the reliability and validity of the instrument, the expert, who gave her opinions on the questionnaire in Phase 3 of the study, was consulted one more time before the distribution of the instrument for data collection (Huck, 2004, p. 89). She stated that the open-ended questionnaire could be used intact.

The questionnaire consisted of two parts. Part 1 elicited demographic information on the participants. Part 2 comprised questions addressing the perceived strengths and weaknesses of the writing rubric, the efficacy of the rubric in fair assessment of students' written work, and the raters' confidence level in using the rubric. There were 7 questions in total. In each part, there were fill-in questions followed by short-answer questions in order to gather in-depth information on the aforementioned topics. See Appendix 21 for the instrument.

### 3.6.4. Procedure and data collection

At the end of the one-month period which teacher-raters were given to rate the fifty student essays for the completion of the Phase 4 of the study (Psychometric analysis of the new rubric through MFRM), they were also requested to answer the short-answer questions in the open-ended questionnaire designed to discover their perspectives on the different aspects of the new writing rubric. The researcher who was also an instructor at BUU-SFL-IEP distributed the questionnaires to thirteen teacher-raters on the day they would turn in the scores they assigned for the fifty essays they rated using the new writing rubric. The participants were requested to hand in the filled-in questionnaires in a ten-day period. 13 questionnaires were distributed in total, and 13 filled-in questionnaires were received at the end of the ten-day period.

### 3.6.5. Data analysis

The procedure used in Phase 1 (Exploration of raters' perspectives on the current writing rubric) and Phase 3 (Trial and refinement of draft rubric) were used for the analysis of the qualitative data gathered through the open-ended questionnaire. See Sections 3.2.5 and 3.4.5. Broadly, there were two types of questions in the open-ended questionnaire: fill-in questions and short answer questions. In order to analyze the data gathered through the open-ended questionnaire including fill-in questions and short-answer questions, the researcher first familiarized herself with the data by reading the whole data several times, as recommended by Creswell (2012) and Lacey and Luff (2007). For the analysis of the fill-in questions, the frequency distribution of the responses for a particular question was counted, and the data was presented in tables in the results and discussion section. For the analysis of the short-answer questions, as Knoch (2009) suggests, the broad, overarching coding categories or themes were devised a priori based on the content of the short-answer questions (such as the strengths and weaknesses of the draft rubric). As it was in Phase 1, the responses of the 5 participants for each short-answer question were read thoroughly and subcategories under the broad categories were identified and highlighted with the help of color-coding as follows: Each subcategory was assigned a particular color, and participants' responses were gone through by color-coding them accordingly. At the final stage of the data analysis process, emerging subcategories highlighted through color-coding were finalized, and the relationships between the finalized categories and subcategories were explored. Due to the manageable

amount of data obtained from a relatively small number of participants, the data was analyzed by hand rather than a software program like Maxqda or NVivo. Another reason why such a program was not utilized was the determination of overarching categories and themes a priori.

## 3. RESULTS AND DISCUSSION

In order to be able to achieve the overall aim of the study, which was to develop an alternative theoretically-based and empirically-validated multi-trait rubric for the performance-based assessment of EFL writing proficiency at BUU-SFL-IEP, the five phases that were pursued were as follows:

Phase 1: Exploration of raters' perspectives on the current writing rubric, i.e., the adapted version of ESL Composition Profile, and their expectations from an alternative writing rubric,

Phase 2: Development of a new rubric,

Phase 3: Trial and refinement of draft rubric and open-ended questionnaire,

Phase 4: Psychometric analysis of the new rubric through Many Faceted Rasch Measurement (MFRM),

Phase 5: Exploration of raters' perspectives on the new rubric.

Each phase had a different purpose to serve the general aim of the study. Therefore, the results and discussion of each phase was presented separately.

### 4.1. Phase 1: Exploration of raters' perspectives on the current rubric

24 teachers who worked as BUU-SFL-IEP as instructors and raters were the participants in the first phase of the current research (See Section 3.2.2 for in-depth information on the participants). The perspectives of these teacher-raters on the current rubric, i.e., the adapted version of ESL Composition, and their expectations from an alternative writing rubric were explored in this first phase of the study through an open-ended questionnaire. In this section, results of this first phase of the research are presented based on the four areas aimed to be explored by the open-ended questionnaire:

- the strengths and weaknesses of the writing rubric used currently in the proficiency examination,
- the efficacy of the rubric in fair assessment of students' written work,
- the raters' confidence level in using the rubric, as perceived by the experienced EFL instructors at BUU-SFL-IEP, and the last but not the least,
- the categories that should be involved in an effective writing rubric according to the participants.

Each of these areas was allotted a subsection for the ease of following, and findings were discussed in light of the pertinent literature.

### 4.1.1. Strengths of the writing rubric used currently in the proficiency examination

24 instructors who took part in the proficiency examination at BUU-SFL-IEP as raters were participants in the first phase of the study. According to the participants, there were three advantages of the writing rubric that is in use at present:

- Practicality,
- Categorization (of writing constructs that need to be assessed), and
- Objectivity.

As demonstrated in Table 4.1, out of 24 participants, 13 of them stated that practicality is the number one strength of the writing rubric, which was followed by categorization (of writing constructs that need to be assessed), indicated by 11 participants, and objectivity, mentioned by only four of the participants in the current research.

**Table 4.1.** *Strengths of the writing rubric used currently in the proficiency examination (N=24)*

| | ***F*** | ***%*** |
|---|---|---|
| **Practicality** | 13 | 54 |
| **Categorization** | 11 | 46 |
| **Objectivity** | 4 | 17 |

Another group of four participants stated that the current writing rubric did *not* have any strengths.

#### 4.1.1.1. *Practicality*

13 out of 24 participants in the study referred to *practicality* as the number one strength of the writing rubric. According to Participant 14 and Participant 20 (hereafter *P*):

> 1. It is practical and quick to use, so it allows me to save time considering the limited time allocated to grade papers *(P14)*.
> 2. It does not look complicated. It's short (maybe too short ☺) and practical *(P20)*.

However, it is important to note that the majority of the participants who stated practicality as an asset (9 out of 13) referred to it together with accompanying drawbacks in the scoring process. For instance, Participant 6, who thought the writing rubric was

practical, also questioned the importance of fairness over practicality, as shown by the following excerpt:

> 3. The only strength of the current rubric is that it is user-friendly and thus time saving. However, the proficiency exam plays a crucial role in students' exit scores. That's why our primary concern shouldn't be user-friendliness or saving time. Fair evaluation is what should matter in such an important exam *(P6)*.

Although practicality seems to have a priority while evaluating an assessment instrument, factors other than practicality may play an important role in a high-stakes test with a fail or pass result. As stated by Becker (2011, p. 127) in terms of the use of writing rubrics, "what is practical is not always what is best for our students and teachers", as also emphasized in excerpt (3) above.

#### 4.1.1.2. *Categorization*

Following practicality, categorization (of writing constructs that need to be assessed) was stated to be another asset of the writing rubric that is used currently in the proficiency examination. Out of 24 participants, 11 of them considered categorization as a strength. Among these 11 participants, 9 of them perceived categorization as an advantage because, thanks to categorization, raters knew which aspects of writing they needed to assess, as reflected in the excerpts that follow:

> 4. It helps me to evaluate different components of students' writing. It guides me to take these components into consideration and fairly set a point to an essay *(P7)*.
>
> 5. The scale helps me grade students' essays confidently because it tells me where to look at in each essay such as content, organization etc. It has strong guidance for its parts *(P13)*.

Out of 11participants who perceived categorization as a strength 2 of them mentioned the ease of giving feedback when needed as their reason:

> 6. When we are supposed to inform students about their scores, the sections in the scale give us an opportunity to justify the reasons of scoring for each section e.g., vocabulary and grammar (*P17*).
>
> 7. …because we score the written output and write points for each aspect (Content, Organization etc.), students can be explained their strengths and weaknesses in writing, which I believe helps them feel more secure. After writing these, now I feel I favor analytic rubrics (*P19*).

Thus, 11 participants in this study perceived categorization (of writing constructs that need to be assessed) as a strength of the writing rubric since it assisted them in pinpointing the aspects of writing that were required to be assessed and guided them when giving feedback about strong and weak aspects in students' writing, as also emphasized in the relevant literature. Having categories is a defining aspect of analytic/multi-trait scoring as opposed to holistic scoring where performance is assessed globally (Alderson et al., 1995, pp. 189-190). This quality of analytic/multi-trait rubrics makes them advantageous in two ways according to theorists (Brown, 2012, p. 35; Davies et al., 1999, p. 7; Hyland, 2004, p. 230; Knoch, 2011, p. 83; Weigle, 2002, p. 121):

- They provide more exact reporting of written or oral skills development;
- They lead to greater reliability since each test taker is given a number of scores for each category in the rubric.

However, having categories by itself is not sufficient for a reliable and valid scoring process. As suggested by Knoch (2011, p. 81), the most central consideration is what the rubric categories should look like, i.e., the aspects of writing that could be used to form the criteria of the rubric, which was explored thoroughly with question number two in the open-ended questionnaire used as the instrument of this phase of the present study.

#### 4.1.1.3. *Objectivity*

Objectivity was the last advantage of the writing rubric that is in use at present according to four participants in the study:

> 8. First of all, during scoring the essays, the rubric operates our cognition and prevents bias. After reading an essay for the first time, unintentionally, I may think of a holistic score about the paper, but after I apply the levels of the rubric, the score changes, and I feel safer and unbiased as a rater *(P19).*

The main reason why rubrics exist is their assumed assistance in increasing objectivity, reliability, and validity in scoring (Brown, 2012, p. 34; Crusan, 2015, p. 1), yet only four participants out of 24 stated objectivity as a strength of the writing rubric that is currently used in the proficiency exam. The majority of the participants indicated that they did not believe the scoring process was as objective, reliable, or fair as it should be in a high-stakes test like proficiency due to several reasons, which was explained in more detail in the following section under the heading of weaknesses of the writing rubric.

Finally, another four participants out of 24 stated that the writing rubric had *no* strengths at all.

### 4.1.2. Weaknesses of the writing rubric used currently in the proficiency examination

Almost doubling the number of the three strengths listed above, five weaknesses of the writing rubric were listed by the 24 participants of this research who took part in the proficiency examination at BUU-SFL-IEP as raters:

- The number of categories,
- The wordings of descriptors,
- The range of scores within each level,
- The number of score/band levels, and
- The weightings.

As illustrated in Table 4.2, out of 24 participants, 17 of them stated *the number of categories* as the first weakness of the writing rubric, which is followed by *the wording of descriptors*, pointed out by 8 participants. Next comes *the range of scores within each band level*, mentioned by five of the participants in the current research. Following the range of scores within each level is *the number of band levels* mentioned by four participants.

**Table 1.2.** *Weaknesses of the writing rubric used currently in the proficiency examination (N=24)*

| | *F* | *%* |
|---|---|---|
| The number of categories | 17 | 71 |
| The wording of descriptors | 8 | 33 |
| The range of scores within each band level | 5 | 21 |
| The number of band levels | 4 | 17 |
| The weightings | 2 | 8 |

Finally, another group of four participants stated the *weightings* as a shortcoming of the writing rubric utilized presently in the proficiency examination. In addition to the weaknesses listed above, 7 participants (29%) pinpointed *the lack of guidance in using the rubric* as a downside of the rubric; however, because it was considered by the researcher to be an issue related to the scoring procedure rather than the writing rubric itself, it was not included in this section.

#### 4.1.2.1. *The number of categories*

While categorization (of writing constructs that need to be assessed) was stated to be an important asset of the writing rubric by 11 participants in the current research, the number of categories (four, i.e., Content, Organization, and Grammar) was highlighted as a weakness by the majority of the participants. Specifically, out of 24 participants, 17 of them (71%) considered it as a drawback of the writing rubric because they believed there should be more categories to assess writing ability. See Table 4.3 below for the number and percentages of participants who considered an additional category (or more) necessary for a well-functioning writing rubric.

**Table 4.3.** *Additional categories deemed to be necessary for a well-functioning writing rubric (N=17)*

| | ***F*** | ***%*** |
|---|---|---|
| **Mechanics** | 7 | 41 |
| **Coherence-Cohesion** | 5 | 29 |
| **Argumentation** | 4 | 24 |
| **Length** | 4 | 24 |
| **Overall quality** | 2 | 12 |

As shown in Table 4.3, among the 17 participants finding an extra category/categories essential, seven of them stated that *Mechanics* should be one of the categories in addition to the existent categories of Content, Organization, Grammar, and Vocabulary:

> 9. In spite of the fact that the proficiency exam is a high-stakes exam, the rubric we use to evaluate the writing part of this exam is not as detailed as the one we use during the year. For example, it doesn't have the component of mechanics although we pay special attention to teach students correct pronunciation, capitalization, and spelling and assess them in our achievement tests *(P4).*
>
> 10. The current rubric lacks some crucial criteria such as mechanics and cohesion. Mechanics seems especially important to me as punctuation, capitalization, and spelling are the elements students are taught and evaluated throughout the year *(P6).*

Apart from mechanics, *coherence* and *cohesion* were writing constructs that were needed to be assessed but missing in the current writing rubric according to five participants in the study:

> 11. I really don't understand why we don't have coherence and cohesion as two important aspects of writing ability. Students should know that there must be unity in an essay. Every

paragraph must be developed and the whole essay must flow logically with the help of accurate use of cohesive devices i.e., linkers and transitions *(P4)*.

12. Unfortunately, coherence and cohesion –two components of discourse competence– are overlooked in our rubric *(P11)*.

Another category referred to by 4 participants is *argumentation*:

13. In the proficiency exam students are required to write an opinion essay. Therefore, we expect our students to write a counter-argument and refutation of it. However, in the rubric it is stated as a P.S. under organization, and there is not any information about how many points will be scored if an essay lacks it *(P21)*.

Four other participants stated that *length* should also be assessed as a separate category because test-takers are required to write between 250 and 300 words. Finally, two participants emphasized the importance of having a separate category for *overall quality* of an essay, which is considered an unclear concept, "by far the most difficult definition to articulate" in the related literature (Spencer & Fitzgerald, 1993, p. 212). According to Participant 13:

14. In my humble opinion, our rubric should have a category called "overall quality" in addition to other categories such as content, organization etc. This category should weigh at least 10% of the total score and assess the essay as a whole *(P13)*.

As can be seen from the data presented above, the majority of the participants (17 out of 24) stated that there should be more categories to assess different writing constructs such as coherence and cohesion or argumentation. However, there was one component of the writing rubric which was considered unnecessary though it was not a separate category: *irrelevance*. At present, if a student essay written in the writing section of the proficiency examination at BUU-SFL-IEP is considered to be irrelevant, it is awarded 1 point out of 20, which was found too harsh by two of the participants in the study:

15. Writing out of topic shouldn't be graded as 1 but rather should be taken out of Content. Since the topic required is out of the content, I believe being out of topic should be graded out of Content whereby other criteria should be considered.

    Content: 0, other criteria should be considered and graded *(P9)*.

Although it was stated by only two participants here, irrelevance had long been an issue of concern expressed by the raters in staff meetings at BUU-SFL-IEP. Since the writing section of the proficiency examination at BUU-SFL-IEP is a high-stakes test, 1 out of 20 was considered intolerant by the raters where content was off-topic. One of the most prominent figures in the writing assessment literature, Weigle (2002, p. 132) classifies off-topic scripts under the heading of special problems in scoring. She suggests

that in such cases it is the raters who need to decide the extent to which task achievement is essential to scoring, and this decision needs to be based on the purpose of the assessment and the type of scoring that is used. Because the writing section of the exit examination of BUU-SFL-IEP aims at determining test-takers' general writing ability through a ratable sample of writing that will demonstrate control of a variety of writing constructs such as Content, Organization, Grammar etc., the degree to which writers follow instructions exactly will be less important (Weigle, 2002, p. 133). Thus, the other participant who found the criteria for irrelevance too harsh suggested that off-topic scripts should only be penalized for content, as emphasized in excerpt (15) above by Participant 9.

In sum, it can be concluded that categorization (of writing constructs that need to be assessed) played a crucial role for the participants of the present research. As also mentioned above, the data revealed that categorization was considered as one of the strengths of the current writing rubric by the raters, whereas the *number* of categories was deemed to be the greatest weakness. The existent literature supports the participants in terms of the prominence of categorization in performance-based assessment of writing proficiency. According to Weigle (2002, p. 41), the key target of any language test is to "make inferences about language ability"; therefore, language testers need to define what is meant by language ability for a particular test. The language ability desired to be tested is defined as a *construct* by Weigle, and categories in a writing rubric represent (or are supposed to represent) the writing constructs that are deemed to be important for the realization of the expected outcomes of a language program. For the ultimate purpose of the current research, which was to develop an alternative theoretically-based and empirically-validated multi-trait rubric for the performance-based assessment of EFL writing proficiency at BUU-SFL-IEP, what was needed to be determined was the writing constructs that were considered to be necessary by the raters for the assessment of writing quality in the proficiency examination, which was explored in detail by another question (question number 3) in the open-ended questionnaire.

The subsequent section continues with the second weakness of the current rubric as perceived by the participants; that are, *the wording of descriptors* followed by *the range of scores within each band level*, and *the number of band levels*, the three of which are stated to be interrelated in the relevant literature (Brown, 2012; Knoch, 2011; Weigle, 2002).

#### 4.1.2.2. *The wording of descriptors*

Following the number of categories, the wording of descriptors was considered to be the second weakness of the current rubric, as perceived by the participants of this study. As demonstrated in Table 4.3, 33% of the participants (i.e., 8 out of 24) saw it as a flaw of the rubric. The comprehensive excerpt below by Participant 4 recaps the opinions of the eight participants who found the wording of descriptors confusing:

> 16. The wording of descriptors is so vague that it's difficult to understand what is meant. For example, in grammar section, the phrase "few errors of grammar" indicates 3 or 4 points can be given for a paper that is "good to average". However, are these few errors a total sum of articles, prepositions, and tense errors, or a few from each of these grammatical points? Also, should I give the same point to simple errors like articles and prepositions and to more critical ones such as fragments and run-ons that obscure meaning? Furthermore, what's the limit of "few errors" to give students 3 points and what's the limit of "few errors" to give them 4 (as they are in the same band)? *(P4)*

As the excerpt above clearly indicates, the raters had problems in conceiving what is meant by the quantifier "few" because it does not specify the number of errors. North (2003) labels such wording as *abstract formulation* because, as the name implies, it is not clear enough whether there is a significant difference between quantifiers, which makes it difficult for raters to come to a decision. According to Knoch (2011, p. 95), defined rubrics which have descriptions with concrete or objective formulation styles are the most useful.

#### 4.1.2.3. *Range of scores within each band level*

The range of scores within each band level was another weakness of the current rubric, as perceived by five of the participants. As mentioned above, in the original rubric the total score is calculated out of 100, but in the adapted version the total score is 20 since the writing section of the proficiency test comprises 20% of the total score. Thus, the weightings of each category differ in each rubric. In both the original rubric (ESL Composition Profile) and the adapted version, raters are expected to make a decision within a range of scores in a band level. For instance, in the original rubric, the highest band level for the category of Content ranges from 30 to 27, which is considered "Excellent to Very Good". In the adapted version, the highest band level for the category of Content ranges from 8 to 6, which is considered "Good to Average". It is the rater who

needs to make a distinction among different scores within the same band level as the following excerpts by Participant 4 and Participant 7 exemplify:

> 17. …if an essay is relevant to the given topic, we are to give 8 to 6 points in the content category. If I find an essay relevant as a rater, why do I have to consider three different points (8, 7, or 6)? How can I make a distinction? It is so hard really *(P4).*
>
> 18. …since the highest points each category has two or more options, I find it difficult to decide which point I should assign; for example, 8, 7, or 6 for an essay which has good content *(P7).*

While the original rubric by Jacobs et al. (1981) and the adapted version have a range of scores within the same band level, the rubrics used and displayed in the recent research and the seminal work of theorists in the literature of performance-based assessment of writing have one score for each band level (e.g., Banarjee et al., 2015; Becker, 2018; Brown, 2012; Janssen et al., 2015; Knoch, 2007, Shohamy et al., 1992). As indicated by five of the participants in this study, rubrics with one score for each band level are more practical in terms of the raters' decision-making processes. In addition, such a rubric design can also contribute to higher agreement between raters.

#### 4.1.2.4. *Number of band levels*

The number of band levels was another weakness of the writing rubric used currently in the performance writing section of the proficiency examination. Four participants in this study considered it a drawback of the rubric. In the original rubric by Jacobs et al. (1981) there are four band levels: Excellent to Very Good, Good to Average, Fair to Poor, and Very Poor; however, in the adapted version the number of band levels is three, which is even less: Good to Average, Fair to Poor, and Very Poor. The adaptation was made for the sake of practicality; however, three band levels for each category was found insufficient for fair assessment as the following excerpts by Participant 7 and Participant 11 illustrate:

> 19. …I can differentiate a well-written and a very low level writing paper, yet sometimes, I feel like I am underestimating a perfectly written paper because the papers above a certain level get the same points *(P7).*
>
> 20. The descriptors designed in three levels as "good to average", "fair to poor", and "very poor" are too limited to be considered as a reference indicator in a writing rubric for such an important test like proficiency *(P11).*

According to Knoch (2011, p. 92), the decision about the number of band levels should depend on the context in which a rubric is to be used. She goes on to say that the seven (plus or minus two) band level rule is applicable if a writing test is administered to a very different ability group of test takers, which is also the case for the writing performance test in this study.

The last weakness of the writing rubric as perceived by the 24 participants in this research was the weightings; that is, the relative importance of different skills and language which is assigned in the assessment process (Richards & Schmidt, 2010, p. 635).

#### 4.1.2.5. *Weightings*

Two of the participants referred to the weightings of categories as a shortcoming of the current writing rubric. However, there is not a consensus among these participants on the weightings of the four categories i.e., Content, Organization, Grammar, and Vocabulary. While one participant believed that the categories of Content and Organization should weigh more than Grammar and Vocabulary, the other participant stated that Grammar and Vocabulary are more important; therefore, they should weigh more. Another group of participants believed that each category should weigh equally. According to Participant 18:

> 21. Content should have less weighing like 4-5 (out of 20), whereas Grammar and Vocabulary should have more weighting like 5-6 as I believe they are key elements that show student's proficiency in language use (*P18*).
>
> 22. As one researcher put it, writing is like driving a car. You have to do many things at the same time to be successful. In terms of writing this means that you have to think about content and organize it by using appropriate lexis, grammar, and mechanics. One category is not more important than the other for me (*P23*).

As East and Cushing (2016) puts forward, there is an ongoing debate about the facets and sub-facets of writing considered important, for which rubric categories and descriptors are needed. There is also a discussion about how many points of differentiation are essential so that meaningful and valid information on test taker's writing abilities can be provided. The discussion in the literature about the rubric categories was also echoed in the current study, as demonstrated by the participants' excerpts above (21 & 22).

To sum up, there were five weaknesses of the writing rubric currently used for the performance-based assessment of writing proficiency at BUU-SFL-IEP, as perceived by

the twenty-four participants in the study: the number of categories, the wording of descriptors, the range of scores within each band level, the number of band levels, and the weightings.

As also mentioned a few times so forth, the most central consideration in rubric design is what the rubric categories should look like (Knoch, 2011, p. 81). The following section explores what categories were considered to be necessary by the participants in this study for reliable and valid assessment of writing proficiency.

### 4.1.3. Categorization (of writing constructs that need to be assessed)

Categorization of writing constructs that need to be assessed was an issue that was mentioned by the participants in this research in both strengths and weaknesses sections (See sections 4.1.1.2 and 4.1.2.1). According to the majority of the participants, the existence of categories was an asset, whereas the limited number of categories was a drawback of the writing rubric currently used for the performance-based assessment of writing proficiency at BUU-SFL-IEP.

Categories are deemed to be important in the literature of performance-based assessment of writing; however, the facets and sub-facets of writing that are considered to be necessary and for which categories and descriptors are required have been a debatable issue (East & Cushing, 2016; Hamps-Lyons, 2016a). According to Brown (2012, p. 20), there are so many possibilities of categories and subcategories, and the central exercise in choosing categories basically involves being aware of the possibilities and then narrowing them down from many options to a few i.e., whatever number of categories the teachers/raters will find useful in the particular institution involved.

In order to find out what categories were deemed to be important in performance-based assessment of writing but missing in the current rubric, the participants in this study were asked to select from a list of categories the writing constructs that they found important for reliable and valid assessment of writing proficiency. The list was compiled from the categories Brown (2012), Haswell (2007), and Knoch (2011) considered to be necessary for valid and reliable assessment of L2 writing performance. See Table 4.4 for the number and percentages of participants who considered each listed category necessary for a well-functioning writing rubric.

**Table 4.4.** *Categories deemed to be necessary for a well-functioning writing rubric*

| | ***F*** | ***%*** |
|---|---|---|
| **Content** | **24** | **100** |
| **Organization** | **24** | **100** |
| **Grammar** | **24** | **100** |
| **Vocabulary** | **22** | **92** |
| **Mechanics** | **22** | **92** |
| **Argumentation** | **19** | **79** |
| **Coherence** | **19** | **79** |
| **Cohesion** | **18** | **75** |
| **Length** | **11** | **46** |

As displayed in Table 4.4, all 24 participants considered the categories of Content, Organization, and Grammar crucial for a well-functioning rubric, which were followed by Vocabulary and Mechanics. Argumentation, Coherence, and Cohesion were also referred to by the majority of the participants in this research; however, Length did not seem to be as necessary as the other writing constructs listed in Table 4.4.

All the participants in this research referred to Content and Organization as the most important aspects of writing ability as exemplified in the excerpts below:

> 23. Content and Organization form the core of an essay *(P1)*.
>
> 24. From my point of view, content and organization are the most important categories in writing. If the content is not relevant, or if ideas are not logically developed and supported, this makes all the other categories meaningless *(P4)*.
>
> 25. Ideas, facts, and opinions are the building block of an essay. There is nothing to evaluate if there is no relevant building block *(P12*).
>
> 26. Content and organization go hand in hand. Content is necessary since writing is a way of communicating ideas, so it must be meaningful. It is the same with organization. Our students should be able to present their ideas in an organized way by putting their ideas in meaningful clusters *(P15*).

As the excerpts above show, the categories of Content and Organization had utmost importance for the participants in the study. Both categories have been crucial in writing assessment since the beginning of standardized testing as the following quote by David P. Harris, the project director of the TOEFL exam from 1963 to 1965, indicates:

> Although the writing process has been analyzed in many different ways, most teachers would probably agree in recognizing at least the following five general components: Content, Form, Grammar, Style, Mechanics (Harris, 1969, p. 68).

This is not the case only for TOEFL. It is, without exception, possible to see both Content and Organization in all of the scoring rubrics for six tests of ESL writing (Haswell, 2007, p. 111) as *main traits*, while other categories such as Vocabulary and Mechanics are used varyingly, which applies to scoring rubrics in more recent research, too (e.g., Banarjee et al., 2015: Becker, 2018; Brown, 2012; Janssen et al., 2015; Knoch, 2011).

The next category in the list provided in Table 4.4 above was Grammar, and it was as important as Content and Organization according to the 24 participants in the study, as reflected in the excerpts below:

> 27. Grammar is important in terms of conveying ideas clearly and effectively. It is an indicator of language proficiency *(P12).*
>
> 28. Mastery of a language in terms of accuracy is essential for language development *(P15).*

The category of Grammar does exist in the scoring rubrics in the relevant literature mentioned above; however, unlike the categories of Content and Organization, Grammar does not always exist as a main trait. In some cases, it is a subcategory under the main categories of either Accuracy or Language Use, where it is evaluated with other aspects of writing such as Vocabulary or Mechanics. More recent scoring rubrics have the category of Grammar as a main trait mostly (Banarjee et al., 2015; Brown, 2012; Janssen et al., 2015) except for Knoch (2011), where it is assessed under the category of Accuracy and Becker (2018), where it is assessed under the category of Language Use. See Table 2.1 for the main traits of scoring rubrics for six tests of ESL writing (Haswell, 2007).

Following the categories of Content, Organization, and Grammar, which were considered as indispensable for a writing rubric by all the participants in this study, Vocabulary and Mechanics were selected as the two other important aspects of writing by almost all of the participants (i.e., 22 out of 24). The excerpts below exemplify the perceptions of the participants regarding the category of Vocabulary:

> 29. When students do not use correct words/phrases, meaning can easily be obscured. I think accurate use of words is vital *(P4).*

30. Effective word/idiom choice and usage, word mastery, and appropriate register are what we expect to see in good writing. Like grammar, word mastery is an indicator of language proficiency *(P12).*

31. It is crucial for the reason that vocabulary knowledge determines the quality of writing *(P15)*.

Apart from Vocabulary, Mechanics was another category which was considered necessary by nearly all of the participants, specifically 92% of them. The categories of Content, Organization, Grammar, and Vocabulary exist in the current rubric that is used as the writing rubric for the performance-based assessment of EFL writing proficiency at BUU-SFL-IEP. The category of Mechanics, on the other hand, is not included in the current rubric despite its existence in the ESL Profile by Jacobs et al. (1981) from which the current rubric is adapted. More importantly, it exists in the goals and objectives of the writing course at BUU-SFL-IEP but not assessed in the proficiency examination although it is assessed in the achievement tests throughout the year:

*Goal:* Students will be able to use basic writing skills accurately and effectively by focusing on the elements of a good sentence within the context of a paragraph.

*Objective:* Students will be able to write simple sentences with accurate capitalization, punctuation, and word order.

As also mentioned above in the section titled the weaknesses of the current rubric (See 4.1.2.1), the category of Mechanics where spelling, punctuation, and capitalization are assessed was not included in the adapted rubric, and the lack of this category was deemed to be one of the greatest weaknesses of the current rubric by the majority of the participants in this study (See section 4.1.2.1).

32. Mechanics are very important in writing. Otherwise, it would be really hard for the raters to figure out the messages students are trying to convey because incorrect pronunciation, for instance, can make sentences incomprehensible *(P4)*.

33. We spend a great deal of time teaching mechanics because they are important. The rule is simple: Test what you teach *(P7)*.

35. We cannot ignore mechanics while evaluating writing performance. If we are going to ignore these rules, why did we teach them in the first place?! It's inconceivable *(P12).*

36. Mechanics should definitely be a part of the rubric as without relevant punctuation it is almost impossible to give the right message for writers *(P20)*.

As with the categories of Grammar and Vocabulary, the category of Mechanics exists as a main trait in half of the scoring rubrics used for six tests of ESL writing listed by Haswell (2007). See Table 2.1.

Argumentation was the next category considered to be essential for 79% of the participants in this phase of the current research (specifically 19 out of 24 participants) following the categories of Content, Organization, Grammar, Vocabulary, and Mechanics, which existed in both the original rubric by Jacobs et al. (1981) and the adapted version of it except for Mechanics.

What is meant by argumentations is discussing the topic from multiple perspectives by covering counterargument(s) and refutation of it. The reason why argumentation was deemed to be necessary by the participants was that students are required to write an argumentative essay for the assessment of performance writing component of the proficiency exam at BUU-SFL-IEP. When the goals and objectives of the writing course at BUU-SFL-IEP are considered, it is possible to see argumentation in an explicit statement:

*Goal:* Students will be able to write a fully-developed (five-paragraph) argumentative essay.

*Objective:* Students will be able to support their opinion on a controversial issue, and present and refute a counter argument in a full paragraph.

Most of the participants referred to the instructional objectives of the writing course as the reason why argumentation should be a part of a writing rubric:

37. We require our students to write an argumentative essay in the proficiency exam, and throughout an important part of the second semester, we teach them to write a paragraph (the third body paragraph) where they mention about the counter argument by downgrading it and next refute it. Thus, we should definitely assess this properly *(P4).*

38. … As far as I know, argumentation is more important than language use (grammar and vocabulary I mean) in international standardized tests like TOEFL and IELTS. We also spend a great deal of time and energy to teach it during the year. So I think it should be emphasized in the rubric *(P14)*.

79% of the participants who thought argumentation was essential stated that it should be a separate category, while 21% (4 participants out of 19) supported it needed to be assessed within the category of organization:

39. Argumentation is a part of organization. We teach organization because it is not possible to make sense out of a disorganized piece of writing. Argumentation is an essential part in a rubric in this respect *(P12)*.

The latest CEFR Companion Volume with new descriptors (2018: 142) pinpoints the importance of counterargument starting from the level of B1: *Can introduce a counterargument in a discursive text*. In an article devoted to counterargumentation in argumentative writing in a high-stakes test, Lui and Stapleton (2014) also highlight that counterargumentation is a key factor that contributes to the persuasiveness of argumentative essays and put forward that counterargumentation be considered in the writing prompts and scoring rubrics of high-stakes English tests in addition to classroom instruction on argumentative writing, as also emphasized by the participants in this study.

Following the category of Argumentation, the categories of Cohesion and Coherence to assess discourse competence were deemed to be important for a writing rubric by the participants in this study. Out of 24 participants 19 of them considered Coherence necessary, i.e., 79%, and 18 found Cohesion essential, i.e., 78% of all participants. As with the categories of Mechanics and Argumentation, the categories of Cohesion and Coherence are not assessed in the proficiency examination although they exist in the goals and objectives of the writing course at BUU-SFL-IEP:

*Goal:* Students will be able to write a fully-developed (five-paragraph) argumentative essay.

*Objective:* Students will be able to use appropriate transitions to form a cohesive argumentative essay.

*Objective:* Students will be able to generate a general topic for an argumentative essay and brainstorm ideas by asking questions about the topic. They will be able to narrow down their topics and organize their ideas to form a unified and a coherent argumentative essay.

The excerpts below indicate the importance which participants attached Cohesion and Coherence:

40. Students should know that there must be unity in writing. At the very beginning, we try to teach them that in writing it is necessary to use a concept map (or outline). There must be a thesis statement, an introduction, appropriate body paragraphs, and a suitable conclusion. Every paragraph should be well-developed, and the whole essay should follow a logical order with the help of accurate use of linkers and transitions, which makes cohesion another important aspect of writing. Cohesion also affects the tone of writing. Cohesive devices not

only strengthen the link between the sentences but also affect the whole essay in terms of unity. If these two elements are missing in an essay, reading the text becomes too difficult, and the rater might have to guess what the student is trying to say, or s/he has to read certain sections of the essay again and again *(P4)*.

41. Using accurate transitions is a convention in writing, which we emphasize throughout the year. In the same way, without coherence, we cannot talk about the logical flow of ideas *(P12)*.

42. Cohesion is especially difficult for Turkish students. They definitely need to practice more to use cohesive devices accurately and appropriately because my general observation is that they either underuse or overuse them. Scoring the use of cohesive devices will motivate students to use them more carefully (*P14*).

In line with the relevant literature which highlights the importance of cohesion and coherence in writing performace, the participants in the study indicated that cohesion and coherence are important constructs in the performance-based assessment of writing proficiency.

The final category mentioned by 46% (n=11) of the participants is *length*. Unlike the other categories supported by the majority of the participants, the category of length was not considered to be vital as long as students could develop an argument logically and support their views with relevant information, as the following excerpts illustrate:

43. In my opinion, length is certainly important in writing an essay because it's not possible to develop and support your ideas with two or three sentences in a body paragraph. In their writings, we don't actually expect our students to write anything irrelevant just in order to reach the specified word limit; rather, we expect them to develop their ideas in a logical order with sufficient content *(P4)*.

44. I don't think that length should be a separate category as it is a subcategory of content (*P20*).

45. In our institution we reduce points from the content category if the word limit is not met. In other institutions, for example, Bilkent, my previous workplace, students are awarded no points at all if the word limit is not met. My personal opinion is that it should be scored within the category of content (P23).

In support of the perceptions of the participants in the present research, length does not seem have a vital significance in scoring rubrics for six Tests of ESL writing (Haswell, 2007, p. 111) as only one scoring rubric, i.e., Test of Written English (Educational Testing Service), has length as a separate category.

In conclusion, the participants of this study considered five categories indispensable – Content, Organization, Grammar, Vocabulary, and Mechanics – followed by

Argumentation, Coherence, and Cohesion, which are in line with the goals and objectives of the writing course at BUU-SFL-IEP and the pertinent literature.

Following the categories of writing constructs to be assessed, there is a number of other important decisions a rubric developer has to make at the descriptor level (Knoch, 2011), which is the topic of the subsequent section.

### 4.1.4. Number of band levels

Knoch (2011, p. 92) emphasizes that once the categories of writing constructs (to be assessed) are determined, the decisions that need to be made at the descriptor level are as follows:

- How many band levels should the rubric have?
- How will the descriptors differentiate between the band levels?
- How will the descriptors be formulated?

As also mentioned above in section 4.1.2.4, these three vitally important decisions are stated to be interrelated in the relevant literature (Brown, 2012; Knoch, 2011; Weigle, 2002).

Unlike the original version of the ESL Composition Profile (Jacobs et al. 1981) which has four band levels (Excellent to Very Good, Good to Average, Fair to Poor and Very Poor), there are three band levels (Good to Average, Fair to Poor and Very Poor) in each category of the writing rubric used currently for the performance-based assessment of writing proficiency at BUU-SFL-IEP. According to participants, three band levels were not sufficient. The lack of adequate number of band levels was also mentioned by the participants in this study under the heading of the weaknesses of the current rubric (See 4.1.2.4 above). Specifically, out of 24 participants, 16 of them (67%) considered three band levels to be insufficient. 8 of these 16 participants supported that there should be *four* band levels while 4 of them believed *five* band levels were necessary for fair scoring.

> 46. Well, three levels could be more than enough for classroom assessment, but in a high-stakes test such as proficiency, it is simply not enough. For a practical, rater-friendly rubric, we need to have a writing rubric with four band levels, like Excellent (4), Good (3), Fair (2), and Poor (1). Even (0) may be an option for Very Poor essays *(P5)*.
>
> 47. I think four levels could be much better (Very Good, Good, Fair, Poor). We score the essay students write in the proficiency exam out of 20. 1 point does make a lot of difference

under these circumstances. There should definitely be a difference between Very Good and Good (*P22*).

According to Knoch (2011, p. 92), "the decision about the number of band levels should depend on the context in which a rubric is to be used". She goes on to say that the seven (plus or minus two) band level rule is applicable if a writing test is administered to a very different ability group of test takers, which is also the case for the writing performance test in this study. Considering the data gathered from the participants in this research and the relevant literature, it seemed that for fair assessment of writing proficiency in the specific context of BUU-SFL-IEP there should be four band levels at the least (e.g., *Excellent, Good, Passing, Fail*) with a concrete and objective formulation style in which descriptors in each band can be transformed into a checklist of "yes" or "no" questions (e.g., *Organization*: appropriate title, effective introductory paragraph, topic is stated, leads to body, transitional expressions used; supporting evidence given for generalizations; conclusion logical and complete).

### 4.1.5. Wording of descriptors

Together with *the number of band levels* another important issue that needs to be taken into consideration is *the wording of descriptors*. In order to find out the perceptions of the participants in this study on the issue, they were asked whether there were any categories in which the wording of the descriptors needed to be changed. Out of 24 participants taking part in the proficiency examination at BUU-SFL-IEP as raters, 13 of them asserted that the wording of descriptors in all categories must be modified.

48. I don't think that these descriptors help us do fair assessment. They are vague statements that will lead to subjective conclusions. There should be much more detailed descriptors in order to achieve inter-rater reliability. The descriptors shouldn't be open to interpretation. Any given rater should understand the same thing from the descriptors *(P12).*

49. …The descriptors should be more clear-cut. I find them unclear. Now it is not possible to rate without subjectivity I'm afraid (*P20*).

Below is an excerpt for each category taken from the writing rubric currently used for the performance-based assessment of writing proficiency at BUU-SFL-IEP and the results for each category with regard to the wording of descriptors in each of them.

#### 4.1.5.1. *Content*

The categories of Content, Organization, and Grammar were perceived by the participants in this study as the most important writing constructs (See section 4.1.3). Content was also the category with the highest weighting in the rubric (8 out of 20). See Table 4.5 below for the band levels and descriptors for each band level. As mentioned before, the majority of the participants referred to the difficulty of assigning one of the three scores within the same band level before moving onto the wording of the descriptors.

**Table 4.5.** *The descriptors for the category of content in the adapted version of ESL Composition Profile*

| **Content:** |
|---|
| 8-6: Good to average: relevant to the given topic, knowledgeable |
| 5-3: Fair to poor: mostly relevant to the given topic, some knowledge of the topic |
| 2-1: Very poor: partially relevant to the given topic, limited or no knowledge of the given topic |

Regarding the wording of the descriptors, the first concern raised by the participants was the confusion caused by the use of the adjectives *average* in the highest band level and *fair* in the band level that follows it. According to Participant 5 and Participant 23:

> 50. I can't see any difference between average and fair. So I don't understand the reason why they are used in different levels. We can have Very good, Good, Fair, Poor (*P5*).
>
> 51. The words average and fair mean the same thing, don't they? Excellent (4), Good (3), Average (2), Poor (1) would be more straight-forward and easier to score (*P23*).

Another wording found baffling was the use of *partially relevant* and *no knowledge* in the descriptor of the same band level i.e., the lowest band level, as highlighted in the excerpts below.

> 52. There are some contradictory expressions in this category; for example, the last level 2-1 Very poor. If students have limited knowledge, they can write a partially relevant essay (that's OK); however, if they have no knowledge of the topic, how can they write something that is partially relevant? I believe there must be another band level here with its own descriptor. We can say that if students don't have any knowledge of the topic and therefore

write something irrelevant, we should give them 1 point, and when it's partially irrelevant, then they should get 2 points (*P4*).

53. If a student has no knowledge of the given topic, s/he should get 1 or even 0 for his or her writing because no knowledge of the given topic means that he has produced an irrelevant piece of writing (*P23*).

As the excerpts above clearly indicate, using synonymous adjectives in different band levels and contradictory expressions within the same band level created serious problems on the side of the raters.

#### 4.1.5.2. *Organization*

The next category in the rubric was Organization. See Table 4.6 below for the band levels and descriptors for each band level.

**Table 4.6** *The descriptors for the category of organization in the adapted version of ESL Composition Profile*

| **Organization:** |
|---|
| 4-3: Good to average: well-organized, in accordance with the given style |
| 2 : Fair: loosely organized, but still in accordance with the given style |
| 1 : Poor: loosely organized and different from the given style |

The first issue participants highlighted for this category was the lack of the requirements of argumentative writing in the descriptors, as 9 out of 13 participants who believed the wording of descriptors must be changed indicated.

54. The descriptors in this category should refer to the requisites of argumentative essay because throughout the spring semester we focus on the conventions of argumentative writing. What if the thesis statement or one of the predictors stated in the thesis statement are missing? What if there is not a counterargument, or there is a counterargument without refutation? (*P11*).

Another issue participants refer to was the vagueness of the expressions *well-organized, loosely organized, in accordance with the given style,* and *different from the given style*.

55. Being "loosely organized" and being "different from the given style" are two things, aren't they? I'm so confused about it *(P7*).

56. I don't understand what is meant by "in accordance" or "still in accordance with the given style". There might be a checklist to decide on to what extent the student has mastered

applying the giving style, such as a well written thesis statement, an effective introduction, giving a counterargument and refuting it etc. (*P12*).

57. The expression "loosely organized" shouldn't be used in both bands (*P23*).

The final concern participants raised was the necessity to include the elements of cohesion and coherence in this category as the following excerpts reflect.

58. Organization was not only about following the conventions of the given style. Fluent expression of ideas, logical sequencing, cohesion, and being able to offer enough support are all important *(P12)*.

59. The organization category is so vague… It should be reworded or better rewritten so as to reflect the fulfillment of cohesion and coherence elements *(P23)*.

In sum, the lack of the conventions of argumentative writing, the lack of the elements of cohesion and coherence, and the vague wordings of the descriptors were the issues participants referred to for the category of organization.

#### 4.1.5.3. *Vocabulary*

Vocabulary was the third category in the writing rubric following Content and Organization. See Table 4.7 below for the band levels and descriptors for each band level.

**Table 4.7.** *The descriptors for the category of vocabulary in the adapted version of ESL Composition Profile*

| **Vocabulary:** |
|---|
| 4-3: Good to average: appropriate use of words |
| 2 : Fair: limited use of words |
| 1 : Poor: no word mastery at all, or not enough to evaluate |

As with the other categories analyzed so far, the vagueness of expressions was an issue for this category, as well.

60. Here what do we mean by "appropriate use of words"? Do we mean appropriate use of words, or do we mean trying to use synonyms or antonyms or different parts of speech? What does "limited use of words" mean? How about false translations like "sharp vinegar" or "take under" or wrong collocations like "make business with someone"? What about mistakes that obscure meaning? Should incorrectly used lower frequency words that do not obscure meaning be given a lower score than correctly used simple, higher frequency words? (*P4*).

61. Not clear what is meant by "not enough to evaluate". The phrase "no mastery at all" and "not enough to evaluate" shouldn't be in the same band. While the adjective "appropriate" focuses on quality, the phrases "limited use" and "no word mastery" focus on quantity. Shouldn't the descriptors be parallel in this respect? (*P23*).

The second concern of the participants was the lack of "the use of a *variety* of words and expressions" in the descriptors as Participant 4 also referred to above. 10 out of 13 participants who believed the wording of descriptors must be changed referred to variety for this category. See the excerpts below by Participant 12, Participant 19, and Participant 20 raising the same concern.

62. The descriptors are not detailed enough. We cannot evaluate a student's proficiency in English vocabulary by looking at "appropriate use of words" only. What do you mean by "appropriate use of words" in the first place? It's a hazy concept. Is "limited" opposite of "appropriate"? We could make a better selection of words or phrases in the descriptors. Word range, effective word choice, word form mastery, and appropriate register all come into play here *(P12)*.

63. When a student uses the basic or simple words over and over again, but use them appropriately, I don't know how to score. I expect to see something like "varied words used appropriately" *(P19*).

64. Well, I'd like to see more detailed descriptors, such as 'creative and accurate use of a variety of words with correct collocations *(P20)*.

The literature on L2 writing quality also supports this idea of the use of a variety of words or expressions. According to McNamara, Crossley, and McCarthy (2010), in order to become better writers, students may need to become familiar with and have a better command of "a greater diversity of words, less frequent words, and more complex syntactical structures" (p. 75).

As for the category of Vocabulary, vague descriptors and the lack of reference to diversity of words were the issues participants in this study considered important.

#### 4.1.5.4. *Grammar*

The last category in the rubric was Grammar. See Table 4.8 below for the band levels and descriptors for each band level.

**Table 4.8.** *The descriptors for the category of grammar in the adapted version of ESL Composition Profile*

| **Grammar:** |
|---|
| 4-3: Good to average: few errors of grammar |
| 2 : Fair: some errors of grammar |
| 1 : Poor: no mastery of grammar at all, or not enough to evaluate |

Once again, the ambiguity of the descriptors was the number one problem with this category for the participants in this study. Another noteworthy issue raised by the participants was the emphasis put on "errors" in the descriptors rather than the complexity of syntactical structures, as the following excerpts exemplify.

> 65. "Few" and "some" = exactly how many? More importantly, should we count the mistakes? Is it better to write simple and grammatically correct sentences or making a mistake in an effort to use a more complex grammatical structure? For instance, "I have a sister. My sister is 22. She is a student at university.", or "My elder sister which is a student at university, is 22 years old." Should I score the second example in the "fair" band just because there is a mistake in the relative pronoun? (*P4*).
>
> 66. With the descriptors in the grammar category our main focus is on errors. We need to be able to differentiate sentence structures at A1 or A2 level from the more complex ones, don't we? We need expressions like "a good range of patterns with full accuracy" or "a limited range of patterns with frequent errors" (*P5*).
>
> 67. I don't like to call this part grammar. "Language use" might be a better term to use. When we are assessing grammar, are we going to focus on the number of studens' errors only? Is it fair? How are we going to differentiate a piece of writing with effective complex constructions from the one with accurate but simple ones? *(P12).*
>
> 68. I imagine descriptors like these☺:
>
> Poor: No trace of syntactic unity; inaccuracy leading to misunderstanding
>
> Weak: Use of simple structures with a lot of mistakes
>
> Mediocre: Use of simple sentences and a few compound and complex sentences with a few mistakes hindering meaning
>
> Excellent: Use of a variety of syntactic expressions in a way that enrich meaning and reflect creativity (*P20*).

As the excerpts above clearly indicated, the use of ambiguous expressions like quantifiers and the lack of reference to syntactic complexity in this category were the issues that need to be taken into consideration in the process of rubric development.

In conclusion, the general concern of the participants regarding the wording of descriptors was the use of vague phrases or expressions, synonymous adjectives in

different band levels, and contradictory expressions within the same band level. For the category of Organization, the lack of the requirements of argumentative writing and the lack of the elements of cohesion and coherence in the descriptors were the other issues the majority of participants referred to. Similar concerns were raised about the categories of Vocabulary and Grammar, as well. While the lack of "the use of a diversity of words and expressions" in the descriptors was the common problem for the category of vocabulary, the emphasis put on "errors" in the descriptors rather than the complexity of syntactical structures was the main issue for the category of Grammar, which would definitely be taken into consideration during the rubric design process.

The topic of the following section is another crucially important decision in the rubric design: weighting, i.e., the relative importance of different skills and language which is assigned in the assessment process (Richards & Schmidt, 2010, p. 635).

### 4.1.6. Weightings

There is an ongoing debate on how many points of differentiation are essential so that meaningful and valid information on test takers' writing abilities can be provided (East & Cushing, 2016). According to prominent figures in the field of writing assessment, the key factors that need to be taken into consideration are the *context* and the *purpose* in developing and applying rubrics in both classroom and large-scale writing assessment (e.g., Broad, 2003; Brown, 2012; Crusan, 2010, 2015; East & Cushing, 2016; Hamps-Lyons, 2016a, 2016b).

The 20-point is distributed as follows in the writing rubric currently used for the performance-based assessment of writing proficiency at BUU-SFL-IEP: Content: 8, Organization: 4, Vocabulary: 4, and Grammar: 4. When the participants in the study were asked about these uneven weightings of categories, the majority of them (14 out of 24) indicated that content was overrated despite its importance in writing.

> 69. In my opinion, content is not more important than organization, so its weighting shouldn't be that high. These categories can be given equal points because without a clear organizational pattern, readers - I mean raters in our case - can become confused and may lose interest. Also, a well-structured essay helps the rater to draw connections between the thesis and the body. I believe that logical sequencing of ideas is as important as generating ideas because it does not make any sense unless ideas are not supported or developed logically (*P4*).

70. The weightings of content and organization and should be more or less the same because these two categories form the backbone of a writing (*P13*).

71. Content should have less weighing like 4-5 (out of 20), whereas grammar and vocabulary should have more weighting like 5-6 as I believe they are key elements that show student's proficiency in language use (*P18*).

72. As one researcher put it, writing is like driving a car. You have to do many things at the same time to be successful. In terms of writing this means that you have to think about content and organize it by using appropriate lexis, grammar, and mechanics. One category is not more important than the other for me (*P23*).

The debate on weighting in the literature of performance-based assessment of L2 writing was also reflected in the context of BUU-SFL-IEP, as the excerpts above show. It seemed that equal weighting of the categories might be the solution because of the significance of all writing constructs for L2 writing quality.

### 4.1.7. Categories considered to be difficult to score

When the participants were asked the name of the category (ies) they found difficult to apply, the majority replied as all. Out of 24 participants, 15 of them answered the question this way.

73. I think all categories are problematic in terms of wording of the descriptors and presenting us with different score options to choose from for the same band level (Content: 8-7-6: Good to average). I believe each band level should have one score for the sake of inter-rater reliability. With our present rubric I even feel suspicious of my intra-rater reliability (*P6*).

74. I think all the categories are difficult to apply because of ineffective and vague wording. It is not "guiding" at all. A novice teacher might have a lot of trouble applying this rubric. My experience helped me to 'cope with' this rubric (*P14*).

75. Most of them indeed because the descriptors lead me to a more holistic way of rating. In the last proficiency exam, there were a lot of discrepancies between raters. Thus, in my opinion, it is not my problem only (*P20*).

The issues mentioned under the heading of the weaknesses of the writing rubric in section 3.2.6.2 were raised once again in this section: the wordings of descriptors and the range of scores within each level.

In the last part of the open-ended questionnaire participants were asked their general satisfaction level in using the rubric.

### 4.1.8. Participants' general satisfaction level

Finally, 24 participants of the study who took part in the proficiency examination at BUU-SFL-IEP as raters were asked to rate the current writing rubric from 1 to 5 (ranging from very satisfied to very dissatisfied) on:

- the extent to which it facilitates fair assessment of students' written work and
- the participants' confidence level in applying the writing rubric.

#### 4.1.8.1. *Participants' perceptions on fair assessment of students' written work*

The first item regarding the general satisfaction of participants in using the rubric examined the extent to which it facilitates fair assessment of students' written work. Participants were asked to rate the current rubric from 1 to 5:

5. Very satisfied
4. Satisfied
3. Neither satisfied nor dissatisfied
2. Dissatisfied
1. Very dissatisfied

See Table 4.9 below for the number and percentages of participants' ratings.

**Table 4.9.** *Participants' perceptions on fair assessment of students' writing (N=24)*

| | ***F*** | **%** |
|---|---|---|
| ***5*** | **1** | 4 |
| ***4*** | **7** | 28 |
| ***3*** | **9** | 38 |
| ***2*** | **5** | 23 |
| ***1*** | **2** | 8 |

As can be seen in the table 4.9, only one third of the 24 participants were satisfied with the current writing rubric in terms of its facilitation of fair assessment of students' written work while one third of them were dissatisfied with it. The majority of the participants were neither satisfied nor dissatisfied with the rubric, which was not an unexpected result, considering the abundance of the negative criticisms expressed in the other sections of the open-ended questionnaire.

#### 4.1.8.2. *Participants' confidence level in applying the writing rubric*

The second item in this section investigated participants' confidence level in applying the current writing rubric. See Table 4.10 below for the number and percentages of participants' ratings.

**Table 4.10**. *Participants' confidence level in applying the rubric (N=24)*

| | ***F*** | **%** |
|---|---|---|
| ***5*** | **1** | 4 |
| ***4*** | **8** | 32 |
| ***3*** | **10** | 40 |
| ***2*** | **3** | 17 |
| ***1*** | **2** | 8 |

In accordance with the results of the first item, out of 24 participants only 9 of them felt confident in applying the writing rubric. 15 of them were either moderately confident or not confident at all with the rubric despite using it for eight years.

### 4.1.9. Conclusion of Phase 1

The results of the first phase which aimed at exploring the participants' perceptions of the writing rubric currently in use indicated that weaknesses outweighed strengths by almost doubling them. These results were compliant with the satisfaction and confidence levels of the participants the majority of whom were dissatisfied with the rubric (16 out of 24) and did not feel confident in using it (15 out of 24) although they had been using the rubric for a long time.

Based on the results of this first phase of the study, an alternative theoretically-based and empirically-validated multi-trait rubric was designed in Phase 2 of this research for the specific context of BUU-SFL-IEP for the performance-based assessment of EFL writing proficiency. The data gathered during in Phase 1 was recapitulated by following the four steps provided by Weigle (2000, pp. 122-125), which aimed at determining the type of the rubric that was desired, the aspects of writing that were most important and how they would be divided, the number of band levels, and the wording of descriptors. As Knoch (2011, p. 82) supports, "for a rubric to be valid, each of these design options has to be weighed carefully".

Regarding the type of rubric that was desired, out of 24 participants who volunteered to be a part of this phase of the study, 11 of them considered categorization (of writing constructs that need to be assessed) as a strength of the writing rubric because it assisted them in identifying the aspects of writing that were required to be assessed and guided them when giving feedback about strong and weak aspects in students' writing, as also emphasized in the pertinent literature. Therefore, it could be concluded that an assessor-oriented *analytic/multi-trait rubric* was desired for the specific context of this study.

After the type of the rubric was decided on, another vitally important decision that a rubric designer should make was to finalize the rubric categories and divide them up. Based on the relevant literature and the data gathered through the open-ended questionnaire used in this phase of the present research, there would be five categories in the alternative rubric: *Content*, *Organization*, *Grammar*, *Vocabulary*, and *Mechanics*. As emphasized by the participants, the conventions of argumentative writing and the elements of cohesion and coherence would be referred to in the category of organization. Diversity of words and complexity of syntactic structures would be a part of the categories of vocabulary and grammar respectively. The category of mechanics which was missing in the current rubric would take place in the alternative one. When it came to weighting, it seemed that equal weighting of the categories might be the solution because of the difficulty of coming to an agreement on which aspects of writing should weigh more and the significance of all writing constructs for L2 writing quality.

Following the type of the rubric that was desired and the categories that were most important came the number of band levels and the wording of descriptors. Based on the data gathered from the participants in this study, it seemed that *five band levels* with concrete and objective style would be useful for fair assessment of students' writing proficiency. Thus, descriptors in the form of a checklist was planned to be formulated for the purposes of the current research.

In conclusion, an assessor-oriented analytic/multi-trait rubric that had five categories with five-band levels and concrete descriptors was decided to be designed for the specific context of BUU-SFL-IEP for the performance-based assessment of EFL writing proficiency based on the results of Phase 1. Apart from the results of Phase 1, the related literature and expert opinion were also consulted during the process of

development of draft rubric, which is the topic of the following section, Phase 2: Development of draft rubric.

## 4.2. Phase 2: Development of draft rubric

The preliminary stage of the rubric design was to design a draft rubric based on the expectations of the participants from a writing rubric, as explained in Phase 1, with the guidance of experts in performance-based assessment of writing proficiency and the relevant literature (Banarjee et al., 2015; Becker, 2011, 2018; Brown, 2012; Chiang, 1999, 2003; Crusan, 2010; East, 2009; Hamp-Lyons, 1991, 1995, 2016a, 2016b, 2018; Hattingh, 2009; Janssen, Meier, &Trace, 2015; Knoch, 2007, 2011; Weigle, 2002; Weir, 2005; Weir & Shaw, 2007).

In this section, the process of designing a draft rubric and results of this second phase of the research are presented. Participants in this phase of the study were 3 experts from different institutions and 5 experienced teacher-raters who worked as instructors at BUU-SFL-IEP (See 3.3.2 for detailed information on the participants of this phase of the study). After the five experienced teacher-raters were consulted for their opinion on the draft rubric and their approval was received, the draft rubric was shared with the three experts so as to learn their opinions on the draft rubric.

Below are the suggestions and recommendations of each expert on various aspects of the draft rubric.

### 4.2.1. Expert opinion

The three experts who accepted to participate voluntarily in the study were requested to state their opinions on the second draft of the new rubric, specifically on the content and the number of categories, the number of band levels, the wordings of descriptors, and the weighting.

#### 4.2.1.1. *First expert*

A non-native English speaking professor specialized in performance-based assessment of EFL proficiency and employed at an ELT department of an English-medium Turkish state university was the first expert whose opinion was consulted. The professor stated that the content and the number of categories were adequate to assess writing proficiency. He went on to state that the number of band levels, the range of scores

within each level, and the weighting were institutional decisions that should be made depending on the specific assessment context, which was the rationale behind Phase 1 of the study. This opinion of his is in accordance with the extant literature on the performance-based assessment of writing proficiency which, from a socio-cognitive perspective to assessment, puts the emphasis on the context and the purpose in developing and applying rubrics in both classroom and large-scale writing assessment (e.g., Broad, 2003; Brown, 2012; Crusan, 2010, 2015; East & Cushing, 2016; Hamps-Lyons, 2016a, 2016b). He recommended some modifications on the wording of descriptors so that they could be easy to interpret, simple to use, and accurate. Also, he highlighted some expressions in the categories of Content and Organization to avoid overlapping descriptors. See Appendix 10 for his modifications in the wording of descriptors which were solely on word level.

#### 4.2.1.2. *Second expert*

The second expert whose opinion was consulted was a native English speaking professor also specialized in performance-based assessment of EFL proficiency and educational assessment and evaluation and employed at a TEFL department of an English-medium Japanese university. Regarding the content and the number of categories, he stated that five content areas given as a construct of writing were fairly standard and reflected overall objectives and focuses of instruction in the language program. Therefore, scorers were expected to easily assess student work. In congruence with the first expert, the second expert also referred to some overlap across different elements of the construct that could be clarified. Specifically, he indicated that "conciseness" seems to be treated twice, as a property of Content and Vocabulary. He thought that it would be better to reserve this concept to Content and added that making it part of the Vocabulary component might be problematic because that component also included the concept of variety, which might be in conflict with conciseness at times. Additionally, within the Vocabulary component, "conciseness" might also already overlap with "preciseness", making it an unnecessary addition to the component. He went on to state that some aspects of what was being considered under Organization might be better placed under Content. Specifically, the presence of "Support for Arguments" and the "Counterargument" might be better considered aspects of Content rather than Organization since he thought that in the current rubric, Organization covered many

aspects of the text (title, paragraph structure, thesis statement, transitions, support for arguments, counterargument, and conclusion) while Content covered relatively few (topicality, development of ideas, conciseness, and thoughtfulness/effort). He recommended that "Support for arguments" should be distinguished from "Development of ideas", which argued for this being treated in only one place. He also stated that the counterargument seemed to be more of a content issue than an organizational one. If the counterargument does not have appropriate content, its presence is organizationally inappropriate by default.

In relation to the number of band levels, he stated that 5 bands might allow neat alignment to traditional 5-level grading (A, B, C, D, F). However, he expressed his discontent with the labels of bands (Very Good; Good; Moderate; Poor; Very Poor) for some reasons. Firstly, he referred to a possible problem in distinguishing Poor from Very Poor across the scale. He also highlighted that these labels might not be as helpful to raters and students as they could be. Hence, he suggested some alternatives (along with some reworking of descriptors) that might help raters in their assessments and provide useful feedback to students to support their use of the rubric to learn. The alternative labels that he put forward were Exceeds Expectations; Meets Expectations; Approaches Expectations; Needs Development; Off Topic/Did Not Try because he believed these labels acknowledged that there were standards (expectations) for performance that were not perfection; thus, students could meet or exceed them, or be at different points on the way to meeting them. In other words, the labelling and the bands became developmental, which seemed difficult to disagree because doing this would also clarify the distinction between the 1 and 0 bands in each component. 0 band was for students whose essays were completely off topic or who showed no effort; in other words, 0 was reserved for failures to address the task. 1 was for students who tried but whose abilities were not at a passing level. This might also support a clearer distinction between the 2 and 1 bands. The former was for work that was acceptable but not at the expected level, while the latter was for work that was not yet acceptable, despite whatever efforts the student made.

Finally, the professor expressed his opinions regarding the weighting of each category. As also mentioned in the results and discussion section of Phase 1 (Section 4.1.2.5), he referred to the ongoing debate on how many points of differentiation were needed in order to provide students with meaningful and valid information on their writing abilities (East & Cushing, 2016). In line with the related literature, the results of

Phase 1 also indicated that the 24 teacher-raters had a varying opinions on the weighting of the categories. Still, the professor stated his views regarding the weighting as follows. He said that he would give greater weight to Organization because the component as defined covered so many more aspects of the text. He also added that while it was almost traditional to weigh Grammar and Vocabulary equally, he would, if possible, give more weight to Vocabulary because he believed that if students mastered the aspects of vocabulary use covered in that component, their grammatical ability should improve along with it. Also, he added that putting more weight on Vocabulary might lead students to put more effort into it, which he believed would benefit the development of their language ability more than similar time spent focused on Grammar (See Appendix 11 for his modifications).

#### 4.2.1.3. *Third expert*

The third expert whose opinion was consulted was a native English speaking freelance teacher trainer who was very familiar with the Turkish EFL teaching context, especially with EFL instruction at tertiary level and specialized in performance-based assessment of writing in addition to continuous professional development of EFL instructors. Regarding the content and the number of categories, he agreed with the two other experts on the importance of the alignment of the learning outcomes (of the writing course) with the categories in the rubric. He stated that a quick analysis showed that the categories in the rubric touched many of the bases in what would be described as "convention wisdom" for this type of task and an analytic/multi-trait rubric. Still, he added that Voice could also be a part of the new rubric. With regard to the wording of descriptors, he said that they should "mirror" any outcomes used to inform the learning and teaching process, as it was with the categories in the rubric and added that he found some of the descriptors somewhat "harsh" or somewhat "negative", which was unavoidable as the level lowered. In terms of the number of band levels, he referred to the nature of the debate on the optimal number of bands. While a large number of test writers preferred the 5-level model, he said he did not – but mostly because of the practical issues of moderation when scoring (an area, he believed, that had not received a great deal of attention in the literature). He preferred a 4-level model as it "drove" raters to make a decision (SOLID PASS / PASS and FAIL / POOR FAIL – based on any required standard – in this case a summative proficiency test). In his extensive personal experience,

fewer teachers “floated” or “drifted” to the “borderline PASS” represented, in this case, 2 MODERATE…and a higher degree of rater alignment occured. Finally, he expressed his opinions on the weighting of the categories. The overall weighing of the writing paper was 20% - again, in line with conventional wisdom in the ELT world. There was no distinction made between the various categories, however, and depending on the range and quantity of learning outcomes, this was something that could be suggested – i.e., allocating a larger weight to a category / categories that required more attention and / or effort on the part of the learner, as highlighted by the second expert, as well.

Overall, it can be concluded that all three experts found the content and number of categories adequate for the performance-based assessment of writing proficiency, while one of them suggested Voice could also be a part of the new rubric. Because it is not emphasized in the learning outcomes of the writing course at present, it is not an option for the time being; however, it needs to be taken into consideration when a review of the outcomes is made since there has been an emphasis recently on authorial voice in the literature of L2 writing (e.g. Matsuda, 2015; Zhao, 2014). In terms of the number of band levels, two of the experts found 5 levels appropriate, while one of them preferred 4 levels due to practical issues of moderation. The results of Phase 1 of the study indicated that the majority of the 24 participants (16) were not satisfied with the current rubric with 3 levels, and stated that 4 levels at the least would be much more useful for the context of this study. In line with the recommendation made by the second expert, the labels of the band levels were modified, and the labels Very Good; Good; Moderate; Poor; Very Poor were replaced with Exceeds Expectations; Meets Expectations; Approaches Expectations; Needs Development; Off Topic/Did Not Try. Depending on the results of Phase 3 of the study where the new rubric with 5 band levels was piloted, refinements could be carried out regarding the number or labels of band levels, which also applied to other aspects of the new rubric such as the weighing of the categories or the wording of descriptors.

In terms of weighting, all three experts emphasized the significance of the context and learning outcomes once again. The second expert indicated that he would give Organization and Vocabulary more weight because the category of Organization in the rubric covered so many more aspects of the text, and Vocabulary might drive students to put more effort into it, which he believed would assist the development of their language ability more than similar time spent focused on Grammar. Due to the lack of consensus

on weighting in not only the literature of performance-based assessment of L2 writing but also the context of BUU-SFL-IEP and the significance of all writing constructs for L2 writing quality, equal weighting of the categories was adopted, which again might be subject to change based on the results of the piloting process in Phase 3.

Finally, all three experts recommended revisions on the wording of some of the descriptors so that they could be easy to interpret, simple to use, and accurate. These revisions were made where applicable; however, as Brown (2012, p. 23) puts forward, while it may be problematic to get raters to use rubrics where adjectives and adverbs are shifted in order to agree on difference like those between "Well-organized", "Fairly well organized", and "Somewhat organized", it seems inevitable to use such distinctions for wide bands of ability. Henceforth, the draft rubric did occasionally have such distinctions. See Appendix 11 for the third draft of the new rubric which was modified on the basis of the recommendations of the three experts.

### 4.2.2. Conclusion of Phase 2

The second phase of the study aimed at designing a draft rubric for the performance-based assessment of writing proficiency relying on the contextual requirements of BUU-SFL-IEP, the related literature, and the expert opinion. The recommendations of the three experts on the content and the number of categories, the number of band levels, the wordings of descriptors, and the weighting brought in the third draft of the new rubric that was used in Phase 3 of the study in order to pilot the new rubric, which is the topic of the following section.

## 4.3. Phase 3: Trial and refinement of draft rubric

Five experienced teacher-raters who worked at BUU-SFL-IEP as EFL as instructors and did not participate in the previous phases were the participants in this phase of the study. In this section results and discussion of the third phase of the study which aimed at trialing and refining the draft rubric were presented and discussed in two parts. The first part was devoted to the results and discussion of the Many Faceted Rasch Measurement (MFRM) analysis which was based on the quantitative data gathered through the scoring process. In the second part the results of the open-ended questionnaire used to collect the qualitative data of this phase of the study were presented and discussed in light of the extant literature.

### 4.3.1. Results and discussion of MFRM analysis

This section presents the statistical results of Phase 3 of the study. Following the MFRM model which the analysis was based on was displayed, the global fit of the data; that is, whether the data fit the MFRM model usefully or not was elucidated. Then, the variable map which displayed the joint calibration of examinees, raters, rubric categories, and band levels was explained. Afterwards, detailed measurement results for each facet were separately presented through fit statistics. Finally, yet importantly, the functioning of the rubric was discussed.

For the ease of following, each of these analyses was allotted a subsection, and findings were discussed in view of the related literature.

#### 4.3.1.1. *MFRM model in the Study*

In this study, three facets were used: the proficiency of examinees, the severity of raters, and the difficulty of the criterion, i.e., the rubric categories. Thus, the equation used in this study was the expression of a *three-facet rating scale model* (Linacre & Wright, 2002). Following Eckes (2009, 2015), the MFRM model used to examine the writing performance sample data in this research can be specified as follows:

$ln\,[p_{nijk}/p_{nijk-1}] = \theta_n - \beta_i - \alpha_j - \tau_k$

$p_{nijk}$ = probability of examinee $n$ receiving a rating of $k$ on criterion $i$ from rater $j$,

$p_{nijk-1}$ = probability of examinee $n$ receiving a rating of $k$-1 on criterion $i$ from rater $j$,

$\theta_n$ = proficiency of examinee $n$,

$\beta_i$ = difficulty of criterion $i$,

$\alpha_j$ = severity of rater $j$,

$\tau_k$ = difficulty of receiving a rating of $k$ relative to rating of $k$-1

Eckes (2009, 2015) gives a clear explanation of the equation and states that a MFRM model is an additive linear model which is built on a logistic transformation of observed ratings to a logit or log-odds scale (In = natural logarithm). He goes on to state that the logistic transformation of ratios of consecutive category probabilities (log odds) can be considered as the dependent variable and the three facets as the independent variables that affect these log odds.

When running FACETS analyses, it is customary to center all facets except one to establish a common origin, usually zero (Engelhard & Myford, 2003). If more than one facet is noncentered, then ambiguity may result since the frame of reference is not

sufficiently constrained (Linacre, 1998). Following Eckes (2015), to establish the origin of the logit scale and make the model identifiable, the rater and criterion facets were centered, which means these facets were constrained to have a mean element measure of zero. The examinee facet was the only facet that was left non-centered.

It is of great importance to note that researchers are able to draw useful, diagnostically informative comparisons among the various facets only if the rating data show sufficient *fit* to the model (Engelhard & Myford, 2003), which is the topic of the next session.

#### 4.3.1.2. *Global model fit*

As Rasch models are idealizations of empirical observations, empirical data will never fit a given Rasch model perfectly. The key issue is the *practical utility* of the model; that is, whether the data fits the model usefully or not, and, when misfit is detected, how much misfit there is and where it stems from (Eckes, 2009, p. 27; Eckes, 2015, p. 69).

One way to assess overall data-model fit is to check the differences between responses that were observed and responses that were expected on the basis of the model. These differences between observed and expected responses are generally indicated as *standardized residuals*.

Linacre (2008) states that satisfactory model fit is indicated when about 5% or less of standardized residuals ≥ 2, and about 1% or less of standardized residuals ≥ 3.
Considering the writing performance data in this third phase of the study that was gathered by using the draft rubric, there was a total of 250 valid responses (5 raters × 10 essays × 5 categories = 250), which were used for estimation of model parameters. Totally, there were 11 unexpected responses that did not fit the expectations of the model. See Appendix 22 for the table presenting unexpected responses. Of these, 10 responses (or 4%) were associated with standardized residuals ≥ 2, and 1 response (or 0.4%) was associated with standardized residuals ≥ 3. Following Linacre (2008), it can then be concluded that satisfactory model fit was achieved.

#### 4.3.1.3. *Variable map*

A key feature of the results is a graphical display that illustrates the calibration of the facets involved in the assessment process. As mentioned above in the data analysis section, this graphical display is called a variable map (a logit scale or the Wright map)

(Eckes, 2009, 2015). It is also called a vertical ruler (Brown & Edmonds, 2012). One can see how examinees' abilities varied, how severe or lenient raters were, and how difficult categories were on this common logit scale. All measures of the facets included in the assessment process are positioned vertically on the same latent dimension, with logits as measurement units. The logit measures represent the range of scores on a true interval scale as opposed to raw test scores where the distances between intervals may not be equal (Edmonds, 2012). Figure 4.1 demonstrates the variable map representing the calibrations of examinee proficiencies, rater severities, category difficulties, and five-level scale as raters used it to score examinee essays in Phase 3 of the study.

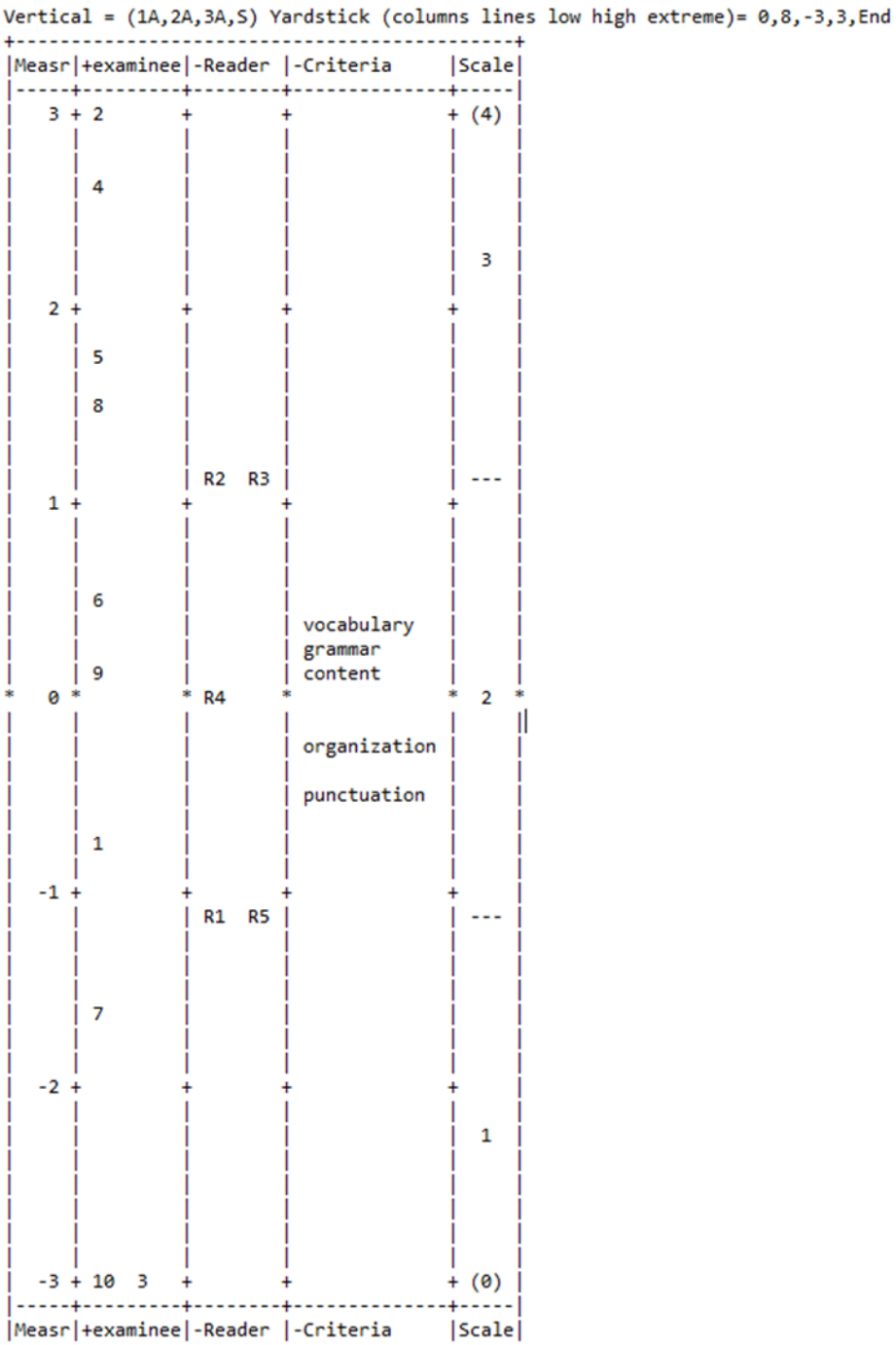


**Figure 3.1.** *Variable map for the dratf rubric*

As can be seen in Figure 4.1, the first column is the *logit (measurement) scale*. This scale is shown in logit scores where the mean is 0 and the range is -3.00 to +3.00 (in this case). On the logit scale a higher score equals a positive logit, that is, a higher measure. In the same way, a lower score equals a negative logit, that is, a lower measure.

The second column, which is labeled “Examinee”, shows the estimates of the examinee proficiency parameter. Each star represents an examinee. As the plus sign before the examinee parameter $\theta_n$ in the equation above indicated, the examinee facet is positively oriented. Thereby, higher-scoring examinees appear at the top of the column, and lower-scoring examinees appear at the bottom. According to Figure 4.1, it is possible to rank the variation in examinee proficiency ranging from -3.00 to 3.00 on the logit scale. The examinees are extended widely along the measure, with more than half of them (6) above 0.00 logits, and the rest (4) positioned at or below 0.00 logits. For instance, an examinee whose proficiency calculation was 3.00 logits on the logit scale is likely to get the highest raw score (four in this case) in all categories when s/he is assessed by an average-severity rater.

The third column labeled “Rater” displays rater severity; that is, compares the raters in terms of the level of severity or leniency each practiced when rating essays. Each of the 5 raters was assigned a number from 1 to 5 as R1 and so on. Unlike the examinee facet, the rater facet has a negative orientation, as the minus sign before the rater parameter $\alpha j$ in the equation above pointed out. It indicates that the higher the rater measure the lower the raw score. Thus, more severe raters appear higher in the scale, and more lenient raters appear lower in the scale.Figure 4.1 shows that almost half of the raters (2) are situated above 0.00 logits, which means they tended to rate the student essays severely (R2 and R3). One of the raters was of average severity whose severity measure is 0.00 on the logit scale (R4). The remaining two raters are positioned below 0.00 logits, the more lenient end of the scale (R1 and R5).

The fourth column labeled “Category” shows the variation in category difficulties; that is, it compares the five rubric categories in terms of their relative difficulties. As it is with the rater facet, the criteria facet is negatively oriented, which means categories situated higher in the logit scale were more difficult than those appearing lower. Hence, the higher the difficulty measure of a specific category, the more difficult it was for examinees to receive a high score on that category. As displayed in Figure 4.1, this column has the least variation of all, with all the categories clustered more or less around the mean 0.00 logits. Vocabulary and Grammar were more difficult than the other categories, while Content and Organization were similarly difficult. Punctuation, Spelling, and Mechanics was the easiest of all categories. The findings of the qualitative part of this phase of the study support the quantitative findings by indicating that the

category of Mechanics was the easiest to score while they believed the categories of Vocabulary and Grammar were the most difficult to score for several reasons. The findings of Edmonds (2012) indicate similar results regarding the category of Mechanics, and for Edmonds these findings are not too surprising as most of the teacher-raters felt most confident in this category.

The last column presents the five-level scale for the five categories and how they relate to the logit scores; that is, distances between each raw scale as they were calibrated onto the true interval logit scale. These demonstrate what any student writer at any proficiency level on the scale are likely to receive. Each horizontal dashed line indicates +0.5 score points, i.e., *category thresholds*, or *Rasch-half-score-point thresholds* (Eckes, 2015: 15). For example, on Figure 4.1 one can see that an examinee whose ability estimate was 3.00 logits on the logit scale is likely to receive almost all correct on their raw scores when s/he is assessed by an average-severity rater. One can also notice that not all five score ranges are the same despite being very similar. While the score ranges of 1, 2, and 3 are almost the same, the score range of 0 and 4 are a little narrower, which indicates that the entire range of 1 to 3 was being used in approximately equal intervals by the raters. Still, there is an almost equal distribution of each band level among the 10 examinees, where 2 examinees are awarded each score level from 0 to 4 (5 score levels). This finding may result from the essay selection criteria used in this phase of the study. As mentioned above in section 3.4.3.2, 10 student essays were selected by the researcher in a way that could represent each score level from 0 to 4 since this was the piloting phase of the draft rubric. Therefore, essays which could be considered irrelevant or lack a few paragraphs were also included in the batch of the 10 essays unlike the essays selected for the fourth phase of the study.

In addition to the summary map provided in the variable map displayed in Figure 4.1, detailed measurement results for each facet are separately presented through fit statistics, which is the topic of the next section.

#### 4.3.1.4. *Fit statistics*

In order for MFRM analysis to interpret examinee proficiency, rater severity, or category difficulty, the data need to fit the MFRM model. For this reason, the summary fit statistics which FACETS provide for examinee fit, rater fit, and category fit will

initially be presented and discussed before the in-depth measurement results for each facet.

#### 4.3.1.4.1. *Summary MFRM results*

Table 4.11 shows the summary MFRM results for the writing performance data in this study. The upper half of the table gives the means and standard deviations of the examinee, rater, and category measures, along with the mean standard errors of the respective measures, the root mean-square measurement errors, and the adjusted (true) standard deviations. The lower half of the table displays the separation statistics, i.e., several group-level statistical indicators for the present three-facet sample data including ♯ Misfit, Homogeneity Index, Separation Ratio, Separation Index, Separation Reliability, and Chi-square (fixed).

**Table 4.11.** *Summary MFRM statistics of the writing performance data*

| **Statistic** | **Examinees** | **Raters** | **Categories** |
|---|---|---|---|
| Mean (*M*, measure) | -.38 | .00 | .00 |
| Standard Deviation (*S.D.*, measure) | 3.30 | 1.08 | .37 |
| Standard Error (*S.E.*) | .46 | .23 | .23 |
| RMSE | .31 | .23 | .23 |
| Adj. (true) *S.D.* | 1.85 | 1.06 | .29 |
| ♯ Misfit | 1 | 0 | 0 |
| Homogeneity index (*Q*) | 325.5** | 87.8** | 10.3 |
| *df* (degrees of freedom) | 9 | 4 | 4 |
| Separation ratio (*G*) | 4.97 | 4.60 | 1.26 |
| Separation (strata) index (*H*) | 8.79 | 6.47 | 2.01 |
| Separation reliability (*R*) | .96 | .95 | .61 |
| Chi-square (fixed) | *p* = .00 | *p* = .00 | *p* = .04 |

** p < .01

As mentioned above in Section 3.4.3.7, the ♯ *Misfit* indicates the number of examinees, raters, or categories that "did not fit the general pattern of responses in the matrix, and can thus be classified as relatively misfitting…" (McNamara, 1996, p. 171). According to Brown and Edmonds (2012), various factors may cause misfitting responses such as examinees with language proficiency levels that are outside the possible range of

scores, raters with inadequate training or attention to the task, or categories with poorly formulated descriptors. Out of 10 examinees, there was one misfitting examinee, which means this examinee did not fit the measurement model as his/her response pattern was unexpected. There were not any misfitting raters or categories in the writing performance data.

Before moving onto the explanations for MFRM summary results, an important issue to be considered is that FACETS output provides both population and sample versions of separation statistics. Eckes (2015) emphasizes that population statistics should be used if it can be presumed that the element list includes the entire population of elements; if not, sample statistics should be used. Since the writing performance data in this study do not comprise the whole population, sample statistics are reported for each facet.

Firstly, *RMSE* refers to *root mean-square standard measurement error* (Eckes, 2015). Brown and Edmond (2012) support that lower RMSE values indicate better data fit with the measurement model. Considering the relatively low RSME values in this study ranging from .23 to .31, it can be concluded that the three facets are fitting the model as desired.

*Homogeneity index (Q)* shows whether as a minimum of two elements within each facet have measures that differ in a statistically significant way. As indicated in Table 4.11, while the two of the three facets, i.e., examinee and rater in the writing performance data in the present study had measures that were different in this respect, the categories in the rubric did not; that is, at least two rubric categories shared the same value of the difficulty parameter.

With reference to the *separation ratio (G)*, the value of 4.97 for the examinee facet pointed out that the variability of the examinee proficiency measures was almost five times larger than the precision of those measures. The *G* value attained for the rater facet indicated that the variability of the severity measures was more than four and a half times larger than their precision. In comparison with the examinee facet, the calculation of measures for elements of the category and rater facets rested on a greater number of observations (each measure was calculated based on 50 observations as opposed to 25 observations for the examinee facet). These measures were then calculated with a lower error measurement error (i.e., *RMSE* =.23). Thereby, the examinee separation ratio obtained a value higher than the rater or category separation ratio (Eckes, 2015). Using

the separation ratio, the separation index (the number of statistically different levels of examinee proficiency, rater severity, or category difficulty) can be calculated, which is the topic of the next paragraph and is discussed thoroughly there.

The *separation (strata)* or *index (H)* refers to the degree to which the examinees, raters, and categories spread out relative to their degree of precision (Linacre, 2008, p. 149). According to Brown and Edmonds (2012), the higher the separation index value, the more efficiently each facet is spreading its elements out. In the sample, the value of the examinee proficiency index was 8.79, indicating that among the 10 examinees included in the analysis, there were more than eight-and-a-half statistically different group of examinees. The separation index calculated for the rater facet was 6.47, suggesting the five raters differed significantly from each other in terms of their severity indices. Lower might have been better; however, considering that the raters in this study were not given a training and norming session for the new writing rubric, and also this was a first time of rating, it is satisfactory (Edmonds, 2012, p. 111). More importantly, as emphasized by Brown and Edmonds (2012) and Eckes (2015), the MFRM approach puts the emphasis on the internal consistency among raters since it will make statistical modeling of rater characteristics possible. Besides, this approach considers variability in stable rater characteristics as the reality of life. Thus, as long as raters are internally consistent, such variability can be accounted for through either multiple rating by averaging of scores or the use of more advanced means of MFRM. Finally, the separation index for the category facet was 2.01, a value lower than the number of categories included in the analysis, and also a value less than the separation index values of the other two facets, indicating that there were not much difference among the categories in terms of their difficulty levels. In other words, the five categories were not as efficient as the examinee and rater facets in spreading its elements out.

The next statistic, *separation reliability (R)* needs to be interpreted differently depending on the facet to be discussed (Eckes, 2015). For examinees, the examinee separation reliability indicates how different the examinee proficiency measures are. The high reliability for examinees in the current research (.96) shows that the examinees consistently differ from each other, which Brown and Edmonds (2012, p. 78) believes is the aim of a testing situation like the one described in this study. For raters the interpretation of the separation reliability differs from that of examinees (Eckes, 2015). Eckes (2015, p. 66) explains the difference explicitly with reference to standard approach

to rater reliability. The standard approach to rater variability supports that when raters within a group practiced a highly similar degree of severity, rater separation reliability will be close to 0; thus, it aims for *low* rater separation reliability as this would signal that raters were approaching the actually "impossible" ideal of being interchangeable. In contrast, when raters within a group practiced a highly different degree of severity, rater separation reliability will be close to 1. This means that unlike interrater reliability, which in general terms is an index of how *similar* raters are in terms of their severity, rater separation reliability is an index of how *different* severity measures are. Therefore, as emphasized by Eckes (ibid.), these two kinds of reliability indices needs to be differentiated. In the present study, rater separation reliability was as high as .95, demonstrating a remarkable heterogeneity of severity indices. It is important to note that from the perspective of the standard approach to rater reliability this value would be equal to around .80 (interclass correlation coefficient). The ideal reliability coefficient for performance assessment is between "the .70 to .79 ranges, which is considered sufficient for tests assessing written performance" (Hughes, 1989, p. 39) and shows that raters were functioning almost interchangeably. However, as the rater separation indices and separation reliability statistics clearly displayed, this conclusion would be misguiding. In Eckes's (2015) on words

> "… Actually, traditional group-level reliability statistics often mask non-negligible differences within a group of raters, lulling those in charge of assessment programs into a false sense of security" (p. 66).

Hence, using more sophisticated means of MFRM modelling and monitoring whether the variability in rater severity indices of each rater is within the accepted limits (or not) would provide more reliable information for language programs with a performance assessment component for assessing L2 writing proficiency. See Eckes (2015, pp. 42-50) for a detailed discussion of the pitfalls of the standard approach to interrater reliability.

Finally, the separation reliability for the category facet was 0.61, which points out that the categories differed moderately from each other in terms of difficulty and were doing different things to a moderate degree. According to Eckes (2015), when the set of criteria is designed to cover a wide range of performance features spread out across the underlying difficulty dimension (which was the case in the current study), high values of this statistic would be desirable. Thus, a higher value is expected to be reached at in Phase

4 (for the separation reliability for the category facet) based on the feedback that will be gathered from the analysis of the qualitative data in this phase of the study, which aimed at trialing and refining the draft rubric.

The last statistic in Table 4.11 is the *chi-square (fixed)* values. They indicate that the chi-square statistics for the three facets in the study were significant ($p<.01$), which means that the examinees probably differed from each other in proficiency for reasons other than chance; that the raters did, too; and the categories probably differed from each other for reasons that can only be attributed to chance.

In addition to the variable map and summary Rasch results, the MFRM modelling provides in-depth measurement results for each facet, which will be presented and discussed thoroughly in the following subsections.

#### 4.3.1.4.2. *Fit statistics for each facet*

Another important part of the output of Rasch-based analyses are mean-square *fit statistics* provided for each facet included in the measurement process. Two fit statistics are commonly used: *infit* (information weighted fit statistic) and *outfit* (outlier-sensitive unweighted fit statistic). As the names imply, outfit values include all the observations in the data set, and are hence sensitive to atypical outlying values, whereas infit values indicate the extent of score variability in a given data set which remains after the extreme values (outliers) have been removed (Davies et al., 1999). In other words, infit statistic is more sensitive to variability in the range of observations that are usually of most interest. For this reason, infit is generally deemed to be more important than outfit in estimating model fit (Bond & Fox, 2007; Davies et al., 1999; Edmonds, 2012; Linacre, 2002c, 2008; Myford & Wolfe, 2003).

Infit and outfit statistics close to 0 demonstrates that the scores for a specific examinee, rater, or category falls within the normal range (Davies et al., 1999; Linacre, 2002c, 2008; Myford & Wolfe, 2003). Fit values that are greater than 1 indicate more variation from the norm than expected and tend to *misfit* (underfit) the model as also mentioned above. Conversely, fit values lower than 1 refers to less variation than expected and tend to *overfit* the model. Misfit is considered to be more problematic than overfit (Myford and Wolfe, 2003).

As a general rule for fit statistics, Linacre (2002c, 2008) recommended 0.50 as a lower control limit and 1.50 as an upper control limit for the infit and outfit mean squares.

Thus, mean square values within the range between 0.50 and 1.50 is considered to be useful fit. Values greater than 1.50 indicate significant misfit, and values lower than 0.50 show significant overfit. Other researchers suggested a narrower range: 0.75 and 1.30 (Bond & Fox, 2007; McNamara, 1996). According to Eckes (2009, 2015), the decision depends on the nature of the assessment purpose. For the purposes of this study, the range between 0.50 and 1.50 is adopted; however, as the reader will see in the following subsections, the three facets included in the writing sample data in the current research indicates largely useful fit even if a narrower range is considered to be appropriate.

### *Examinee fit statistics*

As demonstrated in the variable map (Figure 4.1), there was a variation in examinees' writing proficiency despite the small sample size (10 essays). The variation in examinee proficiency ranged from -3.00 to 3.00 on the logit scale, which means examinees were extended along 6 logits. What is more, separation statistics (Table 4.11.) indicated that the examinee separation (strata) index was 8.79, with an examinee separation reliability of .96; that is, examinees were well-differentiated in accordance with their level of writing proficiency. Table 4.12 provides an in-depth measurement of the examinees' writing proficiency.

**Table 2.12.** *Examinee fit statistics*

| **Examinee** | **+ Measure** | **Model *SE*** | **Infit *MnSq*** | **Outfit *MnSq*** |
|---|---|---|---|---|
| 3 | -8.10 | 1.84 | Minimum | |
| 10 | -2.99 | .34 | .85 | .82 |
| 7 | -1.58 | .64 | .64 | .64 |
| 1 | -.77 | .30 | 1.15 | 1.15 |
| 9 | .12 | .30 | 1.39 | 1.38 |
| 6 | .55 | .30 | .84 | .83 |
| 8 | 1.52 | .30 | .73 | .73 |
| 5 | 1.79 | .30 | 1.59 | 1.59 |
| 4 | 2.65 | .32 | 1.01 | 1.04 |
| 2 | 2.97 | .33 | .59 | .68 |
| **Mean** | -.38 | .46 | .98 | .98 |

+: examinee's proficiency positively oriented (high logit values means high proficiency)

From the left, the columns display examinee identification (each of which was numbered for easy retrieval and anonymity), the logit measures of the examinees' proficiency followed by standard error and infit and outfit mean square values. More specifically, Column 1 shows the order of examinees based on the measure of proficiency in Column 2. Column 2 demonstrates that the proficiency span between the lowest scoring examinee (Examinee 3) and the highest scoring examinee (Examinee 4) was 5.96 logits. The differences in proficiency ranged from -2.99 to 2.97 logits (6 logits). Column 3 indicates that the standard error (*SE*) ranged from 0.30 to 0.34. Column 4 and Column 5 show the infit and outfit mean square values; that is, they present the extent to which the data representing each examinee fell within the accepted boundaries (between 0.50 and 1.50). As also demonstrated above in Table 4.11, there is only one examinee that fell out of the useful fit. Table 4.12 shows that Examinee 5 (1.53) had fit statistics that were greater than 1.50, which could be said to be misfitting and not performing as the model predicted. The performance of this examinee needs to be evaluated further to find out what might be the reason behind the non-predictive behavior. It is important to note that misfitting or overfitting does not necessarily mean that the examinees, raters, or categories are problematic; it just means that the examinees did not perform according to the model, or they performed too well within the model (Edmonds, 2012). The remainder of the examinees fell between the reasonable boundaries (between 0.50 and 1.50), and eight of them did so even the narrower range (between 0.75 and 1.30) was adopted.

Overall, based on the examinee proficiency measurement report it can be concluded that the examinees' scores were reasonably well distributed and even though one of the performances was outside the model's expectations, the rest were well within the expectations. The reliability index on the examinee measurement was very high at 0.96 which pinpoints that it is very likely that the examinees would perform in a similar way if another test that aimed at measuring the same construct was administered (Bond & Fox, 2007, p. 40). One drawback of the analysis could be the number of the examinees which was only 10, yet the analysis did indicate that there was a high person reliability in which some of the examinees scored higher and some lower in a consistent manner.

***Rater fit statistics***

The variable map displayed in Figure 4.1 showed clearly that the raters in this study varied substantially in their measures of severity, which was also supported by the separation index (6.47) and the separation reliability (.95) statistics displayed in Table

4.11. However, as mentioned above and as the following fit statistics indicate, there are not any misfitting raters, either. Table 4.13 presents the detailed rater measurement report in a similar way to the examinee report.

**Table 4.13.** *Rater fit statistics*

| **Rater** | **-Measure** | **Model *SE*** | **Infit *MnSq*** | **Outfit *MnSq*** |
|---|---|---|---|---|
| 1 | -1.08 | .23 | 1.11 | 1.14 |
| 5 | -1.08 | .23 | 1.34 | 1.26 |
| 4 | -0.3 | .23 | .79 | .81 |
| 2 | 1.09 | .23 | .75 | .73 |
| 3 | 1.09 | .23 | .98 | .97 |
| **Mean** | .00 | .23 | .99 | .98 |

From the left, the columns demonstrate rater identification in numbers, the logit measures of rater severity followed by standard error and fit statistics. Column 1 displays the rank of the raters based on the measure of severity in Column 2. Column 2 pinpoints that the differences in severity ranged from -1.08 to 1.09, which means the severity range between the most severe raters (R2 and R3) and the most lenient raters (R1 and R5) was 2.89 logits. The standard error was 0.23. Column 4 and Column 5 present the infit and outfit mean square values; that is, they display the extent to which the data representing each rater fell within the reasonable boundaries (between 0.50 and 1.50). These statistics indicate that all 5 raters fell between the reasonable boundaries (between 0.50 and 1.50), and three of them did so even the narrower range (between 0.75 and 1.30) was adopted, which means their rating behavior was independent. As Edmonds (2012, p. 111) puts forward, lower indices might have been preferable; however, considering that the raters in this study were not given a training and norming session for the new writing rubric, and also this was a first time of rating, it is satisfactory. Moreover, as also mentioned above, the MFRM approach considers variability in stable rater characteristics a fact of life (McNamara, 1996) and highlights the internal consistency among raters since it will make statistical modeling of rater characteristics possible (Brown & Edmonds, 2012; Eckes, 2015) and also what causes variability in rater severity still remains a mystery. As Eckes (2009) emphasizes, the research into the stability and change in rater severity and the personal and situational factors that affect rater severity is considerably sparse.

Among some of the factors he lists are teaching and rating experience, demographic and/or personal characteristics, workload, and assessment aim. Considering the teacher-raters who participated in this phase of the current research, teaching-rating experience does not seem to be one of the factors of the variability in rater severity. All five teacher-raters in Phase 3 had more than 10 years of professional experience in the ELT field, and their rating behavior was independent with a reliability of 0.95 and a separation index of 6.47. Regardless of the factors involved, the variability in rater severity could be observed in relatively recent research that used MFRM to test the validity of L2 writing rubrics in spite of the efforts to reach rater agreements with extensive training and norming sessions (e.g., Becker, 2018; Eckes, 2009, 2015; Edmonds, 2012; Hattingh, 2009; Knoch, 2007). Thus, rater training needs to aim at increasing raters' internal consistency and reducing extreme levels of rater severity or leniency rather than trying to reach an inter-rater agreement as Eckes explains in his following words:

> Rater training usually does not succeed in reducing between-rater severity differences to an acceptably low level. Therefore, in most situations, adopting the standard view that rater training needs to achieve maximal between-rater similarity, and eagerly pursuing this objective in rater training sessions, is extremely likely to end up in frustration of those in charge of the training. The constructive alternative to striving after fictitious rater homogeneity is to accept rater heterogeneity within reasonable bounds and to adopt a suitable psychometric modeling approach. Many-facet Rasch measurement provides the tools to probe deeply into the complexities of rater behavior and to use the insights gained for the purposes of making performance assessments as fair as possible (Eckes, 2015, p. 73)

Overall, all 18 teacher-raters, who were included in this phase of the study (5 raters) and Phase 4 of the research (13), were internally consistent despite the lack of a training and a norming session and their unfamiliarity with the new rubric. There were not any misfitting raters who were outside of the model's expectations, and the vast majority of the examinees responses and all of the raters' performances were well within the model's expectations.

***Category fit statistics***

The difficulty measurement report of the five categories, i.e., Content, Organization, Grammar, Vocabulary, and Punctuation, Spelling, and Mechanics, is demonstrated in Table 4.14 in the same way as the examinee proficiency and rater severity reports.

**Table 4.14.** *Category fit statistics*

| Categories | -Measure | Model *SE* | Infit *MnSq* | Outfit *MnSq* |
|---|---|---|---|---|
| Punctuation | -0.52 | .23 | 1.14 | 1.17 |
| Organization | -.20 | .23 | .80 | .83 |
| Content | .06 | .23 | 1.45 | 1.41 |
| Grammar | .22 | .23 | .85 | .82 |
| Vocabulary | .43 | .23 | .70 | .68 |
| **Mean** | .00 | .23 | .99 | .98 |

From the left, the columns present category identification, variance in category difficulty, error, and fit statistics. Column 1 shows the order of categories based on the measure of difficulty in Column 2. Column 2 pinpoints that the differences in category difficulty ranged from – 0.52 to 0.43, which means the difficulty range between the most leniently scored category (Punctuation, Spelling, and Mechanics) and the most severely scored category (Vocabulary) was -0.95. When compared with the logit measure ranges of the examinee (8.79) and the rater (6.47) facets, the logit measure range of the category facet was smaller indicating that the difficulty measures of categories did not vary as much as the examinee proficiency and rater severity measures. Column 3 shows that the standard error was 0.23. Column 4 and Column 5 indicate the infit and outfit mean square values; that is, they display the extent to which the data representing each category fell within the reasonable boundaries (between 0.50 and 1.50). These statistics indicate that all 5 categories fell between the reasonable boundaries (between 0.50 and 1.50) with no misfitting or overfitting data. Moreover, three of them did so even the narrower quality control limits (between 0.75 and 1.30) were adopted. This finding is in compliance with the assumption of psychometric unidimensionality of the set of categories in this study (McNamara, 1996), which means all five categories seemed to relate to the same dimension i.e., examinee writing proficiency, as assumed by MFRM (Eckes, 2015). The difference in category difficulty was small, and the reliability 0.61 indicated that the categories were performing independent of each other at a moderate level, which is expected to be compensated in Phase 4 based on the feedback gathered form the qualitative data in this phase of the study. All in all, the five categories did not vary greatly in terms of difficulty, and none of them displayed any significant misfit and overfit, supporting that the multi-trait rubric behaved to a great extent as the model might expect.

#### 4.3.1.5. *Band level (rating scale) analysis*

In addition to category difficulty, the quality of the five-level scale used by the raters to evaluate examinee proficiency is of great importance. A variety of statistical indices are utilized to investigate rating scale validation, i.e., whether the five band levels in the draft rubric performed as intended. Based on Linacre (2004b) and Bond and Fox (2007), the following three indices are presented by Eckes (2009, 2015) as the indicators of rating scale effectiveness: the *average measure* of each band level, the *mean-square outfit statistic* calculated for each band level, and the *ordering of Rasch-Andrich thresholds*.

The first indicator, the average measure, refers to the average of the examinee proficiency measures modeled to produce the observations in a given band level. It is required that the average measures progress monotonically, which means the higher the band level, the larger the average measure. If this prerequisite is met, it can be concluded that higher ratings equal "more" of the variable that is measured (Eckes, 2009, p. 26).

The second indicator of rating scale effectiveness is the *mean square outfit statistic*, which is the examinee proficiency measure the model estimates for a given level if the data were to fit the model. Generally, this statistic should not be above 2.0 (Eckes, 2009, p. 26).

The final indicator of rating scale effectiveness is *the ordering of the Rasch-Andrich thresholds*. As it is with the average measure, the requirement is that these thresholds should increase monotonically with each level (Eckes, 2009: 26).

Table 4.15 presents the results with regard to these indices. Column 1 shows each band level in the writing rubric, i.e., 0, 1, 2, 3, and 4. Column 2 displays the counts used to estimate the indices. Column 3 indicates the average measure for each band level, Column 4 the mean square outfit statistic again for each band level, Column 5 the Rasch-Andrich thresholds, and Column 6 the standard error.

**Table 4.15.** *Overall category probability statistics*

| **Band Levels** | **Counts** | **Average Measure** | **Outfit Statistic** | **Rasch-Andrich Thresholds** | ***SE*** |
|---|---|---|---|---|---|
| 0 | 17 | -3.15 | .9 | | |
| 1 | 44 | -1.49 | 1.0 | -3.23 | .32 |
| 2 | 71 | .33 | .9 | -1.10 | .22 |
| 3 | 62 | 1.74 | 1.2 | 1.18 | .20 |
| 4 | 31 | 3.04 | .9 | 3.15 | .24 |

As demonstrated in Table 4.15, the average measures of examinee proficiency advanced with each band level. Likewise, values of the mean square outfit statistic were almost equal, or very close, to the desirable value of 1.0. Lastly, there was a monotonic advancement of band level thresholds from -3.23 logits (i.e., the threshold between band levels 1 and 2) to 3.15 logits (i.e., the threshold between band levels 2 and 4). All in all, these indices strongly support that the five band levels of the new rubric were ordered appropriately and functioning as desired.

In addition to the statistics explained above, MFRM provides a graphical illustration, which is called the *probability curves*, for rating scale validation. According to Brown and Edmonds (2012), these curves are beneficial as they graphically demonstrate the degree to which the band levels are distinct or overlapping. Both Eckes (2009: 26) and Brown and Edmonds (2012, p. 80) support that the best would be probability curves that have a "distinct hill-like appearance" with one curve for each band level and some overlap between hills but not too much. Figure 4.2 demonstrates the probability curves for the four-level scale utilized by the raters to rate the examinees on the five-category rubric.

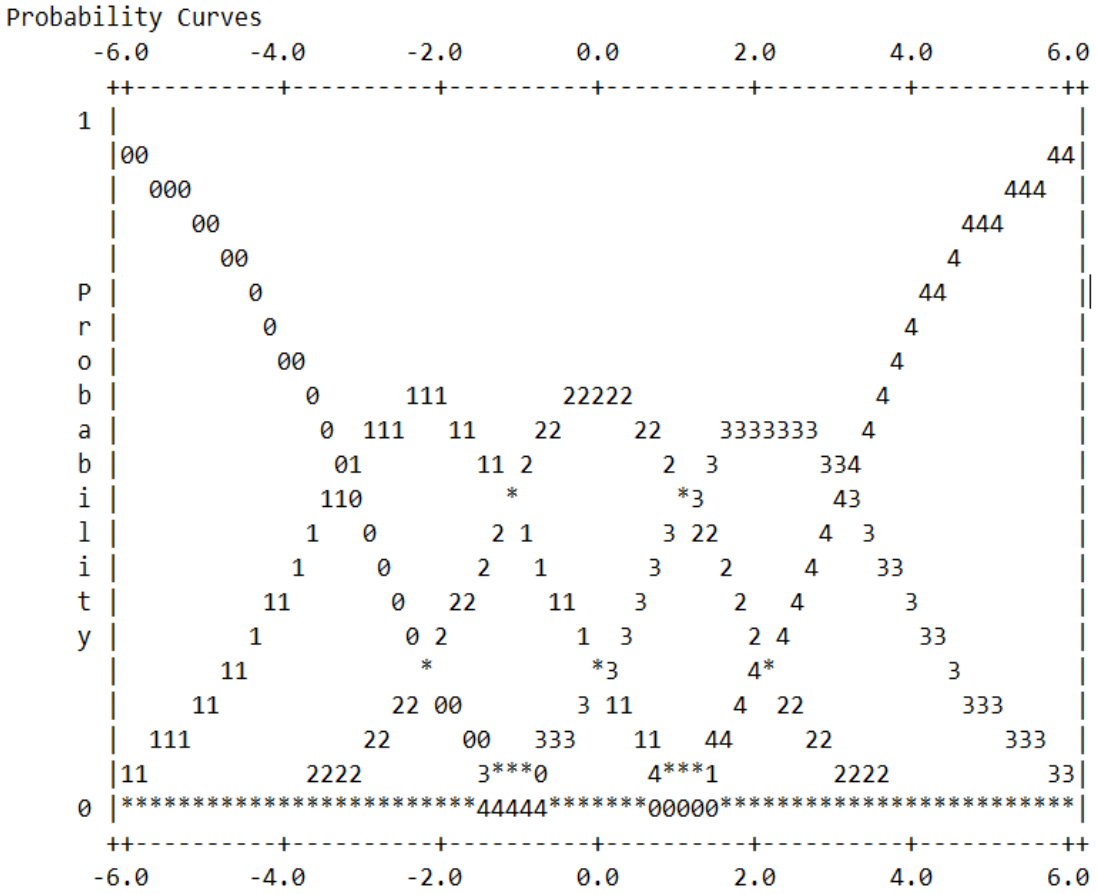


**Figure 4.2.** *Probability curves for the draft rubric*

The horizontal axis gives the examinee proficiency scale and the vertical axis the probability of being rated in each level. As the figure displays, there is a distinct hill for each level with little overlap, and the level thresholds are properly ordered from left to right. It can be concluded that the probability curves the statistical indices in that the levels in the draft rubric were properly ordered and working as expected.

To sum up, fit statistics that are in the form of summary results and fit statistics for each facet accompanied by graphical displays in the MFRM analysis show that the scores the teacher-raters awarded to examinees by using the new rubric to assess their writing proficiency proved to be reliable and valid.

### 4.3.2. Results and discussion of open-ended questionnaire

In addition to the statistical analyses based on the results of the scoring process, the 5 teacher-raters who were participants in this phase of the study were asked to fill in an open-ended questionnaire to find out their perceptions on the draft rubric. See Appendix

21. Specifically, they were requested to share their ideas on the strengths and weaknesses of the draft rubric designed to be used for the assessment of writing performance in the proficiency examination, the efficacy of the draft rubric in fair assessment of students' written work, and their confidence level in using this draft rubric. Each of these areas was allocated a subsection for the ease of following, and findings were discussed in comparison with the first phase of the study which explored the perspectives of the participants on the rubric currently used in the assessment of writing performance, the MFRM analysis carried out in this phase of the study and, the last but not the least, the relevant literature.

#### 4.3.2.1. *Strengths of the draft rubric*

According to the 5 participants of this phase of the study, the strengths of the draft rubric were as follows:

- Comprehensiveness of descriptors,
- Clarity of descriptors,
- Manifestation of the learning outcomes of the writing course in the rubric,
- Allocation of one score for each band level (rather than a range of scores), and
- Equal weighting of each category.

*Comprehensiveness of descriptors* and *clarity of descript*ors were mentioned by all five participants, followed by *manifestation of the learning outcomes of the writing course in the rubric*, which was stated by four participants. Finally, *allocation of one score for each band level* and *equal weighting of each category* were highlighted by three participants.

According to Participant 3 (hereafter P3), who compared the draft rubric with the one currently used at BUU-SFL-IEP:

> 76. The draft rubric is much more detailed and clear in terms of the categories it presents. Also, it makes scoring easier because the rater does not have to decide about let's say whether giving 4, 5, or 6 points within the same band as we have to do now. When you decide on the correct band, you give the exact point that is offered. Moreover, giving the same 4 points to all categories seems so fair because while writing an essay, I believe that any one of the categories is not more important and accordingly does not deserve more points than the others. Finally, goals and objectives of the writing course are wholly reflected in the rubric *(P3).*

Along the same line with Participant 3, Participant 4 lists the advantages of the draft rubric as follows:

> 77. The descriptors in each category are very clear. In the rubric that we are currently using, there are vague terms such as "good", "fair", and "poor". These vague terms are open to interpretation and might cause dramatic differences in the marks the assessors assign. However, with this draft rubric, raters might assign more consistent marks for each category. I find this new rubric quite practical and easy to apply. This stems from the strong wording of descriptors and even weighting of the 20 points among categories *(P4)*.

In addition to the MFRM analysis, the findings gathered by the open-ended questionnaire indicated that the draft rubric might cater for the specific needs of BUU-SFL-IEP to a great extent. However, it was not without its drawbacks, which is the topic of the following section.

#### 4.3.2.2. *Weaknesses of the draft rubric*

The weaknesses of the draft rubric as perceived by the 5 participants in this phase of the study are listed as follows:

- Difficulty of distinguishing between levels 0 and 1,
- Wording of band level 4, "Exceeds expectations", and
- Wording of the label of the Punctuation, Spelling, and Mechanics category,
- Wording of some descriptors.

The first pitfall of the draft rubric, as perceived by 4 participants out of a total of five, was the difficulty of distinguishing between levels 0 and 1. Although the statistical analysis did not indicate any problems in this respect (See Table 4.15 and Figure 4.2), the majority of the participants found it confusing. The other two drawbacks were mentioned each by three participants in this study.

Participant 5 expresses her concerns as follows:

> 78. The most prominent problem I had during marking was separating between 0 and 1 almost in all of the categories. I believe we must give 0 to essays only with a few sentences or students who haven't written an essay. Other than that I think they deserve 1 because of the effort they spend *(P5)*.

Participant 1 shares her opinions and gives a very reasonable recommendation in the following excerpt of hers:

> 79. The expression "Exceeds expectations" is confusing, even a bit intimidating for me. Even a high-quality essay may not "exceed" our expectations due to the language proficiency levels of our students. Another thing I get confused about is levels 0 and 1. I think the solution

> may be having 4 levels labeled 4. Meets expectations 3. Approaches expectations 2. Needs development and 1. Inadequate. This could be much fairer *(P1)*.

The last shortcoming of the draft rubric, as perceived by three participants out of five, was the wording of descriptors, and was directly related to the first pitfall participants referred to, i.e., the difficulty of distinguishing between levels 0 and 1. The majority of the descriptors the participants found difficult to distinguish were in these band levels, as covered in the following sections.

#### 4.3.2.2.1. *Number of band levels*

The lack of adequate number of band levels was mentioned by the participants in Phase 1 of the study under the heading of the weaknesses of the current rubric (See 3.2.6.2 above). Specifically, out of twenty-four participants, 16 of them (67%) considered three band levels to be insufficient. 8 of these 16 participants supported that there should be four band levels while 4 of them believed five band levels were necessary for fair scoring. Regarding the number of band levels, expert opinion was also consulted (Section 3.3.5). Two of the experts found 5 levels appropriate, while one of them preferred 4 levels due to practical issues of moderation. Because this third phase of the study aimed at trialing and refining the draft rubric and necessary changes could be made for Phase 4 based on the results of the current phase, 5 band levels were preferred by the researcher; however, it seemed that neither three nor five band levels were satisfactory for the teacher-raters who participated in Phase 1 and Phase 3 of the study. As also mentioned above in the section allotted to weaknesses of the draft rubric, all 5 participants found it difficult to distinguish between band levels 0 and 1, which meant that 4 band levels with revised labels and descriptors would function more efficiently in the specific context of the current research.

#### 4.3.2.2.2. *Wordings of descriptors in each category*

The third question in the open-ended questionnaire looked into the participants' perceptions of the wordings of descriptors in each of the five categories and whether there were any categories in which participants thought the wording of descriptors needed to be changed. As also mentioned above, out of five participants, four participants referred to the difficulty of distinguishing between the levels 0 and 1, and the majority of the

descriptors the participants found difficult to distinguish were in these band levels, as the following excerpt by Participant 5 indicates:

> 80. Some expressions in the bands are not distinguishable from one another. For example, in the Vocabulary category, what is the difference between *lacks variety* (Level 1) and *no concept of variety* (Level 0)? Some expressions are not clear: Punctuation, *unacceptable to educated readers* (Level 1)? or Organization, *could not be outlined by reader* (Level 0)? Some descriptors overlap: Grammar, *difficult to understand sentences* (Level 1), *unintelligible sentence structure (P5).*

Similar concerns related to the band levels 0 and 1 were mentioned by the other three participants, as well. In addition, Participant 2 stated that she had doubts about wordings of the other band levels and made recommendations about them, as pinpointed by the excerpt below:

> 81. I have doubts about the word *evidence* in the levels 4 and 3 of the Content category. *Ideas* sounds like a better option. Another wording I find confusing is again in the category of Content but this time in band level 2: *essay is somewhat off the topic. Essay may deviate from the topic* could be an option. And in the category of Vocabulary in band level 2 *repetitive use of vocabulary* could be replaced by *repetitive use of basic vocabulary.* I guess this way band levels 1 and 2 may be more distinguishable *(P2)*.

Finally, Participant 4 referred to the label of the category of Punctuation, Spelling, and Mechanics and stated that:

> 82. I would change the label of this category as Writing Conventions and Mechanics or just Mechanics. Mechanics includes "capitalization", "spelling", and "punctuation". It is an umbrella term. I believe it does not make sense to use such a long label when it is possible to use just a word *(P4).*

Regarding the wording of descriptors, the detailed explanations and recommendations of all of the five participants revealed that how diligently they approached the scoring process with the draft rubric. Based on their suggestions, refinements would be carried out before the scoring process in Phase 4 commenced. The next section is devoted to the weighting of the draft rubric.

#### 4.3.2.3. *Weighting*

As mentioned several times within the scope of the current research, the issue of weighting (of categories) has been disputable in the literature of performance-based assessment of writing (East & Cushing, 2016), as reflected in the findings of Phase 1 of this study. Furthermore, the three experts whose opinions were consulted in Phase 2 of the study had also differing perspectives on the issue. Out of five teacher-raters who

participated in this phase of the study, 3 of them found equal weighting advantageous, while two of them were hesitant. The following excerpts by Participant 4 and Participant 1 reflect each point of view successively:

> 83. I believe even weighting is a good idea because it makes the rubric much easier and more practical to apply. I did not have second thoughts *(P4).*
>
> 84. Generally speaking, I would give more weight to Content and Organization because I believe that being able to develop ideas and express them cohesively is more important. On the other hand, equal weighting makes the rater's job easier during scoring I guess *(P1).*

As it was shown in the findings of Phase 1 (Section 3.2.6.6), there was no one right answer for the question of weighting. Equal weighting of the categories still seems to be the best solution because of the controversy on the issue and, more importantly, the significance of all writing constructs for L2 writing quality.

#### 4.3.2.4. *Categories considered difficult to score*

All five participants answered this question in a similar fashion by stating that apart from the band levels 0 and 1 almost in all of the categories and wordings of some descriptors in a few categories (as indicated in previous sections), they generally found the categories very satisfactory, as reflected in their answers to the last two questions of the open-ended questionnaire, which aimed at exploring their general satisfaction level with the draft rubric.

#### 4.3.2.5. *Participants' general satisfaction level*

In the last section of the open-ended questionnaire 5 teacher-raters who participated in this phase of the study were asked to rate the current writing rubric from 1 to 5 (ranging from very satisfied to very dissatisfied) on:

- the extent to which it facilitates fair assessment of students' written work and
- the participants' confidence level in applying the writing rubric.

##### 4.3.2.5.1. *Participants' perceptions on fair assessment of students' written work*

The first item about the general satisfaction of participants in using the draft rubric explored the extent to which it assists in the fair assessment of students' written work. Participants were asked to rate the draft rubric from 1 to 5:

5. Very satisfied

4. Satisfied

3. Neither satisfied nor dissatisfied

2. Dissatisfied

1. Very dissatisfied

As an answer to this questions, one participant stated that she was very satisfied while four of them said they were satisfied. Two of the participants who stated they were satisfied expressed their need for a training and norming session, as the following excerpts by Participant 2 and Participant 4 demonstrate:

> 85. What's ideal is to have norming meetings before each grading process, but having a written form of norming may also enhance fair assessment. Thus, I expect to see more explanations for each level or an extra sheet of paper explaining what to expect in detail *(P2)*.
>
> 86. This rubric could facilitate fair assessment of students' written work on condition that assessors receive effective norming and express willingness and motivation to "use it efficiently" *(P4)*.

Both participants referred to the necessity of having a norming session in their explanations. As mentioned before (Section 3.4.3.1), participants were not given a training and a norming session in order not to affect their decision-making processes in an undesirable way and also to discover the extent which the draft rubric functions effectively on its own. However, a much more detailed assessor guide could be prepared for Phase 4 of the study, following the recommendation of Participant 2. In Phase 1 of this study where the teacher-raters' expectations from a writing rubric were explored, findings indicated a rubric with a concrete and objective formulation style in which descriptors in each band can be transformed into a checklist of "yes" or "no" questions (e.g., *organization*: appropriate title, effective introductory paragraph, topic is stated, leads to body, transitional expressions used; supporting evidence given for generalizations; conclusion logical & complete) (Section 3.2.6.4). Thus, it would be more meaningful to prepare a comprehensive assessor guide which could guide the assessors when needed rather than a condensed rubric which might seem overwhelming.

In addition to training and norming, Participant 4 raised a very important issue, which was the willingness and motivation of the assessors to use the new rubric. Apart from brainstorming and learning from each other, making the volunteering teacher-raters an important part of the different phases of this study was thought to be a way of giving them a voice, acknowledging their experience and expertise, and increase their

enthusiasm for using the rubric. Brown puts the importance of this into words nicely in his following excerpt:

> …In my view, given that such decisions (decisions on rubric design) can differ dramatically in various language courses or programs, such decisions should most often rest with the experts, who most often turn out to be the teachers of the course or program, who will have to use the rubrics, who need to buy into their use, and who can kill a rubric if they do not approve of it (Brown, 2012, p. 21).

Out of 52 English instructors who work at BUU-SFL-IEP, 44 of them volunteered to participate in the different phases of the study. Rubric design is not one-shot; it is, on the other hand, an ongoing process, as Janssen et al. (2015) put forward. Therefore, it is expected that all English instructors will be a part of the revision process when the new rubric is started to be used for the performance-based assessment of writing proficiency at BUU-SFL-IEP.

#### 4.3.2.5.2. *Participants' confidence level in applying the draft writing rubric*

The second item in this section investigated participants' confidence level in applying the new writing rubric. In compliance with the findings in the previous section, out of five participants one participant said that she was very confident, and the remaining four participants stated that they were confident with the draft rubric, which is pleasing as this was the first time they used the draft rubric.

Considering the findings of the open-ended questionnaire, following refinements were carried out in the draft rubric:

- Band level 0 was eliminated from all the categories because it was considered confusing and redundant by the participants,
- The labels of band levels were modified as 4. Meets expectations, 3. Approaches expectations, 2. Needs development, 1. Inadequate so as not to mislead assessors by using the phrase "Exceeds expectations",
- The label of the category of Punctuation, Spelling, and Mechanics was changed into Mechanics because it functioned as an umbrella term for all the elements included in this category,
- Wording of some descriptors were refined, and the last but not the least,

- A much more detailed assessor guide was prepared to be able to give assessors more in-depth information on the writing constructs that were aimed to be evaluated (See the revised Assessor Guide in Appendix 19).

See the revised edition of the rubric in Appendix 20.

### 4.3.3. Conclusion of Phase 3

The aim of the third phase of the study was to pilot the draft rubric through an initial implementation, identify possible weaknesses in the use of it, and refine the rubric based on the feedback that would be received from the raters, as carried out in Knoch (2007) and Hattingh (2009). Another aim of this phase was to pilot the open-ended questionnaire that would be used to explore the perceptions of the raters on the efficacy of the draft rubric and discover the potential problems that may exist in the open-ended questionnaire, such as the clarity of the items. (Zacharias, 2012, p. 71).

The results of the MFRM analysis showed that the draft rubric designed to be used for the performance-based assessment of EFL writing proficiency at BUU-SFL-IEP would function reliably and validly to a great extent, whereas the findings of the open-ended questionnaire indicated the necessity of making slight revisions in the rubric. In order to be able to build a more functional rubric, the modifications were carried out by the researcher of the study (See Appendix 20).

Another purpose of the study was to trial the open-ended questionnaire to be utilized to find out the perceptions of the participants on the new rubric. The depth and width of the data gathered through the open-ended questionnaire demonstrated that the instrument would function as required.

Together with the completion of the third phase of the study, the piloting of the new rubric and the open-ended questionnaire to be used in Phases 4 and 5 of the current research was completed.

## 4.4. Phase 4: Psychometric analysis of the new rubric through Many Faceted Rasch Measurement (MFRM)

13 experienced teacher-raters who worked at BUU-SFL-IEP and who did not participate in the previous phases were the participants in this phase of the present research. The aim of the fourth phase, which yielded quantitative data, was to empirically

validate the alternative multi-trait rubric designed for the performance-based assessment of writing proficiency at BUU-SFL-IEP through statistical analyses, specifically MFRM. This section presents the results of these statistical analyses and their discussion. Following the MFRM model which the analysis is based on is displayed, the global fit of the data; that is, whether the data fits the model usefully or not is elucidated. Then, the variable map which displays the joint calibration of examinees, raters, rubric categories, and band levels is explained. Afterwards, detailed measurement results for each facet are separately presented through fit statistics. Finally yet importantly, the functioning of the rubric is discussed. For the ease of following, each of these analyses is allotted a subsection, and findings are discussed in view of the related literature.

When running FACETS analyses, it is customary to center all facets except one to establish a common origin, usually zero (Engelhard & Myford, 2003). If more than one facet is noncentered, then ambiguity may result since the frame of reference is not sufficiently constrained (Linacre, 1998). Following Eckes (2015), to establish the origin of the logit scale and make the model identifiable, the rater and criterion facets were centered, which means these facets were constrained to have a mean element measure of zero. The examinee facet was the only facet that was left non-centered.

It is of great importance to note that researchers are able to draw useful, diagnostically informative comparisons among the various facets only if the rating data show sufficient fit to the model (Engelhard & Myford, 2003), which is the topic of the next session. See section 4.3.1.1 for an explanation of the MFRM model.

### 4.4.1. Global model fit

As Rasch models are idealizations of empirical observations, empirical data will never fit a given Rasch model perfectly. The key issue is the *practical utility* of the model; that is, whether the data fits the model usefully or not, and, when misfit is detected, how much misfit there is and where it stems from (Eckes, 2009, p. 27; Eckes, 2015, p. 69).

One way to evaluate overall data-model fit is to look into the differences between responses that were observed and responses that were expected on the basis of the model. These differences between observed and expected responses are generally indicated as *standardized residuals*. Linacre (2008) states that satisfactory model fit is indicated when about 5% or less of standardized residuals $\geq 2$, and about 1% or less of standardized residuals $\geq 3$.

Considering the writing performance data in this study that was gathered by using the final draft of the new rubric , there was a total of 3250 valid responses (13 raters × 50 essays × 5 categories = 3250), which are used for estimation of model parameters. Totally, there were 100 unexpected responses. See Appendix 23 for the table presenting unexpected responses. Of these, 81 responses (or 2.49%) were associated with standardized residuals ≥ 2, and 19 responses (or 0.58%) were associated with standardized residuals ≥ 3. Following Linacre (2008), it can then be concluded that satisfactory model fit was achieved.

### 4.4.2. Variable map

A key feature of the results is a graphical display that illustrates the calibration of the facets involved in the assessment process. As mentioned above in the data analysis section, this graphical display is called a logit scale, a variable map, or the Wright map (Eckes, 2009, 2015). It is also called a vertical ruler (Brown & Edmonds, 2012). One can see how examinees' abilities varied, how severe or lenient raters were, and how difficult categories were on this common logit scale. All measures of the facets included in the assessment process are positioned vertically on the same latent dimension, with logits as measurement units. The logit measures represent the range of scores on a true interval scale as opposed to raw test scores where the distances between intervals may not be equal (Edmonds, 2012). Figure 4.3 demonstrates the variable map representing the calibrations of examinee proficiencies, rater severities, category difficulties, and four-level scale as raters used it to score examinee essays in this study.

As can be seen in Figure 4.3, the first column is the *logit (measurement) scale*. This scale is shown in logit scores where the mean is 0 and the range is -2.00 to +4.00 (in this case). On the logit scale a higher score equals a positive logit, that is, a higher measure. In the same way, a lower score equals a negative logit, that is, a lower measure.

The second column, which is labeled "Examinee", shows the estimates of the examinee proficiency parameter. Each star represents an examinee. As the plus sign before the examinee parameter $\theta_n$ in the equation above indicated, the examinee facet is positively oriented. Thereby, higher-scoring examinees appear at the top of the column, and lower-scoring examinees appear at the bottom. According to Figure 4.3, it is possible to rank the variation in examinee proficiency ranging from -1.00 to 3.00 on the logit scale. The examinees are extended along the measure, with the majority of them (forty-three)

above 0.00 logits, and the rest (seven) positioned at or below 0.00 logits. For instance, an examinee whose proficiency calculation was 3.00 logits on the logit scale is likely to get the highest raw score (four in this case) in all categories when s/he is assessed by an average-severity rater.

The third column labeled "Rater" displays rater severity; that is, compares the raters in terms of the level of severity or leniency each practiced when rating essays. Each of the 13 raters was assigned a number from 1 to 13 as R1 and so on. Unlike the examinee facet, the rater facet has a negative orientation, as the minus sign before the rater parameter $\alpha_j$ in the equation above pointed out. It indicates that the higher the rater measure the lower the raw score. Thus, more severe raters appear higher in the scale, and more lenient raters appear lower in the scale. Figure 4.3 shows that almost half of the raters (six) are situated above 0.00 logits, which means they tended to rate the student essays severely (R1, R8, R11, R4, R10, and R5, the last two being the most severe). Two of the raters were of average severity whose severity measures are 0.00 on the logit scale (R3 and R13). Finally, five of the raters are positioned below 0.00 logits, the more lenient end of the scale (R2, R6, R12, R7, and R9, the last one being the most lenient).

The fourth column labeled "Category" shows the variation in category difficulties; that is, it compares the five rubric categories in terms of their relative difficulties. As it is with the rater facet, the criteria facet is negatively oriented, which means categories situated higher in the logit scale were more difficult than those appearing lower. Hence, the higher the difficulty measure of a specific category, the more difficult it was for examinees to receive a high score on that category. As displayed in Figure 4.3, this column has the least variation of all, with all the categories clustered more or less around the mean 0.00 logits. Grammar and Vocabulary were more difficult than the other categories, while Content and Organization were similarly difficult. Mechanics was the easiest of all categories. In Phase 5 of this study the participants support the quantitative findings by stating that they find the category of Mechanics the easiest to score while they believe the categories of Grammar and Vocabulary were the most difficult to score for several reasons. The findings of Edmonds (2012) indicate similar results regarding the category of Mechanics, and for Edmonds these findings are not too surprising as most of the teacher-raters feel most confident in this category.

The last column presents the four-level scale for the five categories and how they relate to the logit scores; that is, distances between each raw scale as they were calibrated

onto the true interval logit scale. These demonstrate what any student writer at any proficiency level on the scale are likely to receive. Each horizontal dashed line indicates +0.5 score points, i.e., *category thresholds*, or *Rasch-half-score-point thresholds* (Eckes, 2015: 15). For example, on Figure 4.3 one can see that an examinee whose ability estimate was 3.00 logits on the logit scale is likely to receive almost all correct on their raw scores when s/he is assessed by an average-severity rater. One can also notice that not all four score ranges are the same. While the score ranges of 2, 3, and 4 are almost the same, the score range of 1 is narrower, which indicates that the entire range of 2-4 was being used in approximately equal intervals by the raters. The reason why the score level 1 was used more narrowly could be due to the essay selection criteria used in this phase of the study.

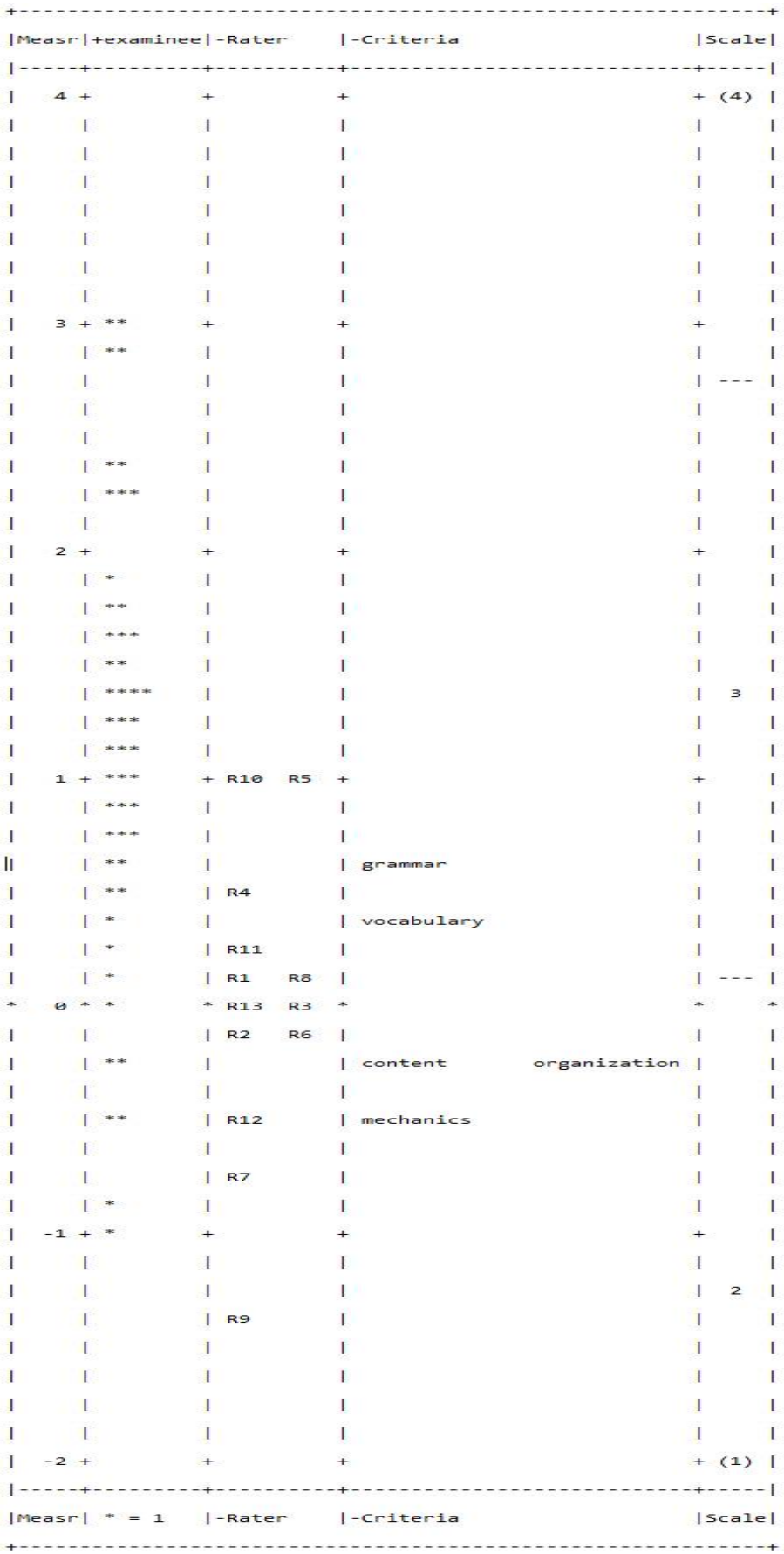


**Figure 4.3.** *Variable map for the new rubric*

As mentioned above in Section 3.5.3.2, fifty student essays which had five paragraphs and were not considered irrelevant were randomly selected by the researcher. Because the score level 1 is generally awarded to essays with missing paragraphs or paragraphs seriously lacking the criteria in the rubric, the raters in this study may not have felt the necessity to use this score level as much as the others.

In addition to the summary map provided in the variable map displayed in Figure 4.3, detailed measurement results for each facet are separately presented through fit statistics, which is the topic of the next session.

### 4.4.3. Fit statistics

In order for Rasch analysis to interpret examinee proficiency, rater severity, or category difficulty, the data need to fit the MFRM model. Thus, the summary fit statistics which FACETS provide for examinee fit, rater fit, and category fit initially are presented and discussed before the in-depth measurement results for each facet.

#### 4.4.3.1. *Summary MFRM results*

Table 4.16 shows the summary MFRM results for the writing performance data in this study. The upper half of the table gives the means and standard deviations of the examinee, rater, and category measures, along with the mean standard errors of the respective measures, the root mean-square measurement errors, and the adjusted (true) standard deviations. The lower half of the table displays the separation statistics, i.e., several group-level statistical indicators for the present three-facet sample data including ♯ Misfit, Homogeneity Index, Separation Ratio, Separation Index, Separation Reliability, and Chi-square (fixed).

As mentioned above in Section 4.4.1, the ♯ *Misfit* indicates the number of examinees, raters, or categories that "did not fit the general pattern of responses in the matrix, and can thus be classified as relatively misfitting…" (McNamara, 1996, p. 171). According to Brown and Edmonds (2012), various factors may cause misfitting responses such as examinees with language proficiency levels that are outside the possible range of scores, raters with inadequate training or attention to the task, or categories with poorly formulated descriptors. Out of fifty examinees, there were only two misfitting examinees, which means these two examinees did not fit the measurement model as their response patterns were unexpected.

**Table 4.16.** *Summary MFRM statistics of the writing performance data*

| **Statistic** | **Examinees** | **Raters** | **Categories** |
|---|---|---|---|
| Mean (*M*, measure) | 1.14 | .00 | .00 |
| Standard Deviation (*S.D.*, measure) | .97 | .66 | .51 |
| Standard Error (*S.E.*) | .21 | .10 | .06 |
| RMSE | .21 | .10 | .06 |
| Adj. (true) *S.D.* | .95 | .65 | .50 |
| ♯ Misfit | 2 | 0 | 0 |
| Homogeneity index (*Q*) | 1002.8** | 446.9** | 244.0** |
| *df* (degrees of freedom) | 49 | 12 | 4 |
| Separation ratio (*G*) | 4.59 | 6.20 | 7.72 |
| Separation (strata) index (*H*) | 6.46 | 8.59 | 10.63 |
| Separation reliability (*R*) | .95 | .97 | .98 |
| Chi-square (fixed) | $p = .00$ | $p = .00$ | $p = .00$ |

** $p < .01$

There were not any misfitting raters or categories in the writing performance data. As it was explained Phase 3 of the study (Section 3.4.3.7), FACETS output provides both population and sample versions of separation statistics. Because the writing performance data in this study did not comprise the whole population, sample statistics were reported for each facet.

To begin with, *RMSE* refers to *root mean-square standard measurement error* (Eckes, 2015). Brown and Edmond (2012) support that lower RMSE values indicate better data fit with the measurement model. Considering the relatively RSME values in this study ranging from .06 to .21, it can be concluded that the three facets, particularly the rubric categories, are fitting the model as desired.

*Homogeneity index (Q)* shows whether as a minimum of two elements within each facet have measures that differ in a statistically significant way. As indicated in Table 4.16, all three facets in the writing performance data in the present study had measures that were different in this respect; that is, at least two rubric categories did not share the same value of the difficulty parameter, after allowing for measurement error.

With reference to the *separation ratio (G)*, the value of 4.59 for the examinee facet pointed out that the variability of the examinee proficiency measures was four and a half times larger than the precision of those measures. The *G* value attained for the rater facet

indicated that the variability of the severity measures was more than six times larger than their precision. In comparison with the examinee and rater facets, the calculation of measures for elements of the category facet rested on a much greater number of observations (each difficulty measure was calculated based on 650 observations). These measures were then calculated with a specifically low error measurement error (i.e., *RMSE* = 0.06). Thereby, the category separation ratio obtained a value higher than the examinee or rater separation ratio (Eckes, 2015). Using the separation ratio, the separation index (the number of statistically different levels of examinee proficiency, rater severity, or category difficulty) can be calculated, which is the topic of the next paragraph and is discussed thoroughly there.

The *separation (strata)* or *index (H)* refers to the degree to which the examinees, raters, and categories spread out relative to their degree of precision (Linacre, 2008, p. 149). According to Brown and Edmonds (2012), the higher the separation index value, the more efficiently each facet is spreading its elements out. In the sample, the value of the examinee proficiency index was 6.46, indicating that among the 50 examinees included in the analysis, there were around six-and-a-half statistically different group of examinees. The new writing rubric, on the other hand, has four band levels to differentiate the examinee proficiency. Because the English proficiency examination serves as an exit examination to pass the BUU-SFL-IEP, and students are placed in four language proficiency levels during the academic year, a writing rubric with six band levels would be not only impractical but also unrealistic in the context of the current research. The separation index calculated for the rater facet was 8.59, suggesting that among the 13 raters included in the analysis there were nearly eight-and-a-half statistically different classes of rater severity. Lower might have been better; however, considering that the raters in this study were not given a training and norming session for the new writing rubric, and also this was a first time of rating, it is satisfactory (Edmonds, 2012, p. 111). More importantly, as emphasized by Brown and Edmonds (2012) and Eckes (2015), the MFRM approach puts the emphasis on the internal consistency among raters since it will make statistical modeling of rater characteristics possible. Besides, this approach considers variability in stable rater characteristics as the reality of life. Thus, as long as raters are internally consistent, such variability can be accounted for through either multiple rating by averaging of scores or the use of more advanced means of MFRM. Finally, the separation index for the category facet was 10.63, a value greater than the

number of categories included in the analysis, and also a value higher than the separation index values of the other two facets, which is not a negative characteristics at all according to Brown and Edmonds (2012), and which is caused by a large number of observations available for each element in this facet (Eckes, 2015), as discussed in the previous paragraph on separation ratio.

The next statistic, *separation reliability (R)* needs to be interpreted differently depending on the facet to be discussed (Eckes, 2015). For examinees, the examinee separation reliability indicates how different the examinee proficiency measures are. The high reliability for examinees in the current research (.95) shows that the examinees consistently differ from each other, which Brown and Edmonds (2012, p. 78) believes is the aim of a testing situation like the one described in this study. For raters the interpretation of the separation reliability differs from that of examinees (Eckes, 2015). Eckes (2015, p. 66) explains the difference explicitly with reference to standard approach to rater reliability. The standard approach to rater variability supports that when raters within a group practiced a highly similar degree of severity, rater separation reliability will be close to 0; thus, it aims for *low* rater separation reliability as this would signal that raters were approaching the actually "impossible" ideal of being interchangeable. In contrast, when raters within a group practiced a highly different degree of severity, rater separation reliability will be close to 1. This means that unlike interrater reliability, which in general terms is an index of how *similar* raters are in terms of their severity, rater separation reliability is an index of how *different* severity measures are. Therefore, as emphasized by Eckes (ibid.), these two kinds of reliability indices needs to be differentiated. In the present study, rater separation reliability was as high as .97, demonstrating a remarkable heterogeneity of severity indices. It is important to note that from the perspective of the standard approach to rater reliability this value would be equal to around .80 (interclass correlation coefficient). The ideal reliability coefficient for performance assessment is between the .70 to .79 ranges, which is considered sufficient for tests assessing written performance (Hughes, 1989, p. 39) and shows that raters were functioning almost interchangeably.

Finally, the separation reliability for the category facet was 0.98, which points out that the categories differed from each other in terms of difficulty to a very high degree and were consistently doing different things. When compared to the results of Phase 3 where the draft rubric was trialed (0.61), 0.98 is a much higher reliability indice.

According to Brown and Edmond (2012), this is certainly a desirable characteristics for the facet of category. Eckes (2015) agrees with their opinion by stating that when the set of criteria is designed to cover a wide range of performance features spread out across the underlying difficulty dimension (which was the case in the current study), high values of this statistic would be desirable.

The last statistic in Table 4.16 is the *chi-square (fixed)* values. They indicate that the chi-square statistics for the three facets in the study were significant ($p<.01$), which means that the examinees probably differed from each other in proficiency for reasons other than chance; and, that the raters and categories did, too.

In addition to the variable map and summary Rasch results, the MFRM modelling provides in-depth measurement results for each facet, which will be presented and discussed thoroughly in the following subsections.

#### 4.4.3.2. *Fit statistics for each facet*

Another important part of the output of Rasch-based analyses are mean-square *fit statistics* provided for each facet included in the measurement process. Two fit statistics are commonly used: *infit* (information weighted fit statistic) and *outfit* (outlier-sensitive unweighted fit statistic). As the names imply, outfit values include all the observations in the data set, and are hence sensitive to atypical outlying values, whereas infit values indicate the extent of score variability in a given data set which remains after the extreme values (outliers) have been removed (Davies et al., 1999). In other words, infit statistic is more sensitive to variability in the range of observations that are usually of most interest. For this reason, infit is generally deemed to be more important than out fit in estimating model fit (Bond & Fox, 2007; Davies et al., 1999; Edmonds, 2012; Linacre, 2002c, 2008; Myford & Wolfe, 2003).

Infit and outfit statistics close to 0 demonstrates that the scores for a specific examinee, rater, or category falls within the normal range (Davies et al., 1999; Linacre, 2002c, 2008; Myford & Wolfe, 2003). Fit values that are greater than 1 indicate more variation from the norm than expected and tend to *misfit* (underfit) the model as also mentioned above. Conversely, fit values lower than 1 refers to less variation than expected and tend to *overfit* the model. Misfit is considered to be more problematic than overfit (Myford and Wolfe, 2003).

As a general rule for fit statistics, Linacre (2002c, 2008) recommended 0.50 as a lower control limit and 1.50 as an upper control limit for the infit and outfit mean squares. Thus, mean square values within the range between 0.50 and 1.50 is considered to be useful fit. Values greater than 1.50 indicate significant misfit, and values lower than 0.50 show significant overfit. Other researchers suggested a narrower range: 0.75 and 1.30 (Bond & Fox, 2007; McNamara, 1996). According to Eckes (2009, 2015), the decision depends on the nature of the assessment purpose. For the purposes of this study, the range between 0.50 and 1.50 is adopted; however, as the reader will see in the following subsections, the three facets included in the writing sample data in the current research indicates largely useful fit even if a narrower range is considered to be appropriate.

#### 4.4.3.2.1. *Examinee fit statistics*

As demonstrated in the variable map (Figure 4.3), there was a variation in examinees' writing proficiency although the essays considered to be irrelevant or lack a few paragraphs were excluded. The variation in examinee proficiency ranged from -1.00 to 3.00 on the logit scale, which means examinees were extended along 4 logits. What is more, separation statistics (Table 4.16) indicated that the examinee separation (strata) index was 6.46, with an examinee separation reliability of .95; that is, examinees were well-differentiated in accordance with their level of writing proficiency. Table 4.17 provides an in-depth measurement of the examinees' writing proficiency.

**Table 4.17.** *Examinee fit statistics*

| Examinee | + Measure | Model *SE* | Infit *MnSq* | Outfit *MnSq* |
|---|---|---|---|---|
| 31 | -1.00 | .21 | .86 | .87 |
| 13 | -.91 | .21 | 1.06 | 1.05 |
| 23 | -.48 | .21 | .90 | .89 |
| 34 | -.44 | .21 | 1.29 | 1.27 |
| 2 | -.23 | .20 | .83 | .83 |
| 33 | -.23 | .20 | 1.05 | 1.03 |
| 30 | .02 | .20 | .99 | .98 |
| 11 | .14 | .20 | 1.53 | 1.50 |
| 7 | .26 | .20 | .72 | .72 |
| 18 | .38 | .20 | .97 | .96 |
| 35 | .46 | .20 | .90 | .89 |

**Table 4.17. (Continued)** *Examinee fit statistics*

| | | | | |
|---|---|---|---|---|
| 29 | .54 | .20 | .58 | .58 |
| 32 | .58 | .20 | 1.09 | 1,08 |
| 20 | .62 | .20 | 1.12 | 1.10 |
| 39 | .70 | .20 | .72 | .71 |
| 46 | .74 | .20 | 1.35 | 1.34 |
| 37 | .78 | .20 | 1.01 | 1.01 |
| 48 | .82 | .20 | .85 | .84 |
| 4 | .90 | .20 | 1.05 | 1.05 |
| 24 | .90 | .20 | 1.46 | 1.49 |
| 5 | .94 | .20 | 1.30 | 1.29 |
| 21 | .94 | .20 | .79 | .80 |
| 15 | 1.02 | .20 | .92 | .92 |
| 27 | 1.06 | .20 | .73 | .73 |
| 10 | 1.14 | .20 | 1.01 | 1.04 |
| 8 | 1.18 | .20 | 1.21 | 1.21 |
| 44 | 1.22 | .20 | .79 | .79 |
| 38 | 1.26 | .20 | .94 | .95 |
| 47 | 1.31 | .20 | 1.15 | 1.14 |
| 1 | 1.39 | .20 | .84 | .85 |
| 50 | 1.39 | .20 | .85 | .85 |
| 16 | 1.43 | .20 | .72 | .73 |
| 26 | 1.43 | .20 | 1.18 | 1.19 |
| 12 | 1.47 | .20 | .89 | .89 |
| 6 | 1.51 | .20 | 1.10 | 1.10 |
| 28 | 1.59 | .20 | .80 | .81 |
| 49 | 1.63 | .20 | 1.18 | 1.20 |
| 45 | 1.67 | .20 | 1.32 | 1.35 |
| 17 | 1.72 | .20 | .90 | .92 |
| 22 | 1.80 | .21 | .71 | .72 |
| 14 | 1.89 | .21 | .66 | .68 |
| 25 | 2.19 | .21 | 1.10 | 1.13 |
| 19 | 2.24 | .21 | .83 | .84 |
| 41 | 2.24 | .21 | 1.23 | 1.27 |
| 40 | 2.42 | .22 | 1.34 | 1.38 |
| 43 | 2.42 | .22 | 1.10 | 1.15 |
| 9 | 2.87 | .23 | .83 | .98 |

**Table 4.17.(Continued)** *Examinee fit statistics*

| | | | | |
|---|---|---|---|---|
| 36 | 2.92 | .23 | .90 | .85 |
| 42 | 2.98 | .23 | 1.39 | 1.59 |
| 3 | 3.03 | .24 | .70 | .71 |
| **Mean** | 1.14 | .21 | 1.00 | 1.01 |

+: examinee's proficiency positively oriented (high logit values means high proficiency)

From the left, the columns display examinee identification (each of which was numbered for easy retrieval and anonymity), the logit measures of the examinees' proficiency followed by standard error and infit and outfit mean square values. More specifically, Column 1 shows the order of examinees based on the measure of proficiency in Column 2. Column 2 demonstrates that the proficiency span between the lowest scoring examinee (Examinee 31) and the highest scoring examinee (Examinee 3) was 4.03 logits.

The differences in proficiency ranged from -1.00 to 3.03 logits (4 logits). Column 3 indicates that the standard error (*SE*) ranged from 0.20 to 0.24. Column 4 and Column 5 show the infit and outfit mean square values; that is, they present the extent to which the data representing each examinee fell within the accepted boundaries (between 0.50 and 1.50). As also mentioned above in Table 4.17, there are only two examinees that fell out of the useful fit. Table 4.17 shows that Examinee 11 (1.53) and Examinee 42 (1.59) had fit statistics that were greater than 1.50, which could be said to be misfitting and not performing as the model predicted. The performance of these examinees needs to be evaluated further to find out what might be the reason behind their non-predictive behavior. It is important to note that misfitting or overfitting does not necessarily mean that the examinees, raters, or categories are problematic; it just means that the examinees did not perform according to the model, or they performed too well within the model (Edmonds, 2012). The remainder of the examinees fell between the reasonable boundaries (between 0.50 and 1.50), and the majority of them (44) did so even the narrower range (between 0.75 and 1.30) was adopted.

Overall, based on the examinee proficiency measurement report it can be concluded that the examinees' scores were reasonably well distributed and even though 2 of their performances were outside the model's expectations, the rest were well within the expectations. The reliability index on the examinee measurement was very high at 0.95 which pinpoints that it is very likely that the examinees would perform in a similar way if another test that aimed at measuring the same construct was administered (Bond & Fox,

2007: 40). One drawback of the analysis could be the number of the examinees which was only 50, yet the analysis did indicate that there was a high person reliability in which some of the examinees scored higher and some lower in a consistent manner.

#### 4.4.3.2.2. *Rater fit statistics*

The variable map displayed in Figure 4.3 showed clearly that the raters in this study varied substantially in their measures of severity, which was also supported by the separation index (8.59) and the separation reliability (.97) statistics displayed in Table 4.16. However, as mentioned above and as the following fit statistics indicate, there are not any misfitting raters, either. Table 4.18 presents the detailed rater measurement report in a similar way to the examinee report.

**Table 4.18.** *Rater fit statistics*

| **Rater** | **-Measure** | **Model *SE*** | **Infit *MnSq*** | **Outfit *MnSq*** |
|---|---|---|---|---|
| 9 | -1.40 | .11 | 1.14 | 1.18 |
| 7 | -.80 | .11 | 1.41 | 1.46 |
| 12 | -.46 | .11 | .92 | .92 |
| 6 | -.17 | .10 | 1.20 | 1.19 |
| 2 | -.16 | .10 | 1.07 | 1.06 |
| 13 | .04 | .10 | 1.32 | 1.30 |
| 3 | .05 | .10 | 1.02 | 1.03 |
| 8 | .07 | .10 | .84 | .83 |
| 1 | .15 | .10 | 1.09 | 1.10 |
| 11 | .19 | .10 | .61 | .61 |
| 4 | .48 | .10 | .80 | .80 |
| 5 | .98 | .10 | .68 | .67 |
| 10 | 1.04 | .10 | .91 | .90 |
| **Mean** | .00 | .10 | 1.00 | 1.01 |

Table 4.18 presents the detailed rater measurement report in a similar way to the examinee report. From the left, the columns demonstrate rater identification in numbers, the logit measures of rater severity followed by standard error and fit statistics. Column 1 displays the rank of the raters based on the measure of severity in Column 2. Column 2 pinpoints that the differences in severity ranged from -1.40 to 1.04, which means the

severity range between the most severe raters (Rater 5 and Rater 10) and the most lenient rater (Rater 9) was -2.44 logits. The standard error range was 0.10 to 0.11. Column 4 and Column 5 present the infit and outfit mean square values; that is, they display the extent to which the data representing each rater fell within the reasonable boundaries (between 0.50 and 1.50). These statistics indicate that all 13 raters fell between the reasonable boundaries (between 0.50 and 1.50), and the majority of them (11) did so even the narrower range (between 0.75 and 1.30) was adopted, which means their rating behavior was independent. As Edmonds (2012: 111) puts forward, lower indices might have been preferable; however, considering that the raters in this study were not given a training and norming session for the new writing rubric, and also this was a first time of rating, it is satisfactory. Moreover, as also mentioned above, the MFRM approach considers variability in stable rater characteristics a fact of life (McNamara, 1996) and highlights the internal consistency among raters since it will make statistical modeling of rater characteristics possible (Brown & Edmonds, 2012; Eckes, 2015) and also what causes variability in rater severity still remains a mystery. As Eckes (2009) emphasizes, the research into the stability and change in rater severity and the personal and situational factors that affect rater severity is considerably sparse. Among some of the factors he lists are teaching and rating experience, demographic and/or personal characteristics, workload, and assessment aim. Considering the teacher-raters who participated in Phase 3 (Trial and refinement of draft rubric) and this phase (Phase 4) of the current research, teaching-rating experience does not seem to be one of the factors of the variability in rater severity. All five teacher-raters in Phase 3 had more than 10 years of professional experience in the ELT field, and their rating behavior was independent with a reliability of 0.95 and a separation index of 6.47. Even higher indices were observed for the thirteen teacher-raters in Phase 4 where both the most severe and the most lenient teacher-raters had professional experience ranging from 17 to 20 years in the ELT field. Regardless of the factors involved, the variability in rater severity could be observed in relatively recent research that used MFRM to test the validity of L2 writing rubrics in spite of the efforts to reach rater agreements with extensive training and norming sessions (e.g., Becker, 2018; Eckes, 2009, 2015; Edmonds, 2012; Hattingh, 2009; Knoch, 2007). Thus, rater training needs to aim at increasing raters' internal consistency and reducing extreme levels of rater severity or leniency rather than trying to reach an inter-rater agreement as Eckes explains in his following words:

> "Rater training usually does not succeed in reducing between-rater severity differences to an acceptably low level. Therefore, in most situations, adopting the standard view that rater training needs to achieve maximal between-rater similarity, and eagerly pursuing this objective in rater training sessions, is extremely likely to end up in frustration of those in charge of the training. The constructive alternative to striving after fictitious rater homogeneity is to accept rater heterogeneity within reasonable bounds and to adopt a suitable psychometric modeling approach. Many-facet Rasch measurement provides the tools to probe deeply into the complexities of rater behavior, and to use the insights gained for the purposes of making performance assessments as fair as possible" (Eckes, 2015, p. 73).

Overall, all 18 teacher-raters, who were included in Phase 3 (5 raters) and this phase of the current research (13 raters), were internally consistent despite the lack of a training and a norming session and their unfamiliarity with the new rubric. There were not any misfitting raters who were outside of the model's expectations, and the vast majority of the examinees responses and all of the raters' performances were well within the model's expectations.

#### 4.4.3.2.3. *Category fit statistics*

The difficulty measurement report of the five categories, i.e., Content, Organization, Grammar, Vocabulary, and Mechanics, is demonstrated in Table 4.19 in the same way as the examinee proficiency and rater severity reports.

**Table 4.19**. *Category fit statistics*

| **Categories** | **-Measure** | **Model *SE*** | **Infit *MnSq*** | **Outfit *MnSq*** |
|---|---|---|---|---|
| Mechanics | -.51 | .07 | .95 | .99 |
| Organization | -.29 | .07 | 1.24 | 1.25 |
| Content | -.27 | .07 | 1.13 | 1.11 |
| Vocabulary | .39 | .06 | .84 | .85 |
| Grammar | .68 | .06 | .82 | .82 |
| **Mean** | 00 | .06 | 1.00 | 1.01 |

From the left, the columns present category identification, variance in category difficulty, error, and fit statistics. Column 1 shows the order of categories based on the measure of difficulty in Column 2. Column 2 pinpoints that the differences in category difficulty ranged from – 0.51 to 0.68, which means the difficulty range between the most leniently scored category (Mechanics) and the most severely scored category (Grammar)

was 1.19. When compared with the logit measure ranges of the examinee (4.03) and the rater (-.2.44) facets, the logit measure range of the category facet was smaller indicating that the difficulty measures of categories did not vary as much as the examinee proficiency and rater severity measures. Column 3 shows that the standard error range from 0.06 to 0.07. Column 4 and Column 5 indicate the infit and outfit mean square values; that is, they display the extent to which the data representing each category fell within the reasonable boundaries (between 0.50 and 1.50). These statistics indicate that all 5 categories fell between the reasonable boundaries (between 0.50 and 1.50) with no misfitting or overfitting data. Moreover, all of them did so even the narrower quality control limits (between 0.75 and 1.30) were adopted. This finding is in compliance with the assumption of psychometric unidimensionality of the set of categories in this study (McNamara, 1996), which means all five categories seemed to relate to the same dimension i.e., examinee writing proficiency, as assumed by MFRM (Eckes, 2015). Even though the difference in category difficulty was small, the reliability 0.98 indicated that the categories were performing consistently independent of each other. All in all, the five categories did not vary greatly in terms of difficulty, and none of them displayed any significant misfit and overfit, supporting that the multi-trait rubric behaved as the model might expect.

#### 4.4.3.3. *Band level (rating scale) analysis*

In addition to category difficulty, the quality of the four-level scale used by the raters to evaluate examinee proficiency is of great importance. A variety of statistical indices are utilized to investigate rating scale validation, i.e. whether the four band levels in the new rubric performed as intended. Based on Linacre (2004b) and Bond and Fox (2007), the following three indices are presented by Eckes (2009, 2015) as the indicators of rating scale effectiveness: the *average measure* of each band level, the *mean-square outfit statistic* calculated for each band level, and the *ordering of Rasch-Andrich thresholds*.

The first indicator, the average measure, refers to the average of the examinee proficiency measures modeled to produce the observations in a given band level. It is required that the average measures progress monotonically, which means the higher the band level, the larger the average measure. If this prerequisite is met, it can be concluded that higher ratings equal “more” of the variable that is measured (Eckes, 2009, p. 26).

The second indicator of rating scale effectiveness is the mean square outfit statistic, which is the examinee proficiency measure the model estimates for a given level if the data were to fit the model. Generally, this statistic should not be above 2.0 (Eckes, 2009, p. 26).

The final indicator of rating scale effectiveness is the ordering of the Rasch-Andrich thresholds. As it is with the average measure, the requirement is that these thresholds should increase monotonically with each level (Eckes, 2009, p. 26).
Table 4.20 presents the results with regard to these indices.

**Table 4.20.** *Overall category fit statistics*

| **Bend Levels** | **Counts** | **Average Measure** | **Outfit Statistic** | **Rasch-Andrich thresholds** | ***SE*** |
|---|---|---|---|---|---|
| 1 | 74 | -.28 | 1.3 | | |
| 2 | 886 | .20 | 1.0 | -.70 | .12 |
| 3 | 1587 | 1.24 | 1.0 | .16 | .05 |
| 4 | 703 | 2.24 | .9 | 2.54 | .05 |

Column 1 shows each band level in the writing rubric, i.e., 1, 2, 3, and 4. Column 2 displays the counts used to estimate the indices. Column 3 indicates the average measure for each band level, Column 4 the mean square outfit statistic again for each band level, Column 5 the Rasch-Andrich thresholds, and Column 6 the standard error. As demonstrated in the table, the average measures of examinee proficiency advanced with each band level. Likewise, values of the mean square outfit statistic were almost equal, or very close, to the desirable value of 1.0. Lastly, there was a monotonic advancement of band level thresholds from -2.70 logits (i.e. the threshold between band levels 1 and 2) to 2.54 logits (i.e. the threshold between band levels 2 and 4). All in all, these indices strongly support that the four band levels of the new rubric were ordered appropriately and functioning as desired.

In addition to the statistics explained above, MFRM provides a graphical illustration, which is called the *probability curves,* for rating scale validation. According to Brown and Edmonds (2012), these curves are beneficial as they graphically demonstrate the degree to which the band levels are distinct or overlapping. Both Eckes (2009: 26) and Brown and Edmonds (2012, p. 80) support that the best would be probability curves that have a "distinct hill-like appearance" with one curve for each

band level and some overlap between hills but not too much. Figure 4.4 demonstrates the probability curves for the four-level scale utilized by the raters to rate the examinees on the five-category rubric.

The horizontal axis gives the examinee proficiency scale and the vertical axis the probability of being rated in each level. As the figure displays, there is a distinct hill for each level with little overlap, and the level thresholds are properly ordered from left to right. It can be concluded that the probability curves the statistical indices in that the levels in the new rubric were properly ordered and working as expected.


**Figure 4.4.** *Probability curves for the new rubric*

### 4.4.4. Conclusion of Phase 4

In conclusion, fit statistics that are in the form of summary results and fit statistics for each facet accompanied by graphical displays in the MFRM analysis show that the scores the teacher-raters awarded to examinees by using the new rubric to assess their writing proficiency proved to be reliable and valid. When compared to the MFRM analysis carried out in Phase 3 of the study where the draft rubric was trialed and refined,

the findings of the MFRM analysis to investigate the new rubric in this phase of the study are even more satisfactory; particularly the ones related to the facet of category difficulty, which could be related to the revisions made in the rubric based on the feedback received from the participants in the piloting phase of the study.

### 4.5. Phase 5: Exploration of raters' perspectives on the new rubric

In this section, results of the fifth and the last phase of this research are presented and discussed based on the three areas aimed to be explored by the open-ended questionnaire; that is, the strengths and weaknesses of the new writing rubric designed to be used for the assessment of writing performance in the proficiency examination, the efficacy of the new rubric in fair assessment of students' written work, and the teacher-raters' confidence level in using this new rubric, as perceived by the13 experienced teacher-raters who were EFL instructors at BUU-SFL-IEP and who were participants in the third phase of the study, as well. Each of these areas is allocated a subsection for the ease of following, and findings are discussed in comparison with the first phase of the study which explored the perspectives of the participants on the rubric currently used in the assessment of writing performance, third phase of the study where the draft rubric was piloted, the fourth phase of the study where psychometric analysis of the new rubric was carried out, and, the last but not the least, the relevant literature.

#### 4.5.1. Strengths of the new rubric

According to the 13 participants of the study who took part in the proficiency examination at BUU-SFL-IEP as raters, there were seven advantages of the new writing rubric:

- Practicality,
- Comprehensiveness of descriptors,
- Clarity of descriptors,
- Categorization (of writing constructs that need to be assessed),
- Provision of diagnostic feedback for students and teachers alike,
- Allocation of one score for each band level (rather than a range of scores), and
- Learner-centeredness.

See Table 4.21 below for the total number and percentages of participants perceiving an aspect of the writing rubric as an asset.

**Table 4.21.** *Strengths of the new writing rubric designed to be used in the proficiency examination (N=13)*

| | ***F*** | ***%*** |
|---|---|---|
| **Practicality** | 9 | 69 |
| **Comprehensiveness of descriptors** | 8 | 62 |
| **Categorization of writing constructs** | 7 | 54 |
| **Clarity of descriptors** | 7 | 54 |
| **Provision of feedback for students and teachers** | 3 | 23 |
| **Allocation of one score for each band level** | 2 | 15 |
| **Learner-centeredness** | 1 | 8 |

As demonstrated in Table 4.21, out of 13 participants, 9 of them stated *practicality* was the number one strength of the new writing rubric, which wass followed by *comprehensiveness of descriptors* indicated by 8 participants, *clarity of descriptors* and *categorization* (of writing constructs that need to be assessed) both mentioned by 7 participants, *provision of feedback for students and teachers*, stated by only three of the participants in the current research. 2 participants referred to *allocation of one score for each band level* and 1 to *learner- centeredness*.

#### 4.5.1.1. *Practicality*

The majority of the participants (that is 9 out of 13) stated *practicality* as the first strength of the new rubric. Practicality was also mentioned as the number one quality of the existing rubric in Phase 1 (Exploration of raters' perspectives on the current writing rubric) of the study by half of the participants in that phase. However, almost all of the participants in Phase 1 interrogated that the writing rubric currently used in the assessment of writing performance provided practicality at the expense of fairness, which jeopardized objectivity and scoring validity of such a high-stakes test serves as an exit exam at the end of the year. All 9 participants who found the new rubric practical, on the other hand, emphasized that the new rubric was practical without risking objectivity. Hence, it could contribute to more valid and reliable assessment of writing proficiency at BUU-SFL-IEP, as reflected in the following excerpts by Participants 5 and 12 (hereafter *P*):

> 87. It provides clear and easy-to-follow guidelines for the rater, and this helps avoid subjectivity. In addition, comprehensive descriptors are provided so that the rater can easily

justify the score assigned for each student. In short, it is clear-cut, useful, and practical by avoiding ambiguity, which makes it powerful in terms of validity as well as reliability (*P5*).

88. … Apart from a few details, it's easy-to-use and easy-to-understand rubric making me feel more confident as a rater *(P12)*.

Overall, 60 % of the participants referred to practicality as the first strength of the new rubric. Practicality is frequently stated as the third consideration in test design together with validity and reliability since a test cannot be utilized in a specific context unless it is not practical to administer in that context, however reliable and valid it may be (Davies et al., 1999). Nonetheless, this should not mean that validity and reliability could be risked for the sake of practicality, particularly, in a high-stakes test with a fail or pass result that can have vital backwash on students' educational lives, as also highlighted by Becker (2011) whose research looked into the role of writing and the types of rubrics used at the IEPs of various US universities.

#### 4.5.1.2. *Comprehensiveness of descriptors*

After practicality, the second strength of the new rubric participants referred to was *comprehensiveness of descriptors*. Out of 13 participants 8 participants stated the detailed descriptors as one of the assets of the new rubric. In Phase 1 (Exploration of raters' perspectives on the current writing rubric) of this study, the wording of descriptors was mentioned as the second weakness of the rubric currently used for the assessment of writing proficiency in the exit examination due to the narrowness of the descriptors. The new rubric, on the other hand, was found to be detailed but not verbose by the participants in this phase as the excerpts below indicate:

89. The descriptions for the numeric values under each criterion guided me to assess writings more carefully. The rubrics lacking a comprehensive description usually leads raters to interpret them subjectively. However, a comprehensive description helps them to make assessments more fairly and independently from their own prejudices towards the essay as a whole (*P1*).

90. The new writing rubric is quite detailed and clear. Especially the top score -4- in every category makes the expectation clear for each of them (*P8*).

Knoch (2011, p. 95) labels such criteria as "defined" rubrics, i.e., rubrics with detailed descriptions with concrete or objective formulation, and she considers them to be the most beneficial ones. However, creating descriptors is easier said than done because while writing a description for each criterion and each performance level is one

of the most important aspects of rubric design, it is also the hardest (Hamp-Lyons, 2018). As an option, an ongoing rubric analysis is highly recommended for fine-tuning the descriptors in a rubric, particularly in contexts that use high-stakes performance assessment, as it is the case in the context of the current research (Janssen, Meier, & Trace, 2015). This way a bridge between teacher-raters' feedback to formal assessment tools can be built, which Hamp-Lyons (2018) believes to be really valuable in reliable assessment of writing performance.

#### 4.5.1.3. *Categorization*

Following *practicality* and *comprehensiveness of descriptors*, two other aspects of the new rubric that participants stated to be advantageous were *categorization* (of writing constructs that need to be assessed) and *the clarity of descriptors*, both of which were stated by more than half of the participants in the study (7 out of 13).

*Categorization of writing constructs* is the determinant aspect of analytic or multi-trait rubrics, where a separate score is awarded for each number of features of a task, as opposed to holistic rubrics where a stretch of written discourse is assessed impressionistically based on its overall properties (Davies et al., 1999). In Phase 1 (Exploration of raters' perspectives on the current writing rubric) of this study, the existence of categories to assess a variety of writing constructs was again reported as the strength of the rubric currently used in the assessment of writing performance in the proficiency examination of the BUU-SFL-IEP. However, the number of categories was highlighted as the number one drawback of the rubric by the majority of the participants in Phase 1. In other words, for participants categorization of writing constructs was deemed to be advantageous, but the lack of sufficient categories to fully assess writing proficiency was a major disadvantage that could affect scoring validity negatively. Based on the feedback received from the participants in Phase 1 of the study, the construct of mechanics which was missing was included as a separate category in the new rubric. The constructs of argumentation and cohesion-coherence were added into the already existing categories of Content and Organization successively. It seemed that the participants in Phase 5 of the study found these modification useful, as reflected in the following excerpts by Participant 1 and Participant 9:

> 91. Analytic scoring makes me feel more confident as a rater. This alternative rubric enabled me to gain a different perspective as some of the essays had a coherent writing system despite

lacking a variety in vocabulary, or some were almost grammatically flawless without a concrete base on which the essays could be structured, though. This might contribute scores to be more fairly distributed because teachers tend to get an overall sense of the writings and consider them either good or bad depending on their impression (*P1*).

92. The new rubric is quite comprehensive. It directed me very well while evaluating the papers. Rubric contained almost all the aspects I thought were missing in the rubric we use at present (*P9*).

As the excerpts 91 and 92 demonstrate, categorization of writing constructs was considered to be an asset of the new rubric by 53% of the participants in this phase of the study. Many prominent authors specialized in performance assessment of writing list the advantages of analytic/multi-trait scoring as more exact reporting of written or oral skills development and greater reliability as each test taker is given a number of scores for each category in the rubric (Brown, 2012, p. 35; Davies et al., 1999, p. 7; Hyland, 2004, p. 230; Knoch, 2011, p. 83; Weigle, 2002, p. 121). Among them is Hamp-Lyons (2016a, 2016b, 2018) who challenges the value of holistic scoring by reminding her readers in different sources that not only is writing a complex and multi-faceted activity, but also the assessment of writing is also complex and multi-faceted. Thus, she supports fervently that analytic/multi-faceted scoring is the best method for assessing writing (Hamp-Lyons, 2018: 63). However, as also mentioned by all the above-mentioned theoreticians and researchers, contextual requirements need to be taken into consideration when making decisions that can have an important backwash on students' educational lives.

#### 4.5.1.4. *Clarity of descriptors*

The wording of descriptors was listed as the second drawback of the writing rubric currently used in the proficiency exam because of the vagueness of descriptors in Phase 1 of the study where raters' perspectives into the rubric in use were explored. Nonetheless, in Phase 5 of the research, together with *categorization*, *clarity of descriptors* was also mentioned by 54% of the participants as one of the assets of the new rubric. See the following excerpts by Participant 2 and Participant 7:

93. I think the most important strength of the new rubric is that the descriptors are easy to understand, clear, and straightforward. Band descriptors are clearly distinguishable for each category which makes it easy for the rater to score each criterion in limited time (*P2*).

94. The strengths of the new writing rubric design are the ease of understanding, straightforward application, and improved visibility. The new rubric was easy to understand. I was able to read through the rankings from each section and comprehend the expectation

> for the four grade levels. It was very easy to read through the rubric once and know what to look for when reading the essays… (*P7*).

The importance of clarity of descriptors for raters were emphasized in both excerpts. As also explained under the title of *comprehensiveness of descriptors* above, a solid and impartial formulation style needs to be adopted for the ease of understanding of and using a rubric (Knoch, 2011). Because it is not easy to achieve at one shot, an ongoing rubric analysis including the teacher-raters in the process would be rewarding to not only increase scoring validity but also bridge the gap between teacher-raters' feedback and formal assessment tools (Janssen, Meier, & Trace, 2015; Hamp-Lyons, 2018).

*Provision of feedback for students and teachers* was another strength of the new rubric stated by only three participants in the study. While the main purpose of a proficiency test is to measure how much of a language someone has learned rather than identify test takers' strengths and weaknesses, standardized tests such as the TOEFL or Cambridge Exams tend to make a washback effect on instruction (Davies et al., 1999, p. 154). Davies et al. righteously support that this achievement-proficiency dynamic accurately leads to new proficiency tests being designed. Based on this proposition, it would not be wrong to state that in contexts where high-stakes performance assessment of writing proficiency is used, it would be very useful to provide feedback to students (who ask for it at the least) so that they can find out their strengths and weaknesses and go on their language development under the light of this feedback. Institutions can also make great use of this feedback in order to make appropriate modifications on writing instruction. Participant 4 emphasizes the importance of such feedback for not only students but also institutions in the following excerpt:

> 95. Since the new rubric is analytical, it can give the assessors and the institution further feedback about the students' writing performance, which can guide them while reflecting on their practice in terms of teaching writing. This means this rubric can serve two functions; first, as a tool to use for assessment and second, as a guide to determine our students' strengths and weaknesses and common mistakes they make and how effective our teaching is (*P4*).

In addition to provision of feedback for students and teachers, *allocation of one score to each band level* and *learner-centeredness* were the other strengths of the new rubric, the first of which was stated by solely two participants and the second by only one participant. When the relevant literature is reviewed, it is possible to see that one score (rather than a range of scores) is allocated for each band level in the rubrics used and

displayed in the recent research and the seminal work of theorists in the literature of performance-based assessment of writing (e.g., Banarjee et al., 2015; Becker, 2018; Brown, 2012; Janssen et al., 2015; Knoch, 2007; Shohamy et al., 1992). In compliance with the related literature, one fifth of the participants indicated their dissatisfaction with having to choose a score within a range of scores for the same band level with the same descriptors (See section 4.1.2.3) in Phase 1 (Exploration of raters' perspectives on the current writing rubric). This feedback was taken into consideration, and one score was allotted for each band level in the new rubric. Finally, one participant stated that the new rubric was learner-centered because the highest attainable level was set "Meets expectations" instead of "Exceeds expectations", which was refined based on the recommendations by the participants in Phase 3 (Trial and refinement of the new rubric) of this study.

To sum up, participants listed seven strengths of the new rubric in Phase 5 of the study in contrast with the results of Phase 1 where the weaknesses of the rubric currently used in the assessment of writing performance in the proficiency exam at BUU-SFL-IEP outweighed the strengths drastically. The basic role of rubrics in performance-based assessment of writing proficiency is explained as their assumed assistance in increasing objectivity, reliability, and validity in scoring (Brown, 2012, p. 34; Crusan, 2015, p. 1). Results show that participants believed that the new rubric fulfilled its function to a great extent.

### 4.5.2. Weaknesses of the new rubric

As opposed to 7 assets of the new rubric as perceived by the participants, only 4 shortcomings were stated in Phase 5 of the study, the majority of which were stated by a very few participants unlike Phase 1 of the study where the number of weaknesses double the strengths with the majority of the participants stating their dissatisfaction. The four drawbacks are listed as follows:

- Difficulty of grading the categories of content and organization (in comparison with the other categories),
- Equal weighting,
- Vague descriptors, and
- Lack of coherence as a separate category.

See Table 4.22 below for the total number and percentages of participants perceiving an aspect of the writing rubric as a weakness.

**Table 4.22**. *Weaknesses of the new writing rubric designed to be used in the proficiency examination (N=13)*

| | ***F*** | ***%*** |
|---|---|---|
| **Difficulty of assessing Content and Organization** | 5 | 38 |
| **Equal weighting** | 3 | 23 |
| **Vague descriptors** | 3 | 23 |
| **Lack of a separate category for coherence** | 2 | 15 |

Table 4.22 shows that 5 participants out of 13 stated the difficulty of assessing the categories of Content and Organization as the first weakness of the new rubric while equal weighting and vague descriptors were listed as the next two drawbacks both of which were mentioned by 23% of the participants. Finally, one participant referred to the *lack of coherence as a distinct category*. Below are the descriptions and explanations of each weakness as perceived by the participants.

#### 4.5.2.1. *Difficulty of assessing the categories of Content and Organization*

5 out of 13 participants stated the *difficulty of assessing the categories of content and organization* compared to other categories i.e., grammar, vocabulary, and mechanics. They have similar concerns as demonstrated in the excerpts by Participant 6 and Participant 13:

> 96. Not a huge problem maybe, but it was difficult to mark the papers which did not have a counterargument (content) or title (organization) although the development of ideas and paragraphs were not so bad (*P6*).
>
> 97. While assessing the essays, I had difficulties in the categories of content and organization. Some essays can be awarded 4 points for the category of content in terms of development of ideas but with poor counterargument or refutation, which means the essay does not "Meet the expectations", right? I couldn't be sure (*P13*).

The excerpts indicate that teacher-raters had difficulties in how many points to deduce when one of the criteria in the categories of Content or Organization was missing in students' essays. Therefore, it can be concluded that the issue was not about the use of the rubric but reaching a consensus on how to score such essays so that consistency was not compromised. Two elements are highlighted in the relevant literature so that the

reliability of rating could be increased: "holding *training* and *norming sessions* to make sure that the rubric is used properly and consistently by the raters (Davies, 1999, p. 53; Hamp-Lyons, 2007, p. 1; Weigle, 2002, p. 108) and *rating expertise* gained through experience as a rater over time". A training and norming session was not held with the raters for the purposes of the current research in order not to affect their decision-making processes in an undesirable way and to to find out the extent to which the new rubric functions effectively on its own. Instead, raters were provided with an assessor guide that gave in-depth information on the criteria of each category (See Appendix T). However, before the actual use of the new rubric for the assessment of written performance at BUU-SFL-IEP, improved training and norming sessions could be carried out "in order to increase the reliability coefficients of each category, especially of the one(s) that are rendered problematic by the raters", as recommended by Hamp-Lyons (1990, p. 70; 2007, p. 1). During the session, instructors might be asked for their opinions regarding each category (e.g., the number of categories, the number of band levels in each category, wording of the descriptors), and changes that are agreed upon could be made accordingly, considering goals and objectives of the writing course. Finally, general ground rules could be compiled under the title of guidelines. and each rater is asked to follow the guidelines regarding the decisions made collectively.

Regarding rating expertise, if teachers are given the opportunity to specialize in performance based assessment of either writing or speaking, they can measure learners' performance more consistently in each language skill. Therefore, it could be concluded that the quality of the rating process might be increased by holding training and norming sessions regularly and through rating expertise gained over time.

#### 4.5.2.2. *Equal weighting*

According to three participants of this phase of the study, the second weakness of the new rubric is *equal weighting*. See the excerpts below by Participants 11 and 13:

> 98. The rubric values accuracy over content, which is usually what is expected in ESL writing rubrics. However, categories such as content and organization assess the learner's critical thinking capacities which prove to be important in both L1 and L2. Therefore, a balance between the two may yield more interesting results, as it would assess the learner's overall writing skills and be more learner-centered (*P11*).
>
> 99. While assessing the essays, I kept questioning if it is necessary to grade Mechanics category with 4 points. As these students are at university level, should we give students

> points because they can capitalize and punctuate correctly? Maybe 2 points is enough for Mechanics and 5 points can be given to Content and Organization to increase their weighting (*P13*).

However, in Phase 1 of the study, uneven weighting was stated to be a drawback of the rubric that is used at present, and the majority of the participants in that phase indicated that content is overrated despite its importance in writing. Hence, equal weighting of the categories was considered as a possible solution because of the significance of all writing constructs for L2 writing quality, and the draft rubric had equal weighting for each of the categories. In Phase 3 of the study where the draft rubric was piloted, three out of five participants expressed their satisfaction with each category having equal weighting.

There is an ongoing discussion on the weighting of categories for meaningful and valid assessment of test takers' writing abilities (East & Cushing, 2016). When the views of the participants in Phase 1, Phase 3, and Phase 5 of the study are reviewed, it would not be wrong to state that the debate on weighting in the literature of performance-based assessment of L2 writing is slightly echoed also in the context of BUU-SFL-IEP; however, the majority of the participants in all three phases of the study are in favor of equal weighting.

#### 4.5.2.3. *Vague descriptors*

Following *equal weighting*, *vague descriptors* were stated to be the third weakness of the new rubric as perceived by 3 participants as opposed to the majority of the participants who think the clarity of descriptors is an asset of the new rubric. According to Participant 6:

> 100. The wording of some statements in the category of content is vague and a little confusing. For example, the expressions in levels 3 and 2 are so close that it is difficult to decide on which one to choose: "ideas could be more fully developed as some evidence may be lacking" = "the main ideas may not be fully supported by the evidence given" or "some extraneous material is present" = "essay is somewhat off the topic". Both mean irrelevant to me (*P6*).

As opposed to Participant 6 who stated to have difficulties in differentiating between the band levels 3 and 2, Participant 1 said that he had a difficulty in awarding the scores in the highest (4) and the lowest bands (1):

> 101. I felt like it was more challenging for me to assign scores from both end. In other words, I was more comfortable scoring the essays 3 or 4 rather than 1 or 4. Particularly, scoring the

> essays 1 was quite difficult for me. I believe that this difficulty arouse due to the adjectival phrases in the descriptions such as "inadequate effort". Having options within a small range might also have contributed to this belief of mine as I started calculating what the scores would mean in a 0-100 range (*P1*).

It is important to note that it is inevitable to use adjectives and adverbs on difference in a rubric for assessing wide bands of ability even though raters may have some hesitance in using them (Brown, 2012, p. 23). As mentioned several times within the scope of this research, in order not to affect the decision-making processes of the raters and to find out to what extent the new rubric functions effectively on its own, participants were not given a training and norming session but an assessor' guide that explains what is expected from the test takers in each category. The issues Participants 1 and 6 highlighted such as the difference between the band levels 3 and 2 or the adjectival phrases in the rubric may again be clarified in a training and norming session with samples exemplifying each case. Another option is to refine the wording of descriptors as part of ongoing rubric analysis so that they could be more concrete and more objective for the raters (Janssen, Meier, & Trace, 2015; Knoch, 2011).

Finally, two participants referred to *lack of coherence as a separate category* because the flow ideas is a very important aspect of L2 writing ability, too. Participants' high satisfaction level of categories in the new rubric is particularly important because determining the categories of a rubric is stated to be the most important decision in the rubric development process (Knoch, 2011, p. 92; Weigle, 2002, p. 41). Apart from the two participants, the rest of the participants did not mention any dissatisfaction with the number or scope of categories in the new rubric, which was determined according to *the contextual needs of BUU-SFL-IEP* including goals and objectives of the writing course and teacher-raters' expectations from a writing rubric, *the related literature*, and the last but not the least, *the opinions of experts in performance-based assessment of writing*.

In conclusion, a minority of participants referred to the difficulty of grading the categories of Content and Organization, equal weighting, and vague descriptors, which can be elucidated in a training and norming session where discussions and negotiations take place through a further explanation of the new rubric and the assessment of sample essays. If it does not suffice, the descriptors that are found unclear by the participants can be fine-tuned cooperatively.

### 4.5.3. Number of band levels

Following the strengths and weaknesses of the new rubric, the second question in the open-ended questionnaire explored participants' perspectives on the number of band levels, which is another important aspect of the rubric design at the descriptor level (Knoch, 2011, p. 92). Participants were asked to state and expand on their opinions regarding the four band levels in each category, which are labelled as *Meets Expectations*, *Approaches Expectations*, *Needs Development*, and *Inadequate*.

Out of 13 participants, 10 of them (77 %) stated that they found four band levels adequate. One of the remaining 3 participants was unsure, whereas 2 participants expressed that they would feel more confident with a rubric with five band levels. Participants 7 and 10 express their opinions as follows:

> 102. The four levels in each category are very helpful in assessing a student's writing ability. Each of the four levels clearly states the requirements and is useful as a means of evaluating writing proficiency with confidence. Additionally, it's very easy for anyone to understand a student's competency in different areas of writing competency by reviewing the student's scores for each category (*P7*).
>
> 103. These four levels are adequate. We can measure and define students' writing proficiency effectively with them. They include specific descriptors. The detailed descriptive language outline what I need to see in my students' writings very explicitly. Also, because the number of levels isn't too many or too few, it wouldn't cause students to feel confused and gives the basic necessary feedback. It justifies even small differences adequately *(P10*).

Participant 1 who is not quite sure about the number of band levels explains his ideas as follows:

> 104. Firstly, I can say that these four levels were enough to cover my expectations and impression of the essays. However, the highest score in the rubric does not reflect perfection and the lowest does not reflect total failure. In this sense, the rubric is quite similar to those used for the assessments in international exams. The highest score not reflecting perfection seems more realistic and student-friendly. Due to these reasons, I felt like there could be another level for the lowest score because the description still implies that there is still effort to some extent. My explanation here might be somewhat difficult to understand. This is because I did not always feel the latter. I just felt I needed a lower score for maybe 2 or 3 essays (*P1*).

According to Participant 8 who thinks four band levels are inadequate:

> 105. Having 4 levels in each category means every level has 25% share. I think this is high. If it was 20% which means 5 levels in each category, it would be easier for me to grade. For a very weak paper, I may prefer to give a mark below 5, which means lower than 25 out of 100 (*P8*).

As the excerpts above indicated, while 77% of the participants believed that four band levels were adequate, the rest of them had different opinions regarding the number of band levels. Knoch (2011, p. 92) supports that as it is with many of the decisions made during the rubric design process, the decision about the number of band levels should also depend on the context in which a rubric is to be used. She goes on to say that the seven (plus or minus two) band level rule is appropriate if a writing test is administered to a very different ability group of test takers, which was also the case for the writing performance test in this study. The original version of the ESL Composition Profile (Jacobs et al., 1981) has four band levels (Excellent to Very Good, Good to Average, Fair to Poor and Very Poor); however, there are three band levels (Good to Average, Fair to Poor and Very Poor) in each category of the writing rubric used currently for the performance-based assessment of writing proficiency at BUU-SFL-IEP, which is an adaptation of the ESL Composition Profile. According to the majority of the 24 participants of Phase 1 of this research, three band levels were not sufficient. The lack of adequate number of band levels was also mentioned by the participants in Phase 1 under the heading of the weaknesses of the current rubric (See 3.2.6.2 above). Specifically, out of twenty-four participants, 16 of them (67%) considered three band levels to be insufficient. 8 of these 16 participants supported that there should be *four* band levels while 4 of them believed *five* band levels were necessary for fair scoring.

Based on a variety of sources (such as the extant literature, goals and objectives of the writing course, the results of Phase 1 where participants' perspectives on the writing rubric in use at present were explored, and the opinions of experts in performance-based assessment of writing), a draft rubric with five band levels was designed in Phase 2 of the study, and this draft rubric was piloted with 5 experienced raters in Phase 3 in order to trial the draft rubric, identify possible weaknesses in the use of it, and refine the rubric based on the feedback received from the raters as carried out in Knoch (2007) and Hattingh (2009). Regarding the band levels, the results of Phase 3 indicated that out of 5 participants 4 of them thought a rubric with four band levels would be more practical than a rubric with five band levels because it would be more clear-cut and less confusing. Thus, the data gathered from different participants in different phases of the study i.e., Phases 1, 3, and 5 verified that a rubric with four band levels was appropriate for the assessment of writing proficiency at BUU-SFL-IEP.

Following the number of band levels, the next question in the open-ended questionnaire explored participants' perceptions on the wordings of descriptors in each category in order to be able to gather more in-depth feedback on the wording of descriptors in each category.

### 4.5.4. Wordings of descriptors in each category

The third question in the open-ended questionnaire looked into the participants' perceptions of the wordings of descriptors in each of the five categories and whether there were any categories in which participants think the wording of descriptors needed to be changed.

#### 4.5.4.1. *Content*

As mentioned above in the sub-section titled weaknesses of the new rubric, out of 13 participants 5 of them stated that they had some difficulty in assessing the categories of Content and Organization. Based on their explanations, it is possible to state that the difficulty they had mostly arose from how many points to deduce when one of the criteria in the categories of Content or Organization was missing in students' essays, such as the refutation in the content or the title in the organization rather than the descriptors of these categories in the rubric.

When the participants asked whether there were any categories in which they thought the descriptors needed to be changed, only 3 of them out of 13 expressed such a need.

See Table 4.23 below for the descriptors in the category of content.

**Table 4.23**. *The descriptors for the category of content in the new rubric*

| ***Content*** |
|---|
| ***4. Meets Expectations*:** Essay addresses the assigned topic; ideas are concrete and thoroughly developed with supporting evidence provided; counterargument with a sound refutation; no extraneous material; essay reflects carefully planned thought |
| ***3. Approaches Expectations:*** Essay addresses the assigned topic but misses some points; ideas could be more fully developed as some evidence may be lacking; counterargument with an acceptable refutation; some extraneous material is present |
| ***2. Needs Development:*** Development of ideas not complete or essay may deviate from the topic; the main ideas may not be fully supported by the evidence given; counterargument with weak or no refutation |
| ***1. Inadequate:*** Ideas incomplete due to lack of supporting data; essay does not reflect careful thinking or was hurriedly written; inadequate effort in area of content |

Out of the three participants who stated the wording of the descriptors needed to be changed, two participants (Participants 2 and 8) explicitly wrote what they think should be modified. According to Participant 2, in Band 2 the phrase *somewhat off the topic* needed to be expanded. S/he justifies her argument as follows:

> 106. I think the phrase somewhat off the topic in band 2 is confusing because if the essay is mainly a bit off-topic, it would be fair to give only 1 point. If it is the existence of irrelevant information or examples in minor supporting sentences, it will be more straightforward to use familiar wording to the one in band 4: 2. Needs Development: Development of ideas not complete or essay contains some extraneous material that distracts reader off the main topic. Something like this maybe… *(P2).*

Another participant, Participant 8, recommended an addition to Band 2:

> 107. 2. Needs Development: '*frequent repetition of ideas*' may be added to this band. In some of the papers, I saw the same idea was repeated for many times *(P8).*

When the modifications are made as recommended by the two participants, Band 2 may prove to be clearer for not only the two participants who suggested these modifications but also the other participants who stated that they had difficulty in assessing the categories of Content and Organization in the section titled weaknesses of the new rubric.

#### 4.5.4.2. *Organization*

As it was with the category of Content, the category of Organization was stated to be difficult to assess by five participants in the section titled the weaknesses of the new rubric. The difficulty mostly stemmed from the lack of agreed-upon decisions on how many points to deduce when one of the criteria in the categories of Content or Organization was missing in students' essays rather than the ambiguity of the wording of the descriptors. For instance, Participant 6 asks:

> 108. One of the confusing parts was that there were some papers whose organization approached expectations in general but e.g. the order of the paragraphs in the essay contradicted the order of the predictors given in the thesis statement. What score should I assign for the organization, 3 or 2? *(P6).*

As also mentioned above, the difficulty could be easily overcome by holding a training and norming session, making ground rules that could be compiled under the title of guidelines in this session through negotiation, and asking each rater to follow these guidelines to ensure consistency.

Apart from such concerns, again out of 13 participants, 3 participants expressed a need to revise the wording of some descriptors in the category of Organization.
See Table 4.24 below for the descriptors in the category of Organization.

**Table 4.24.** *The descriptors for the category of organization in the new rubric*

| ***Organization*** |
|---|
| ***4. Meets Expectations*:** An informative title that covers the topic; effective introductory paragraph, topic is stated in a clear thesis statement, leads to body with three full paragraphs; appropriate transitional expressions used; conclusion logical & complete |
| ***3. Approaches Expectations:*** Adequate title, introduction & conclusion; body of essay is acceptable with three paragraphs that may not be fully developed; sequence is logical but few transitional expressions may be absent, misused, or overused |
| ***2. Needs Development:*** Mediocre or scant introduction, body, or conclusion; paragraphs aren't divided exactly right; problems with the order of ideas in body; problems of organization interfere |
| ***1. Inadequate:*** Shaky or minimally recognizable introduction; organization can barely be seen; severe problems of ordering ideas |

Two of the three participants who felt the need to revise the wording of descriptors stated that "title" should either be included in the descriptors of all bands or be excluded from all of them without exception. Title was particularly emphasized in the top two bands, Bands 4 and 3, to indicate that the organization part of an essay could be scored 4

or 3 only if it includes all the components of an essay starting from the title till the end. However, title could surely be included in the last two bands, Bands 2 and 1, in a way that represents the criteria in these two bands.

#### 4.5.4.3. *Grammar*

The category of Grammar was stated to be the most comprehensible hence the easiest to assess category by all of the 13 participants who took part in Phase 5 of the study as reflected in the following excerpt by Participant 7:

> 109. The grammar category is very well-designed, and the levels are divided appropriately. It makes a meaningful assessment of a student's abilities possible with the descriptors in each level *(P7).*

Below in Table 4.25 are the descriptors for the category of grammar.

**Table 4.25.** *The descriptors for the category of grammar in the new rubric*

| ***Grammar*** |
|---|
| ***4. Meets Expectations*:** Proficient in English grammar with almost no grammar errors; all/majority of simple, compound, and complex sentences are correct and appropriate; mostly accurate and appropriate use of prepositions, modals, articles, verb forms, and tense sequencing; no fragments or run-on sentences |
| ***3. Approaches Expectations:*** Rather proficient in English grammar; few grammar problems which don't distort meaning; almost no fragments or run-on sentences |
| ***2. Needs Development:*** Somewhat proficient in English grammar; grammar problems distort meaning; fragments or run-on sentences present |
| ***1. Inadequate:*** Limited proficiency in English grammar; grammar problems hinder meaning; difficult to understand sentences |

There was, on the other hand, one important issue stressed by two participants regarding the wording of descriptors in Bands 3 and 2. According to Participant 6:

> 110. In some papers there was a tedious use of the same simple structures. The meaning was not distorted but their grammar cannot be labeled as 'rather proficient', either! *(P6).*

Participant 2, who raised the same issue, suggested a revision in the wording of descriptors and thus a solution for the problem, as follows:

111. There should be a difference between the scores assigned to the students who played in the safe zone, their linguistic comfort zone, and the students who took risks to build a better text. We can add the following phrases to make the distinction maybe:

3 Approaches Expectations: ... attempts sentence variety with few minor problems which don't distort meaning …

2 Needs Development: ... overwhelming use of simple sentences … *(P2)*

Finally, Participant 8 recommended an addition to the top band, Band 4:

112. I think 'word order' is one of the most frequent mistakes we see in student papers. It can be a good idea to add 'correct word order' in this category *(P8)*.

The category of Grammar could be even more straightforward for the raters with the revisions suggested by the teacher-raters themselves.

#### 4.5.4.4. *Vocabulary*

Following the category of Grammar, the category of Vocabulary was also mentioned to be straightforward and easy to assess by the majority of the participants. Below in Table 4.26 are the descriptors for the category of vocabulary.

**Table 4.26.** *The descriptors for the category of vocabulary in the new rubric*

| ***Vocabulary*** |
|---|
| ***4. Meets Expectations*:** All/majority of vocabulary use is appropriate, accurate, and varied |
| ***3. Approaches Expectations:*** Appropriate and accurate use of vocabulary with few errors; attempts variety |
| ***2. Needs Development:*** Some vocabulary misused; repetitive use of basic vocabulary with little variety |
| ***1. Inadequate:*** Serious problems of vocabulary; lacks variety |

While the majority of the participants expressed their satisfaction with the wording of descriptors, 3 participants out of 13 stated that they found the descriptors less explanatory when compared to the descriptors in the other categories, as described in the following excerpt by Participant 8:

113. 4. Meets Expectations parts for all the other categories are more detailed and clear. I think Vocabulary should be this way, too. Appropriate to the audience, the tone of the paper etc. *(P8).*

As previously mentioned, instead of a training and norming session which could have an undesirable effect on the rating process, teacher-raters were provided with an assessor guide that gives in-depth information on the criteria of each category (See Appendix T). The three pillars of the category of Vocabulary i.e., appropriacy, accuracy, and variety were explained there due to the limited space rubrics generally have. Considering the fact that raters may not desire to refer to the assessor guide while rating, the aspects listed under the category of Vocabulary could be condensed into the space allocated to that category in the rubric, as suggested in excerpt 112 by Participant 8.

#### 4.5.4.5. *Mechanics*

The fifth and the last category in the rubric was Mechanics. Regarding the wording of the descriptors in the category of Mechanics, the overwhelming majority of the participants stated that they found the descriptors clear (12 participants out of 13). See Table 4.27 below for the descriptors in the category of mechanics.

**Table 4.27.** *The descriptors for the category of mechanics in the new rubric*

| ***Mechanics*** |
|---|
| ***4. Meets Expectations*:** Correct use of English writing conventions; all needed capitals, paragraphs indented, almost no punctuation and spelling errors; very neat |
| ***3. Approaches Expectations:*** Some problems with writing conventions or punctuation; occasional spelling errors; paper is neat |
| ***2. Needs Development:*** Uses general writing conventions but has errors; spelling problems distract reader; punctuation errors interfere with ideas |
| ***1. Inadequate:*** Complete disregard of English writing conventions ; severe spelling problems; errors in sentence-final punctuation; obvious capitals missing |

Only 1 participant, Participant 2, recommended a minor revision in the wording of descriptors:

> 114. The phrase "general writing conventions" could be expanded in not only Band 4 but also the other bands. I am not sure but what I understand from general writing conventions is the accurate use of capitals and the indentation of paragraphs based on Band 4. This can be clarified *(P2).*

As a matter of fact, Participant 2 seemed to have understood what general writing conventions entailed. However, she seemed to have needed to be reassured which was reasonable, particularly when the decisions are high-stakes. Apart from Band 4, what

general English writing conventions included was also explained in the assessor guide. As it was the case in the category of Vocabulary, it seemed that raters did not prefer to consult the assessor guide most probably to save time. Therefore, it seemed that they expected to see as comprehensive as possible descriptors.

As mentioned earlier, raters do not seem to have a problem with the wording of descriptors in the category of mechanics. However, few of the participants are concerned about its weighting, which is the subject of the next section.

### 4.5.5. Weighting

The majority of the participants i.e., 9 out of 13 believe that equal weighting was an asset of the new rubric while 4 of them supported some categories should have higher weighting than the others. According to Participants 2 and 4:

> 115. I think there is no problem with the weighting because all these criteria are complementary regarding to the construction of an essay *(P2)*.
>
> 116. This even weighting shows that no certain concept is judged more heavily than others. Since the writing products of our students that we are going to assess using this rubric is part of their proficiency exam that they take at the end of their language education in prep class, I believe using such a rubric with even weighting is the right thing to do. Our proficiency exam is a summative test, and it covers everything students have learned throughout the academic year *(P4)*.

One of the 4 of the participants who believed the weighting of each category needed to be different explained his/her views as follows:

> 117. We are evaluating how proficient the students are in writing, and of course all these categories are crucial in this. However I am not sure mechanics is as crucial as content and organization. I agree that we teach indenting, punctuation, capitalization, but I do not think they have the same importance with how students express and organize their ideas. We, both students and teachers, spend more time and effort to teach and learn and practice how to use language to express ourselves. To sum up, mechanics can have a lower value – maybe 2 or maximum 3 *(P8)*.

While the 4 participants agreed that the weighting of categories should be different, their ideas differed when it came to the categories that should have higher weighting than the others, as the following excerpts indicate:

> 118. Mechanics could be allotted for 3 while content may be allocated for 5 since the richness of the ideas and their communicative power is highly crucial *(P5)*.

> 119. As I mentioned before 4 is too much for mechanics. 2 points can be adequate for it. 5 points could be given to vocabulary and grammar each *(P13).*

As highlighted a few times before within the scope of this research, the weighting of categories has been a debatable issue in the literature on the performance based assessment of writing proficiency (East & Cushing, 2016). The context of this study was not an exception in this respect, either. Nonetheless, as opposed to few participants who did not favor the equal weighting of categories and could not come to an agreement on which categories needed to outweigh others, the majority of the teacher-raters who were participants in different phases of this research found equal weighting advantageous mostly because it was fair.

The next question in the open-ended questionnaire explored whether there were any categories the participants found difficult to score.

### 4.5.6. Categories considered to be difficult to score

Out of 13 participants 5 of them stated that there was/were a category/categories they had difficulty in scoring. While 4 of them found the categories of Content and Organization more challenging to score, one of them stated she had difficulty in scoring the category of Vocabulary. As also stated in the section titled weaknesses of the new rubric (See 3.6.6.2), teacher-raters had difficulties in how many points to deduce when one of the criteria in the categories of Content or Organization was missing in students' essays rather than the use of the rubric. Participant 8 expressed his/her opinions as follows:

> 120. I sometimes found it difficult to give or take points off the paper about the thesis statement. I couldn't be sure if it should be in the content or organization category. In the rubric, it is in the organization. It is generally OK. However, I sometimes felt that it is also related to the content. I felt/thought just the opposite with 'counter argument and refutation'. A few students wrote 3 body paragraphs with no counter argument or refutation, which was confusing for me *(P8).*

As highlighted before, the difficulty could be easily overcome by holding a training and norming session, making ground rules that could be compiled under the title of guidelines in this session through negotiation, and asking each rater to follow these guidelines to ensure consistency.

Another participant, Participant 7 stated that s/he had difficulty in scoring the category of vocabulary:

> 121. Vocabulary was the hardest category to grade. The level explanations were the least descriptive, and I think it was the most subjective category *(P7).*

A similar concern was expressed by Participant 8 in the section titled the weaknesses of the new rubric (See 3.6.6.2). As also put forward in that section, appropriacy, accuracy, and variety were the three pillars which the descriptors in the category of vocabulary were built on. Due to the limitations in space, what each pillar entailed was described in the assessment guide provided for the teacher-raters rather than the rubric. However, because raters may not desire to refer to the assessor guide during the process of rating, the three aspects listed under the category of Vocabulary could be condensed into the space allocated to that category in the rubric.

Whether teacher-raters had difficulties in scoring any category or categories was also explored in Phase 1 for the rubric which is currently used in the assessment of writing performance and in Phase 3 for the draft rubric. The majority of participants in Phase 1 (15 out of 24) stated that all categories in the rubric we use at present are problematic mostly because of the wording of descriptors and the range of scores within each band level (See 3.2.6.2). The draft rubric the design process of which was explained in Phase 2 (3.3.3.2) was designed by taking this feedback into consideration in addition to the existing literature, goals and objectives of the writing course, and the opinions of experts in performance-based assessment of writing. Then the draft rubric was piloted with 5 experienced raters in Phase 3 so that possible weaknesses could be identified in the use of it, and it could be refined based on the feedback received from those experienced five teacher-raters. Three of the participants' basic concern was the number of band levels, which was five. The feedback received from the participants in Phase 1 and Phase 3 informed the process of the rubric design in this study. The results of Phase 5 indicated that the majority of the participants were satisfied with the new rubric, and they stated that with few refinements it could contribute to the consistent and fair assessment of writing performance at BUU-SFL-IEP.

In the final part of the open-ended questionnaire participants were asked their general satisfaction level in using the rubric.

### 4.5.7. Participants' general satisfaction level

Finally, 13 participants of the fifth phase of the study who take part in the proficiency examination at BUU-SFL-IEP as raters were asked to rate the new writing rubric from 1 to 5 (ranging from very satisfied to very dissatisfied) on:

- the extent to which it facilitates fair assessment of students' written work and
- the participants' confidence level in applying the writing rubric.

#### 4.5.7.1. *Participants' perceptions on fair assessment of students' written work*

The first item about the general satisfaction of participants in using the new rubric explored the extent to which it assisted in the fair assessment of students' written work. Participants were asked to rate the new rubric from 1 to 5:

5. Very satisfied
4. Satisfied
3. Neither satisfied nor dissatisfied
2. Dissatisfied
1. Very dissatisfied

See Table 4.28 below for the number and percentages of participants' ratings.

**Table 4.28.** *Participants' perceptions on fair assessment of students' writing (N=13)*

| | ***F*** | **%** |
|---|---|---|
| ***5*** | 7 | 54 |
| ***4*** | 6 | 46 |
| ***3*** | - | - |
| ***2*** | - | - |
| ***1*** | - | - |

As can be seen in the table, the majority of the participants were very satisfied with the new writing rubric in terms of its facilitation of the fair assessment of students' written work. The rest of the participants were satisfied with the new rubric. None of the participants had a negative or even a neutral attitude towards the new rubric in general apart from few refinements they thought necessary to make it function more effectively. Unlike it was with the new rubric, in Phase 1 of the study only one third of the 24

participants were satisfied with the current writing rubric in terms of its facilitation of fair assessment of students' written work while one third of them were dissatisfied with it. The majority of the participants were neither satisfied nor dissatisfied with the rubric, which was not an unexpected result considering the abundance of the negative criticisms expressed in the other sections of the open-ended questionnaire.

#### 4.5.7.2. *Participants' confidence level in applying the writing rubric*

The second item in this section investigated participants' confidence level in applying the new writing rubric. See Table 4.29 below for the number and percentages of participants' ratings.

**Table 4.29.** *Participants' confidence level in using the new rubric (N=13)*

| | ***F*** | **%** |
|---|---|---|
| ***5*** | 4 | 31 |
| ***4*** | 9 | 69 |
| ***3*** | - | - |
| ***2*** | - | - |
| ***1*** | - | |

In accordance with the results of the first item, out of 13 participants 9 of them felt confident in applying the new writing rubric, and the other 4 participants felt very confident as opposed to only 9 participants (out of 24) in Phase 1 of this study who felt confident in using the rubric.

### 4.5.8. Conclusion of Phase 5

As stated a few times within the scope of this research, rubric design is not a one-shot activity. In addition to acknowledging relevant research and theory, it requires the engagement of different parties such as teachers, experts, and students in the process so that a context-sensitive rubric that may function reliably and validly could be developed. It also requires an ongoing evaluation so that the different needs and expectations of the stake-holders from the performance-based assessment could be fulfilled. The findings of the MFRM analysis in Phase 4 indicated that the new rubric designed to be used for the

performance-based assessment of writing proficiency at BUU-SFL proved to be statistically reliable and valid. The results of this fifth phase of the study where the 13 teacher-raters' perceptions on the new rubric were explored support the quantitative findings indicating that all participants were satisfied with the new rubric despite considering some revisions necessary before it is started to be used for assessment.

## 6. CONCLUSION

This research aimed at developing an alternative theoretically-based and an empirically-validated multi-trait writing rubric which may serve to measure writing proficiency more validly and reliably in the writing section of the English language proficiency examination administered at the end of each academic year at BUU-SFL-IEP.

### 5.1. Summary of the study

The process of rubric design consisted of five phases each of which had a different purpose in order to serve for the general aim of the study. The results of the first phase which aimed at exploring the participants' perceptions of the writing rubric currently in use indicated that weaknesses outweigh strengths by almost doubling them. These results were compliant with the satisfaction and confidence levels of the participants the majority of whom were dissatisfied with the rubric (16 out of 24) and did not feel confident in using it (15 out of 24) although they had been using the rubric for a long time. As a result of this phase of the study, an assessor-oriented analytic/multi-trait rubric that had five categories with five-band levels and concrete descriptors was decided to be designed for the specific context of BUU-SFL-IEP for the performance-based assessment of EFL writing proficiency.

The second phase of the study aimed at designing a draft rubric for the performance-based assessment of writing proficiency relying on the contextual requirements of BUU-SFL-IEP, the related literature, and the expert opinion. The recommendations of the three experts on the content and the number of categories, the number of band levels, the wordings of descriptors, and the weighting brought in the third draft of the new rubric that was used in the third phase of the study in order to pilot the new rubric.

The aim of the third phase of the study was to pilot the draft rubric through an initial implementation, identify possible weaknesses in the use of it, and refine the rubric based on the feedback that would be received from the raters as carried out in Knoch (2007) and Hattingh (2009). Another aim of this phase was to pilot the open-ended questionnaire that would be used to explore the perceptions of the raters on the efficacy of the draft rubric and discover the potential problems that may exist in the open-ended questionnaire, such as the clarity of the items. (Zacharias, 2012, p. 71). The results of the MFRM analysis showed that the draft rubric designed to be for the performance-based assessment of EFL

writing proficiency at BUU-SFL-IEP would function reliably and validly to a great extent, whereas the findings of the open-ended questionnaire indicated the necessity of making slight revisions in the rubric. In order to be able to build a more functional rubric, the modifications were carried out by the researcher of the study (See Appendix U). Another purpose of this third phase of the study was to trial the open-ended questionnaire to be utilized to find out the perceptions of the participants on the new rubric. The depth and width of the data gathered through the open-ended questionnaire demonstrated that the instrument would function as required. Together with the completion of the third phase of the study the piloting of the new rubric and the open-ended questionnaire to be used in the fifth phase of the current research were completed.

The aim of the fourth phase, which yielded quantitative data, was to empirically validate the alternative multi-trait rubric designed for the performance-based assessment of writing proficiency at BUU-SFL-IEP through statistical analyses, specifically MFRM, a general psychometric modelling approach that is described "particularly well-suited to dealing with many-facet data typically generated in rater-mediated assessments" (Eckes, 2015, p. 19). Fit statistics that were in the form of summary results and fit statistics for each facet accompanied by graphical displays in the MFRM analysis showed that the scores the teacher-raters awarded to examinees by using the new rubric to assess their writing proficiency proved to be reliable and valid. When compared to the MFRM analysis carried out in the third phase of the study where the draft rubric was trialed and refined, the findings of the MFRM analysis to investigate the new rubric in this phase of the study were even more satisfactory; particularly the ones related to the facet of category difficulty, which could be related to the revisions made in the rubric based on the feedback received from the participants in the piloting phase of the study.

The results of this fifth phase of the study where the 13 teacher-raters' perceptions on the new rubric were explored supported the quantitative findings indicating that all participants were satisfied with the new rubric despite considering some revisions necessary before it was started to be used for assessment.

There were three research questions to realize the general aim of this study which was to design an alternative theoretically-based and an empirically-validated multi-trait writing rubric so that writing proficiency could be measured more validly and reliably in the writing section of the English language proficiency examination administered at the end of each academic year at BUU-SFL-IEP. The first research question explored the

raters' perspectives on the rubric that is currently used for the performance-based assessment of writing proficiency. Supporting the views that teacher-raters had voiced in different platforms, the results indicated a serious and restless dissatisfaction. In order to be able to satisfy the needs of the specific context of the current research and meet the expectations of the teacher-raters, the learning outcomes related to writing instruction in the institution were revisited, the related literature was reviewed, and the perceptions of the teacher-raters for a well-functioning rubric were also explored so that an alternative multi-trait rubric could be designed. After receiving the opinions of the three experts who were specialized in performance-based assessment of writing, the draft rubric was finalized.

The second research question of the study investigated the extent to which the alternative theoretically-based and empirically-developed multi-trait rubric of academic writing (that was newly designed) valid and reliable for the measurement of performance-based assessment of EFL writing proficiency at BUU-SFL-IEP. According to the results of the MFRM analysis, the draft rubric with five categories and five band levels proved to be statistically reliable and valid. However, the teacher-raters in the study found five band levels unnecessary due to the ambiguity caused by the 0 band level and misguidance of the labels that each band level had while they stated their complete satisfaction with the five categories. Thus, based on the findings obtained through the piloting of the draft rubric, the newly designed rubric had five categories but four band levels with different labels. The results of the MFRM analysis indicated that the new rubric with four band levels and different labels proved to be even more statistically reliable and valid for the context of the current research.

Finally, the last research question looked into the teacher-raters' perspectives on the use of this alternative multi-trait writing rubric that was newly designed. The results indicated complete satisfaction on the side of the teacher-raters apart from few minor changes.

In conclusion, an alternative theoretically-based and empirically-developed multi-trait rubric that could serve reliably and validly for the measurement of performance-based assessment of EFL writing proficiency at BUU-SFL-IEP was designed based on a variety of validity evidence.

## 5.2. Limitations of the study

Although the research reported in this research was carefully designed, several shortcomings need to be acknowledged. Firstly, the writing rubric designed for the purposes of this research was based on a specific task type. This surely limits the generalizability of the resulting ratings to other contexts. When task effects are embedded in the descriptors, the ratings are based on the tasks that the test takers performed. Thus, the ratings brought about a context dependent measure that does not generalize. According to Knoch (2007, p. 290), a case for generalizability can be made from the perspective of the task in use. If the task is generally representative of what is expected in the target use domain; that is, the academic setting, this would establish generalizability even if the rubric is based on only this task.

The second limitation of the study relates to sample size. Only thirteen raters were used in Phase 4 of the study, and these raters rated only fifty essays each. However, a fully-crossed design where each rater rated all fifty essays was chosen. This is generally considered as aiding to increase the stability of the statistics yielded by FACETS (Myford & Wolfe).

As emphasized within the limited scope of this research a few times, the opinions of all stakeholders including students need to be consulted during the process of rubric development. While the opinions of experienced-teacher raters and experts were referred to at the different phases of the current research, students were not included in the process. If the rubric is decided to be used for the performance-based assessment of writing proficiency in an actual proficiency examination, students' could also be included in the process.

Finally, although the rubric proved to be reliable and valid in the context of the current research, the full impact that it may have on the accuracy, consistency, and reliability of ratings can only be examined once the rubric is implemented by all the raters for the writing performance of all learners who sit the proficiency exam at the end of the academic year.

## 5.3. Implications and suggestions for further research

Acknowledging its limitations, this study has important implications for rubric development and validation. The development and modification of writing rubrics is rarely discussed in the language assessment literature in general (Banarjee et al., 2015;

Knoch, 2009, 2011; Lallamamode, Daud, & Abu Kassim, 2016) and Turkish EFL context in particular (Hatipoğlu, 2015). To the best knowledge of the researcher, this study is the first to develop and validate a rubric theoretically and empirically for the performance-based assessment of EFL writing proficiency that is administered in an IEP of a Turkish-state university at the end of each academic year as part of the proficiency examination. Considering the consequences of the examination for students, it might be concluded that it is a high-stakes test. Most IEP's at Turkish state and private universities have performance-based assessment of writing as part of their proficiency tests, which makes calls for work on validation frequent. Ongoing rubric analysis to validate and modify the rubric when necessary could be a solution to prevent the construct-irrelevant variance that threatens the validity and fairness of assessment outcomes in these contexts due to the subjective nature of performance-based assessment. The model used in this study could readily be adapted to determine the validity and reliability of the rubrics utilized for the assessment of writing proficiency.

As emphasized a few times within the scope of the current research, validation requires the evaluation of an instrument based on a variety of quantitative and qualitative forms of evidence to support inferences from test scores (Weir, 2005, p. 15). For the quantitative analysis of the scores assigned through the use of the draft rubric and the new rubric, Many Faceted Rasch Measurement (MFRM), a general psychometric modelling approach, was utilized. Using a resourceful approach like MFRM which allows the researcher to closely examine each of the facets included in the performance-based assessment of writing proficiency and their interrelationships proved to be very useful to evaluate the psychometric quality of many-facet data, as shown by the depth of information presented in the results and discussion section of the current research. To the best knowledge of the reseacher, there does not exist a context-specific rubric design and validation process in Turkish EFL context that used a theoretical framework and MFRM. Thus, the present study could provide guidance also in this respect.

In order for such a validation scheme to be realized, testing office and continuous professional development unit members may initially be trained for both theoretical and empirical aspects of the validation process. As Hatipoğlu (2010) puts forward, teachers involved in test design and development are held accountable for all the aspects of language testing and assessment even though they may not have the necessary skills and knowledge, which is directly related to the language assessment literacy of those teachers.

Recent research into the language assessment literacy of language teachers working in IEP's of Turkish state and private universities indicates that pre-service education has some limitations in this respect, and pre-service teachers are not equipped with necessary knowledge in pre-service education related to language testing and assessment (Ölmezer-Öztürk, 2018). Hence, as suggested by Ölmezer-Öztürk (2018), the content of the course in pre-service education could be considered to be revised, the number of courses related to language assessment might be increased, and more practical hands-on practice can be incorporated into these courses in not only pre-service education but also in-service training. The last but not the least, what characterizes language test development and language testing research is the necessary complementarity of applied linguistics and measurement expertise (McNamara & Knoch, 2012). Therefore, psychometric training needs to be a part of any language testing and assessment training program so that test designers and/or researchers can seize the opportunity to use psychometric models and programs, such as MFRM, in order to be able to explore their potential application in language testing and assessment contexts.

As a product of this study, a theoretically-based and an empirically-validated multi-trait writing rubric was designed for the specific context of this research. If the new rubric is fully implemented, impact and washback studies could be conducted in order to investigate the consequential validity of the rubric. Furthermore, as the most important stakeholders in the process, students could also be included in the rubric validation process in addition to the others, i.e., teacher-raters, test designers, experts, and academic coordinators.

## REFERENCES


Akşit, Z. (2018). *Validating aspects of a reading test.* (Unpublished PhD Thesis), Middle East Technical University.

Alderson, J. C. (1991b). Bands and scores. In J. C. Alderson, & B. North (Eds.), *Language testing in the 1990s* (pp. 71-86). London: McMillan.

Alderson, C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. London: Continuum.

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

Ataman, F. (1999). *A study on reliability and validity studies of METU School of Foreign Languages, Department of Basic English, June 1998 and September 1998 proficiency tests.* (Unpublished PhD Thesis), Middle East Technical University.

Bacha, N. N. (2010). Teaching the academic argument in a university EFL environment. *Journal of English for Academic Purposes, 9*, 229-241.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*: *Designing and developing useful language tests*. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language testing in practice: Developing language assessments and justifying their use in the real world.* Oxford: Oxford University Press

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly, 2*(1), 1-34.

Banarjee, J., Yan, X., Chapman, M., & Elliot, H. (2015). Keeping up with the times: Revising and refreshing a rating scale. *Assessing Writing, 26*, 5-19.

Barber, J. P., Walczak, K. K. (2009). Conscience and critic: Peer debriefing strategies in grounded theory research. Paper presented at *the Annual Meeting of the American Educational Research Association,* San Diego, California**.**

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

Becker, A. (2011). Examining rubrics used to measure writing performance in U.S. intensive English programs. *The CATESOL Journal, 22*(1), 113-130.

Becker, A. (2018). Not to scale? An argument-based inquiry into the validity of an L2 rating scale. *Assessing Writing, 37*, 1-12.

Behizadeh, N., & Engelhard, G. (2011). Historical view of the influences of measurement and writing theories on the practice of writing assessment in the United States. *Assessing Writing, 16*(3), 189-211.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Lawrence Erlbaum Associates Publishers.

Broad, B. (2003). *What we really value: Beyond rubrics in teaching and assessing writing*. Logan: Utah State University Press.

Brown, J. D. (2004). Performance assessment: Existing literature and directions for research. *Second Language Studies, 22*(2), 91-139.

Brown, J. D. (2012). Introduction to rubric-based assessment. In J. D. Brown, (Ed.), *Developing, using, and analyzing rubrics in language assessment with case studies in Asian and Pacific languages* (pp. 1-9). Honolulu: University of Hawai'i, National Foreign Language Resource Center.

Brown, J. D., & Edmonds, CKA. (2012). Issues in analyzing rubric-based results. In J. D. Brown, (Ed.), *Developing, using, and analyzing rubrics in language assessment with case studies in Asian and Pacific languages* (pp. 32-49). Honolulu: University of Hawai'i, National Foreign Language Resource Center.

Brown, J. D., & Rodgers, T. S. (2002). *Doing second language research*. Oxford: Oxford University Press.

Bukta, K. (2014). *Rating EFL written performance*. London: Versita.

Canale, M. (1983). From communicative competence to communicative language pedagogy. In J. C. Richards & R. Schmidt (Eds.), *Language and communication* (pp. 2-27). London, UK: Longman.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*, 1-47.

Chalhoub-Deville, M. (1997). Theoretical models, assessment frameworks and test construction. *Language Testing, 14*(1), 3–22. https://doi.org/10.1177/026553229701400102

Chan, S. (2011). *Demonstrating the cognitive validity and face validity of PTE Academic Writing items Summarize Written Text and Write Essay*. UK: Pearson.

Chiang, S. Y. (1999). Assessing grammatical and textual features in L2 writing samples: the case of French as a foreign language. *Modern Language Journal, 83*(2), 219–232.

Chiang, S. Y. (2003). The importance of cohesive conditions to perceptions of writing quality at the early stages of foreign language learning. *System, 31*, 471-484.

Council of Europe. (2001). *Common European framework of reference for languages (CEFR): Learning, teaching and assessment.* Strasbourg.
Council of Europe. (2018). *Common European framework of reference for languages (CEFR): Learning, teaching and assessment.* Strasbourg.

Creswell, J. W. (2012). *Educational research*. Boston: Pearson.

Cronbach, L. J. (1971). Test validation. In Robert L. Thorndike (Ed.). *Educational Measurement* (pp. 443-507). Washington, DC: American Council on Education.

Cronbach, L. J. (1984). *Essentials of psychological testing*. New York: Harper and Row.

Crossley, S., & McNamara, D. (2012). Predicting second language writing proficiency: the roles of cohesion and linguistic sophistication. *Journal of Research in Reading, 35*(2); 115-135.

Crossley, S., & McNamara, D. (2014). Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. *Journal of Second Language Writing, 26*, 66-79.

Crusan, D. (2010). *Assessment in the second language writing classroom*. Ann Arbor, MI: The University of Michigan Press.

Crusan, D. (2015). Dance, ten; looks, three: Why rubrics matter. *Assessing Writing, 26*, 1-4.

Davies, A. (2008). Assessing academic English: Testing English proficiency 1950–89: the IELTS solution. Cambridge: Cambridge University Press.

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 for foreign language writing. *Assessing Writing, 14*, 88-115.

East, M., & Cushing, S. (2016). Innovation in rubric use: Exploring different dimensions. *Assessing Writing, 30*, 1-2.

Eckes, T. (2011). *Language testing and evaluation 22: Introduction to many-facet Rasch measurement*. Frankfurt am Main: Peter Lang.

Eckes, T. (2015). *Introduction to Many-Facet Rasch Measurement: Analyzing and Evaluating Rater-Mediated Assessments* (2nd ed.). New York: Peter Lang.

Eckes, T., Müller-Karabil, A., & Zimmermann, S. (2017). In D. Tsagari & J. Banarjee (Eds.), *Handbook of Second Language Assessment* (pp. 61-76). Berlin: DeGruyter Mouton.

Edmonds, CAK. (2012). Maori Language Proficiency in Writing: The Kaiaka Reo Year Eight Writing Test. In J. D. Brown, (Ed.), *Developing, using, and analyzing rubrics in language assessment with case studies in Asian and Pacific languages* (pp. 63-76). Honolulu: University of Hawai'i, National Foreign Language Resource Center.

Educational Testing Service. (1989). *TOEFL test of written English guide*. Princeton, NJ: ETS.

Engelhard, G., & Myford, C. M. (2003). *Monitoring faculty consultant performance in the Advanced Placement English Literature and Composition Program with a many-faceted Rasch model* (Research Rep. 03-01). Princeton, NJ: Educational Testing Service.

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

Geranpayeh, A. & Taylor, L. (2013). *Examining listening*: *Research and practice in assessing second language writing.* Cambridge: Cambridge University Press.

Ghanbari, B., Barati, H., & Moinzadeh, A. (2012). Rating scales revisited: EFL writing assessment context of Iran under scrutiny. *Language Testing in Asia, 2*, 83-99.

Gürsoy, S. (2013). *The English proficiency exam in EFL context: A validation study*. (Unpublished PhD Thesis), Çağ University.

Hamp-Lyons, L. (1991). Assessing second language writing in academic contexts. *Modern Language Journal, 77*(2), 229-240.

Hamp-Lyons, L. (1995). Research on the rating process. Rating non-native writing: the trouble with holistic scoring. *TESOL Quarterly, 29*(4), 759-762.

Hamp-Lyons, L. (2007). Worrying about rating. *Assessing Writing, 12*, 1-9.

Hamp-Lyons, L. (2016a). Farewell to holistic scoring? *Assessing Writing, 27*, A1-A2.

Hamp-Lyons, L. (2016b). Farewell to holistic scoring. Part two: Why build a house with only one brick? *Assessing Writing, 29*, A1-A5.

Hamp-Lyons L (2017) Language assessment literacy for learning-oriented language assessment, *Papers in Language Testing and Assessment, 6*(1), 88-111.

Hamp-Lyons, L. (2018). *Handbook of assessment for language teachers*. TALE Project, Erasmus +.

Harris, D. P. (1969). *Testing English as a second language*. McGraw-Hill Book Co.: New York.

Harris, M. & McCann, P. (1994). *Assessment*. Oxford: MacMillan Heinemann English Language Teaching.

Haswell, R.H. (2007). *Researching teacher evaluation of second language writing via prototype theory*, retrieved March 18, 2011, from http://www.writing.ucsb.edu/wrconf08/Pdf_Articles/Haswell-Article.pdf.

Hatipoğlu, Ç. (2010). Summative evaluation of an English language testing and evaluation course for future English language teachers in Turkey. *ELTED*, 13, 40– 51.

Hatipoğlu, Ç. (2015). English language testing and evaluation (ELTE) training in Turkey: Expectations and needs of pre-service English language teachers. *ELT Research Journal, 4*(2), 111–128.

Hattingh, K. (2009). *The validation of a rating scale for the assessment of compositions in ESL* (Unpublished PhD Thesis). The North-West University.

Hawkey, R., & Barker, F. (2004). Developing a common scale for the assessment of writing. *Assessing Writing, 9*, 122-159.

He, L. & Jiang, Z. (2020). Assessing second language listening over the past twenty years: A review within the socio-cognitive framework. In V. Aryadoust and T. Eckes (Eds.), *Frontiers in Language Assessment and Testing*, (pp. 147-161). Lausanne: Frontiers Media SA.

Hirvela, A. (2017). Argumentation & second language writing: Are we missing the boat? *Journal of Second Language Writing, 36*, 69-74.

Huck, S. W. (2004). *Reading statistics and research* (4th Ed.). Boston: Pearson Education.

Hughes, A. (1989). *Testing for language teachers*. Cambridge and New York: Cambridge University Press.

Huhta, A. (2008). Diagnostic and formative assessment. In B. Spolsky and Francis M. Hult (Eds.), *The Handbook of Educational Linguistics* (pp.469-483). Oxford, UK: Blackwell Publishing.

Hyland, K. (2004). *Second language writing*. Cambridge: Cambridge University Press.

Hymes, D.H. (1972). On Communicative Competence. In: J.B. Pride and J. Holmes (Eds.) *Sociolinguistics. Selected Readings* (pp. 269-293). Harmondsworth: Penguin.

Jacobs, H., Zinkgraf, S., Wormuth, D., Hartfiel, V., & Hughey, J., (1981). *Testing ESL composition: A practical approach.* Rowley, MA: Newbury House.

Janssen, G., Meier, V., & Trace, J. (2015). Building a better rubric: Mixed methods rubric revision. *Assessing Writing, 26*, 51-66.

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

Kalajahi, S. A. R., & Abdullah, A. N. (2015). Discourse connectors and cohesion in writing. *Mediterranean Journal of Social Sciences, 6*(3), 441-447.

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527-535.

Kane, M. (2006). Validation. In Robert L. Brennan (Ed.), *Educational measurement*, (pp. 18-64). Washington, DC: American Council on Education and Praeger.
Kane, M. (2013). Validating the interpretation and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73.

Khalifa, H., & Weir, C. J. (2009). *Examining reading: Research and practice in assessing second language reading*. Cambridge: UCLES/Cambridge University Press

Kim, J. (2014). Predicting L2 writing proficiency using linguistic complexity measures: A corpus-based study. *English Teaching, 27*(1), 27-51.

Knoch, U. (2007). *The development and validation of an empirically-developed rating scale for academic writing* (Unpublished PhD Thesis). University of Auckland.

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing, 26*(2), 275-304.

Knoch, U. (2011). Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from? *Assessing Writing, 16*, 81-96.

Kohn, A. (2006). The trouble with rubrics. *Language Arts, 3*, 12-15.

Kuhn, D. (1991). *The skills of argument*. Cambridge, UK: Cambridge University Press.

Kuhn, D. (2005). Education for thinking. Cambridge, MA: Harvard University Press.

Kutevu, E. (2001). *Investigating the validity of achievement and proficiency tests at Bilkent University School of English Language*. Middle East Technical University.

Küçük, F. (2017). *Assessing academic writing skills in Turkish as a foreign language*. (Unpublished Master's Thesis), Boğaziçi University.

Lacey, A., & Luff, D. (2007). *Qualitative data analysis*. The East Midlands: National Institute for Health Research.

Lallmamode, S. P., & Daud, N. M., & Kassim, N. L. (2016). Development and initial argument-based validation of a scoring rubric used in the assessment of L2 writing electronic portfolios. *Assessing Writing, 30*, 44-62.

Lantolf, J. P., & Frawley, W. (1985). Oral-Proficiency Testing: A Critical Analysis. The *Modern Language Journal, 69*(4), 337-345.

Lee, Y., Gentile, C., & Kantor, R. (2010). Toward automated multi-trait scoring of essays: Investigating links among holistic, analytic, and text feature scores. *Applied Linguistics, 31*(3), 391-417.

Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing, 28*(4), 543-560.

Linacre, J. M. (2002c). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16*, 878.

Linacre, J. M. (2004b). Rasch model estimation: Further topics. *Journal of Applied Measurement, 5*(1), 95–110.

Linacre, J.M. (2008). Assessing Statistically and Clinically Meaningful Construct Defiency/Saturation: Recommended Criteria for Content Coverage and Item Writing. http://www.rasch.org/rmt/rmt174d.htm.

Linacre, J. M. (2017). *A user's guide to FACETS Rasch-Model computer programs*. Available online www.winsteps.com.

Linacre, J. M., & Wright, B. D. (2002). Construction of measures from many-facet data. *Journal of Applied Measurement, 3*(4), 486–512.

Liu, F., & Stapleton, P. (2014). Counterargumentation and the cultivation of critical thinking in argumentative writing: Investigating washback from a high-stakes test. *System, 45*, 117-128.

Matsuda, P. K. (2003). Second language writing in the twentieth century: A situated historical perspective. In B. Kroll (Ed.), *Exploring the dynamics of second language writing* (pp. 15-34). New York: Cambridge University Press.

Matsuda, P. K. (2015). Identity in Written Discourse. *Annual Review of Applied Linguistics, 35*, 140-159.

McKay, S. L. (2006). *Researching second language classrooms*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
McNamara, T. (1996). *Measuring second language performance*. Harlow, Essex: Pearson Education.

McNamara, T. (2002). Discourse and assessment. *Annual Review of Applied Linguistics, 22*: 221-242.

McNamara, D. S., Crossley, S. A., & McCarthy, P.M. (2010). Linguistic features of writing quality. *Written Communication, 27*(1), 57-86.

McNamara, T., & Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing, 29*(4), 555-576.

Medve, V. B., & Takac, V. P. (2013). The influence of cohesion and coherence in text quality: A cross-linguistic study of foreign language learners' written production. In E. Piechurska and E. Szymanska-Czaplak (Eds.), *Language in cognition and affect* (pp. 111-133), Berlin: Springer-Verlag.

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. American Psychologist, *30*(10), 955-986.

Messick, S. (1980). Test validity and ethics of assessment. American Psychologist, *35*(11), 1012-1027.

Messick, S. (1989). Validity. In Robert L. Linn (Ed.), *Educational measurement* (pp. 13-103). New York: Macmillan.

Modhish, A. S. (2012). Use of discourse markers in the composition writings of Arab EFL learners. *English Language Teaching, 5*(5), 56-61.

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

Nakatsuhara, F. (2013). *Language testing and evaluation 30: The co-construction of conversation in group oral tests*. Frankfurt am Main: Peter Lang.

North, B. (2003). Scales for rating language performance: Descriptive models, formulation styles, and presentation formats. *TOEFL Monograph, 24*.

Plakans, L., & Gebril, A. (2017). An assessment perspective on argumentation in writing. *Journal of Second Language Writing, 36*, 85-86.

Ölmezer-Öztürk, E. (2018). *Developing and validating langauge assessment knowledge scale (LAKS) and exploring the knowledge of EFL teachers.* (Unpublished PhD thesis), Anadolu University.

Qin, J., & Karabacak. E. (2010). The analysis of Toulmin elements in Chinese EFL university. *System, 38*, 444-456.

Richards, J. C., Li, B., & Tang, A. (1998). Exploring pedagogical reasoning skills. In J. C. Richards (Ed.), *Beyond training: Perspectives on language teacher education* (pp. 86-102). New York: Cambridge University Press.

Richards, J. C., & Schmidt, R. (2010). *Longman dictionary of language teaching and applied linguistics*. Harlow: Pearson Education Limited.

Shaw, S., & Falvey, P. (2008). *The IELTS writing assessment revision project: Towards a revised writing scale*. Cambridge: University of Cambridge ESOL Examinations.

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

Shohamy, E. (1995). Performance assessment in language testing. *Annual Review of Applied Linguistics, 15*, 188-211.

Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The effects of raters' background and training on the reliability of direct writing tests. *The Modern Language Journal, 76*(1), 27-33.

Skehan, P. (1998). *A Cognitive Approach to Language Learning.* Oxford: Oxford University Press.

Spencer, L. S. &, Fitzgerald, J. (1993). Validity and structure, coherence, and quality measures in writing. *Journal of Reading Behavior, 25(*2), 209-231.

Stapleton, P., & Wu, Y. A. (2015). Assessing the quality of arguments in students' persuasive writing: A case study analyzing the relationship between surface structure and substance. *Journal of English for Academic Purposes, 17*, 12-23.

Tanskanen, S. K. (2006). *Collaborating towards coherence*. Amsterdam: John Benjamins Publishing Company.

Taylor, L. & Galaczi, E. D. (2011). Scoring validity of speaking tests. In L. Taylor & C. Weir, *Examining speaking: Research and practice in assessing second language speaking*. Cambridge: Cambridge University Press

Toulmin, S. (1958). *The uses of argument*. Cambridge: Cambridge University Press.

Toulmin, S. (2001). *Return to Reason*. Cambridge and London: Harvard University Press.

Tsagari, D. and Vogt, K. (2017). Assessment Literacy of Foreign Language Teachers around Europe: Research, Challenges and Future Prospects. *Papers in Language Testing and Assessment*, 6 (1), 41-64.

Tsui, A. B. (2003). *Understanding expertise in teaching: Case studies of ESL teachers.* New York: Cambridge University Press.

Tsui, A. B. (2005). Expertise in teaching: Perspectives and issues. In K. Johnson (Ed.), *Expertise in second language learning and teaching* (pp. 167-189). New York: Palgrave Macmillan.

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

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Oxford: Palgrave Macmillan.

Weir, C. J., Hawkey, R., Green, A., & Devi, S. (2009). *The cognitive processes underlying the academic reading construct as measured by IELTS*. IELTS Research Reports Volume 9 (Vol. 9). Bedfordshire.

Whittaker, R., Llinares, A., & McCabe, A. (2011). Written discourse development in CLIL at secondary school. *Language Teaching Research, 15*(3), 343-362.

Widdowson, H. G. (1978). *Teaching language as communication*. Oxford: Oxford University Press.

Wind, A . & Peterson, M. E. (2018). A systematic review of methods for evaluating rating quality in language assessment. *Language Testing, 35*(2), 161-192.

Wingate, U. (2012). 'Argument!' helping students understand what essay writing is about. *Journal of English for Academic Purposes, 11*, 145-154.

Xi, X. & Davis, L. (2017). Quality factors in language assessment. In D. Tsagari & J. Banarjee (Eds.), *Handbook of Second Language Assessment* (pp. 61-76). Berlin: DeGruyter Mouton.

Yang, W. & Sun, Y. (2012). The use of cohesive devices in argumentative writing by Chinese EFL learners at different proficiency levels. *Linguistics and Education, 23*, 31-48.

Yapar, T. (2003). *Study of the predictive validity of the Başkent University English proficiency exam through the use of the two-parameter IRT model`s ability estimates*. (Unpublished PhD Thesis), Middle East Technical University.

Yeğin, O. P. (2003). *The predictive validity of Başkent University proficiency exam (buepe) through the use of the three-parameter IRT model`s ability estimates*. (Unpublished PhD Thesis), Middle East Technical University.

Zacharias, N.T. (2012). *Qualitative research methods for second language education*. New Castle upon Tyne: Cambridge Scholars Publishing.

Zhao, Y. (2014). *Who's afraid of the big bad dragon? Why China has the best (and worst) education system in the world*. San Francisco: Jossey-Bass.

APPENDIX-1. Adapted Version of the ESL Composition Profile

| |
|---|
| Content:<br><br>8-6: Good to average: relevant to the given topic, knowledgeable<br>5-3: Fair to poor: mostly relevant to the given topic, some knowledge of the topic<br>2-1: Very poor: partially relevant to the given topic, limited or no knowledge of the given topic |
| Organization:<br><br>4-3: Good to average: well-organized, in accordance with the given style<br>2 : Fair: loosely organized, but still in accordance with the given style<br>1 : Poor: loosely organized and different from the given style |
| Vocabulary:<br><br>4-3: Good to average: appropriate use of words<br>2 : Fair: limited use of words<br>1 : Poor: no word mastery at all, or not enough to evaluate |
| Grammar:<br><br>4-3: Good to average: few errors of grammar<br>2 : Fair: some errors of grammar<br>1 : Poor: no mastery of grammar at all, or not enough to evaluate |

APPENDIX-2. Consent Form for the Participants (Teacher-raters)

ARAŞTIRMA GÖNÜLLÜ KATILIM FORMU

Değerli Meslektaşlarım,

"Bursa Uludağ Üniversitesi Yabancı Diller Yüksekokulu İngilizce Hazırlık Okulu'nda İngilizce Yazılı Anlatım Becerisini Ölçmek Üzere Geliştirilen Çok Boyutlu Notlandırma Ölçeğinin Güvenirlik ve Geçerliliği" başlıklı bu doktora tez araştırma çalışması, Bursa Uludağ Üniversitesi Yabancı Diller Yüksekokulu İngilizce Hazırlık Okulu'nda her akademik yıl sonunda uygulanan İngilizce yeterlik sınavının yazılı anlatım bölümünde öğrencilerden yazmaları istenen kompozisyonun değerlendirilmesi için kuramsal ve deneysel olarak geçerliği ve güvenirliği sağlanacak çok boyutlu bir notlandırma ölçeği geliştirme amacını taşımaktadır. Çalışma, Prof. Dr. Fatma Hülya Özcan danışmanlığında Okutman Aliye Evin Yörüdü tarafından yürütülmekte ve sonuçları ile Bursa Uludağ Üniversitesi Yabancı Diller Yüksekokulu İngilizce Yeterlik sınavının önemli bileşenlerinden birini oluşturan yazılı anlatım bölümünün geçerlik ve güvenirliğini artırmayı amaçlamaktadır.

- Bu çalışmaya katılımınız gönüllülük esasına dayanmaktadır.
- Çalışmanın amacı doğrultusunda, üç aşamada katılımınız beklenmektedir. Birinci aşamada hâlihazırda kullanılmakta olan çözümsel ölçeğe dair görüşleriniz açık uçlu sorulardan oluşan bir anket yardımıyla alınacaktır. İkinci aşamada, 2017-2018 akademik yılı sonunda uygulanacak olan Bursa Uludağ Üniversitesi Yabancı Diller Yüksekokulu İngilizce Yeterlik sınavı yazılı anlatım bölümünde öğrencilerden yazmaları beklenen 50 kompozisyon gerekli izinler alındıktan sonra rasgele olarak seçilecek ve sizden bu 50 kompozisyonu çalışmayı yürüten Öğretim Görevlisi Aliye Evin Yörüdü tarafından kuramsal ve deneysel olarak geçerlik ve güvenirliği kanıtlanmak üzere geliştirilecek çok boyutlu notlandırma ölçeğini kullanarak değerlendirmeniz beklenecektir. Üçüncü ve son aşamada ise çok boyutlu notlandırma ölçeğinin kullanımına dair görüşleriniz yine açık uçlu bir anket yardımı ile alınacaktır.

- İsminizi yazmak ya da kimliğinizi açığa çıkaracak bir bilgi vermek zorunda değilsiniz/araştırmada katılımcıların isimleri gizli tutulacaktır.
- Araştırma kapsamında toplanan veriler, sadece bilimsel amaçlar doğrultusunda kullanılacak, araştırmanın amacı dışında ya da bir başka araştırmada kullanılmayacak ve gerekmesi halinde, sizin (yazılı) izniniz olmadan başkalarıyla paylaşılmayacaktır.
- İstemeniz halinde sizden toplanan verileri inceleme hakkınız bulunmaktadır.
- Sizden toplanan veriler güvenli bir ortamda korunacak ve araştırma bitiminde arşivlenecek veya imha edilecektir.
- Veri toplama sürecinde/süreçlerinde size rahatsızlık verebilecek herhangi bir soru/talep olmayacaktır. Yine de katılımınız sırasında herhangi bir sebepten rahatsızlık hissederseniz çalışmadan istediğiniz zamanda ayrılabileceksiniz. Çalışmadan ayrılmanız durumunda sizden toplanan veriler çalışmadan çıkarılacak ve imha edilecektir.

Gönüllü katılım formunu okumak ve değerlendirmek üzere ayırdığınız zaman için teşekkür ederim. Çalışma hakkındaki sorularınızı Bursa Uludağ Üniversitesi Yabancı Diller Yüksekokulu İngilizce Hazırlık Programı'ndan Aliye Evin Yörüdü' ye yöneltebilirsiniz.

Araştırmacı Adı :
Adres :
İş Tel :
Cep Tel :

Bu çalışmaya tamamen kendi rızamla, istediğim takdirde çalışmadan ayrılabileceğimi bilerek verdiğim bilgilerin bilimsel amaçlarla kullanılmasını kabul ediyorum.
*(Lütfen bu formu doldurup imzaladıktan sonra veri toplayan kişiye veriniz.)*

Katılımcı Ad ve Soyadı:
İmza:
Tarih:

APPENDIX-3. Demographic Information on the Participants in Phase 1 of the Study

| Instructor | Gender | Native/Nonnative | Age | Education | Active Period of Teaching |
|---|---|---|---|---|---|
| Participant 1 | Female | Non-native | 30-39 | MA in Women Studies | 9 |
| Participant 2 | Female | Non-native | 30-39 | MA in ELT | 12 |
| Participant 3 | Female | Non-native | 30-39 | BA in English Language and Literature | 16 |
| Participant 4 | Female | Non-native | 30-39 | BA in ELT | 16 |
| Participant 5 | Female | Non-native | 30-39 | MA in ELT | 17 |
| Participant 6 | Female | Non-native | 30-39 | BA in ELT | 10 |
| Participant 7 | Female | Non-native | 30-39 | MA in ELT | 15 |
| Participant 8 | Female | Non-native | 40-49 | MA in ELT | 19 |
| Participant 9 | Male | Non-native | 30-39 | BA in ELT | 10 |
| Participant10 | Female | Non-native | 40-49 | MA in ELT | 17 |
| Participant11 | Female | Non-native | 30-39 | BA in ELT | 16 |
| Participant12 | Female | Non-native | 30-39 | MA in ELT | 17 |
| Participant13 | Female | Non-native | 40-49 | BA in ELT | 20 |
| Participant14 | Female | Non-native | 40-49 | MA in Translation Studies | 19 |
| Participant15 | Male | Non-native | 30-39 | MA in ELT | 18 |
| Participant16 | Female | Non-native | 30-39 | BA in ELT | 16 |
| Participant17 | Male | Non-native | 50+ | MA in Educational Management | 33 |
| Participant18 | Female | Non-native | 40-49 | BA in Translation Studies | 16 |
| Participant19 | Female | Non-native | 30-39 | MA in ELT | 7 |

| | | | | | |
|---|---|---|---|---|---|
| Participant20 | Female | Non-native | 30-39 | MA in ELT | 13 |
| Participant21 | Female | Non-native | 30-39 | BA in ELT | 15 |
| Participant22 | Female | Non-native | 30-39 | MA in ELT | 17 |
| Participant23 | Female | Non-native | 40-49 | BA in ELT | 26 |
| Participant24 | Male | Non-native | 20-29 | MA in ELT | 5 |

APPENDIX-4. Open-ended Questionnaire (Phase 1)

RATERS' PERSPECTIVES ON THE WRITING RUBRIC CURRENTLY USED IN THE PROFICIENCY EXAMINATION

PART A

*Please, put a tick (√) inside the suitable box.*

1. Male □ Female □

2. Native Speaker □ Non-native Speaker □

3. Age: □ 20-29 □ 30-39 □40-49 □ 50+

4. Educational Background:

| Degree | Name of the University | Department |
|---|---|---|
| Bachelor (Lisans) | | |
| Master (Yuksek Lisans) | | |
| Doctor (Doktora) | | |

5. Active teaching period: _____ years

6. Active teaching period at Bursa Uludağ University, Preparatory School: _____ years

PART B

I. *Please, answer the following questions.*

1. What are the strengths and weaknesses of the writing rubric used currently in the proficiency examination?

*Strengths:*

_____

_____

_____

_____

_____

*Weaknesses:*

2. Below is a list of categories that may be part of a writing rubric according to the relevant literature. Put a tick *(√)* by the categories that you think should be included in a writing rubric and explain your reason(s) in the space provided below or on the back of the page.

| √ | *CATEGORIES* | *REASONS* |
|---|---|---|
| | Content: meaningfulness, relevance, and logical development of ideas | |
| | Organization: development of a thesis statement with ideas and details with an introductory, three body, and a conclusion paragraphs | |
| | Argumentation: multiple perspectives covering counter argument(s) and refutation of it | |
| | Grammar: accurate use of structures | |
| | Vocabulary: accurate use of words/phrases | |
| | Cohesion: accurate use of transitions at sentence level | |
| | Coherence: fluency i.e. links between ideas to create meaning at textual level (not sentence level) | |
| | Mechanics: punctuation, capitalization, and spelling | |
| | Length: amount of language produced within the specified word limit | |

Other (Please specify others here)

3. There are three levels in each category of the writing rubric used currently (Good, Fair, and Poor).

*These three levels in each category help me assess the students' writings effectively.*

To what extent do you agree with this statement? Dou you think these three levels are adequate? Please explain your reason(s).

4. Are there any categories in which the wording of the descriptors needs to be changed? If yes, please explain your reason(s).

Content:

8-6: Good to average: relevant to the given topic, knowledgeable

5-3: Fair to poor: mostly relevant to the given topic, some knowledge of the topic

2-1: Very poor: partially relevant to the given topic, limited or no knowledge of the given topic

Organization:

4-3: Good to average: well-organized, in accordance with the given style

2 : Fair: loosely organized, but still in accordance with the given style

1 : Poor: loosely organized and different from the given style

Vocabulary:

4-3: Good to average: appropriate use of words

2 : Fair: limited use of words

1 : Poor: no word mastery at all, or not enough to evaluate

Grammar:

4-3: Good to average: few errors of grammar

2 : Fair: some errors of grammar

1 : Poor: no mastery of grammar at all, or not enough to evaluate

5. The 20-point is distributed as follows with different weightings: Content: 8, Organization: 4, Vocabulary: 4, and Grammar: 4. What do you think about this uneven weighting?

6. Are there any categories that you have found difficult to apply? If yes, please explain your reason(s).

_____

_____

II. *Please, answer the following questions.*

1. To what extent does the writing rubric used currently in the proficiency examination facilitate fair assessment of students' written work? Please, rate it from 1 to 5.

   1 ___ 2 ___ 3 ___ 4 ___ 5 ___

2. How confident do you feel in applying the writing rubric? Please, rate your confidence level from 1 to 5.

   1 ___ 2 ___ 3 ___ 4 ___ 5 ___

APPENDIX-5. Uludağ Üniversitesi Etik Kurul İzni

Ana- Üni. Evrak Tarih ve Sayısı:04/04/2018-E.22414

T.C.
**ULUDAĞ ÜNİVERSİTESİ REKTÖRLÜĞÜ**
**Genel Sekreterlik**

Sayı: 26468960-044/12368 04/04/2018
Konu: Aliye Evin YÖRÜDÜ 'nün Uygulama İzini

İlgi : 15.03.201 8 tarihli ve 63784619-605.01-E.34456 sayılı yazınız.

İlgi yazınızda bahsi geçen Universitemiz Eğitim Bilimleri Enstitüsü Yabancı Diller Eğitimi Anabilim Dalı İngilizce Oğretmenliği Doktora Programı öğrencisi Aliye Evin YÖRÜDÜ'nün "Uludağ Üniversitesi Yabancı Diller Yüksekokulu İngilizce Hazırlık Okulunda İngilizce Yazıl! Anlatım Becerisini Ölçmek Üzere Geliştirilen Çok Boyutlu Notlandırma Olçeğinin Güvenirlik ve Geçerliliği" başlıklı tez çalışması kapsamında Üniversitemiz Yabancı Diller Yüksekokulu İngilizce Hazırlık Programı öğrenci ve öğretim elamanlarına anket uygulama isteği Rektörlüğümüzce uygun bulunmuştur.

Bilgilerinize arz ve rica ederim.

imza

Prof. Dr. Eray ALPER
Rektör a.
Rektör Yardımcısı

Bu Belge, 5070 sayılı Kanun hükümlerine uygun olarak elektronik imza ile imzalanmışım.

U.Ü. Rektörlüğü Görükle Kampusu 16059 Nilüfer/BURSA Bilgi İçin:Çiğdem ŞENOL
Tel : 0224 294 OO 86 Faks: 0224 294 00 37 şef e-posta : uugs@uludag.edu.tr Elektronik Ağ:
www.uludag.edu.tr Tel : 0224 294 OO 38Bu belge UDOS İle hazırlanmıştır, Teyit için:
https://udos.uludag.edu.tr/teyiÜ?YLJNs0au9k2Bg22clsuczg

APPENDIX-6. Demographic Information on the Participants in Phase 2 of the Study

| Instructor | Gender | Native/Nonnative | Age | Education | Active Period of Teaching |
|---|---|---|---|---|---|
| Participant 1 | Female | Non-native | 30-39 | MA in ELT | 16 |
| Participant 2 | Female | Non-native | 30-39 | MA in ELT | 14 |
| Participant 3 | Female | Non-native | 30-39 | BA in ELT | 16 |
| Participant 4 | Female | Non-native | 30-39 | BA in ELT | 18 |
| Participant 5 | Female | Non-native | 30-39 | MA in ELT | 17 |

APPENDIX-7. Information Document for the Experts (Phase 2)

Dear Professor,

I am an English instructor at Bursa Uludağ University, School of Foreign Languages, Intensive English Program (BUU-SFL-IEP) and a PhD candidate at Anadolu University, Faculty of Education, ELT Department.

I am in the process of writing my PhD dissertation titled 'An Alternative Multi-Trait Rubric for the Performance-Based Assessment of EFL Writing Proficiency at BUU-SFL-IEP'. The overarching aim of my mixed-methods PhD dissertation is to develop an alternative theoretically-based and an empirically-validated multi-trait writing rubric for the performance-based assessment of writing proficiency at BUU English proficiency examination (EPE) administered at the end of each academic year as an exit exam.

BUU-EPE comprises four sections: language use, listening, reading, and writing. The writing section is 50 min long and offers test takers a choice of three argumentative essay prompts. They choose one and are expected to write between 200-250 words. The prompts require test takers to give their opinion on a statement, justify their opinion using supporting details, present the counter-argument and refute it. The writing section of the exam which is an example of performance-based assessment writing proficiency comprises 20% of the test-takers' total score, making it high-stakes. A context-sensitive rubric validation where the type of rubric to be used needs to be judiciously decided is a requirement for institutions that use high-stakes performance assessment of writing proficiency. BUU-SFL-IEP is also in need for such a validation process for the assessment of writing performance carried out at EPE.

An overview of the process, which consists of five phases, is presented below:

Phase 1: Exploration of raters' perspectives on the current writing rubric, i.e. the adapted version of ESL Composition Profile and their expectations from an alternative writing rubric,

Phase 2: Development of a new rubric with the guidance of the relevant literature, expected learning outcomes of the writing course, raters' perspectives, and expert opinion,

Phase 3: Trial and refinement of draft rubric and the open-ended questionnaire as a pilot scheme,

Phase 4: Psychometric analysis of the new rubric through Many Faceted Rasch Measurement (MFRM),

Phase 5: Exploration of raters' perspectives on the new rubric.

I am about to complete the second phase of the process. I would be really grateful if you could support this endeavor with your invaluable opinions on the draft rubric regarding the following issues:

- The content and number of categories, i.e. the writing constructs to be assessed
- The wordings of descriptors,
- The number of score/band levels, and
- The weighting.

Thank you in advance.

Öğr. Gör. Aliye Evin Yörüdü

APPENDIX-8. The First Draft of the New Rubric

| |
|---|
| *Content*<br>*4. Very Good:* Essay addresses the assigned topic; ideas are concrete and thoroughly developed; no extraneous material; essay reflects careful thought<br>*3. Good*: essay addresses the issues but misses some points; ideas could be more fully developed; some extraneous material is present<br>*2. Moderate:* Development of ideas not complete or essay is somewhat off the topic; paragraphs aren't divided exactly right<br>*1. Poor*: Ideas incomplete; essay does not reflect careful thinking or was hurriedly written; inadequate effort in area of content<br>*0. Very Poor*: Essay is completely inadequate and does not reflect college level work; no apparent effort to consider the topic carefully |
| *Organization*<br>*4. Very Good*: Appropriate title, effective introductory paragraph, topic is stated in a well-expressed thesis statement, leads to body, transitional expressions used; supporting evidence given for generalizations; counterargument with a sound refutation; conclusion logical & complete<br>*3. Good:* Adequate title, introduction & conclusion; body of essay is acceptable but some evidence may be lacking, some ideas aren't fully developed; counterargument with an acceptable refutation, sequence is logical but transitional expressions may be absent or misused<br>*2. Moderate*: Mediocre or scant introduction, or conclusion; problems with the order of ideas in body; the generalizations may not be fully supported by the evidence given; counterargument with weak or no refutation; problems of organization interfere<br>*1. Poor:* Shaky or minimally recognizable introduction; organization can barely be seen; severe problems of ordering ideas; lack of supporting data<br>*0. Very Poor:* Writer has not made any effort to organize the composition (could not be outlined by reader) |

*Grammar*

*4. Very Good*: Proficient in English grammar; correct and appropriate use of compound and complex sentences, prepositions, modals, articles, verb forms, and tense sequencing; no fragments or run-on sentences

*3. Good:* Rather proficient in English grammar; some grammar problems don't influence communication, although the reader is aware of them; no fragments or run-on sentences

*2. Moderate*: Ideas getting through to the reader, but grammar problems are apparent and have a negative effect on communication; run-on sentences or fragments present

*1. Poor*: Numerous serious grammar problems interfere with communication of the writer's ideas; grammar review of some areas are clearly needed; difficult to read sentences

*0. Very Poor*: Severe grammar problems interfere greatly with the message; reader can't understand what the writer is trying to say; unintelligible sentence structure

*Vocabulary*

*4. Very Good:* Precise vocabulary usage; use of parallel structures; concise; register good

*3. Good*: Attempts variety; good vocabulary; not wordy; register acceptable; style fairly concise

*2. Moderate*: Some vocabulary misused; lacks awareness of register; may be too wordy

*1. Poor*: Poor extension of ideas; problems of vocabulary; lacks variety

*0. Very Poor:* Inappropriate use of vocabulary; no concept of register or variety

*Punctuation, Spelling, & Mechanics*

*4. Very Good*: Correct use of English writing conventions; all needed capitals, paragraphs indented, almost no punctuation and spelling errors; very neat

*3. Good*: Some problems with writing conventions or punctuation; occasional spelling errors; paper is neat and legible

*2. Moderate*: Uses general writing conventions but has errors; spelling problems distract reader; punctuation errors interfere with ideas

*1. Poor*: Serious problems with format of paper; parts of essay not legible; errors in sentence-final punctuation; unacceptable to educated readers

*0. Very Poor*: Complete disregard of English writing conventions; paper illegible; obvious capitals missing, no margins, severe spelling problems

APPENDIX-9. The Second Draft of the New Rubric

| |
|---|
| *Content*<br>*4. Very Good:* Essay addresses the assigned topic; ideas are concrete and thoroughly developed; supporting evidence provided; no extraneous material; essay reflects carefully planned thought<br>*3. Good*: Essay addresses the assigned topic but misses some points; ideas could be more fully developed; some evidence may be lacking; some extraneous material is present<br>*2. Moderate:* Development of ideas not complete or essay is somewhat off the topic; paragraphs aren't divided exactly right; the main ideas may not be fully supported by the evidence given<br>*1. Poor*: Ideas incomplete; essay does not reflect careful thinking or was hurriedly written; lack of supporting data; inadequate effort in area of content<br>*0. Very Poor*: Essay is completely inadequate and does not reflect college level work; no apparent effort to consider the topic carefully |
| *Organization*<br>*4. Very Good*: An informative title that covers the topic, effective introductory paragraph, topic is stated in a clear thesis statement, leads to body with two full paragraphs, appropriate transitional expressions used; counterargument with a sound refutation; conclusion logical & complete<br>*3. Good:* Adequate title, introduction & conclusion; body of essay is acceptable with two paragraphs that aren't fully developed; counterargument with an acceptable refutation, sequence is logical but transitional expressions may be absent, misused, or overused<br>*2. Moderate*: Mediocre or scant introduction, body, or conclusion; problems with the order of ideas in body; counterargument with weak or no refutation; problems of organization interfere<br>*1. Poor:* Shaky or minimally recognizable introduction; organization<br>can barely be seen; severe problems of ordering ideas<br>*0. Very Poor:* Writer has not made any effort to organize the composition (could not be outlined by reader) |
| *Grammar*<br>*4. Very Good*: Proficient in English grammar with almost no grammar errors; all/majority of simple, compound, and complex sentences are correct and appropriate; mostly accurate and appropriate use of prepositions, modals, articles, verb forms, and tense sequencing; no fragments or run-on sentences<br>*3. Good:* Rather proficient in English grammar; few grammar problems which don't distort meaning; no fragments or run-on sentences<br>*2. Moderate*: Somewhat proficient in English grammar; grammar problems distort meaning; run-on sentences or fragments present |

| |
|---|
| *1. Poor*: Limited proficiency in English grammar; grammar problems hinder meaning; difficult to understand sentences<br>*0. Very Poor* Unintelligible sentence structure |
| *Vocabulary*<br>*4. Very Good:* All/majority of vocabulary use is appropriate, accurate, and varied<br>*3. Good*: Appropriate and accurate use of vocabulary with few errors; attempts variety<br>*2. Moderate*: Some vocabulary misused; repetitive use of vocabulary with little variety<br>*1. Poor*: Serious problems of vocabulary; lacks variety<br>*0. Very Poor:* Inappropriate and inaccurate use of vocabulary; no concept of variety |
| *Punctuation, Spelling, & Mechanics*<br>*4. Very Good*: Correct use of English writing conventions; all needed capitals, paragraphs indented, almost no punctuation and spelling errors; very neat<br>*3. Good*: Some problems with writing conventions or punctuation; occasional spelling errors; paper is neat<br>*2. Moderate*: Uses general writing conventions but has errors; spelling problems distract reader; punctuation errors interfere with ideas<br>*1. Poor*: Serious problems with format of paper; parts of essay not legible; errors in sentence-final punctuation; unacceptable to educated readers<br>*0. Very Poor*: Complete disregard of English writing conventions; paper illegible; obvious capitals missing, no margins, severe spelling problems |

APPENDIX-10. The Third Draft (Refinements of the First Expert)

| |
|---|
| *Content*<br>*4. Very Good:* Essay addresses the assigned topic; ideas are concrete and thoroughly developed; supporting evidence provided; no extraneous material; essay reflects carefully planned thought<br>*3. Good*: Essay addresses the assigned topic but misses some points; ideas could be more fully developed; some evidence may be lacking; some extraneous material is present<br>*2. Moderate:* Development of ideas not complete or essay is somewhat off the topic; paragraphs aren't divided exactly right; the main ideas may not be fully supported by the evidence given<br>*1. Poor*: Ideas incomplete; essay does not reflect careful thinking or was hurriedly written; lack of supporting data; inadequate effort in area of content<br>*0. Very Poor*: Essay is completely inadequate and does not reflect college level work; no apparent effort to consider the topic carefully |
| *Organization*<br>*4. Very Good*: An informative title that covers the topic, effective introductory paragraph, topic is stated in a clear thesis statement, leads to body with two full paragraphs, appropriate transitional expressions used; counterargument with a sound refutation; conclusion logical & complete<br>*3. Good:* Adequate title, introduction & conclusion; body of essay is acceptable with two paragraphs that aren't fully developed; counterargument with an acceptable refutation, sequence is logical but transitional expressions may be absent, misused, or overused<br>*2. Moderate*: Mediocre or scant introduction, body, or conclusion; problems with the order of ideas in body; counterargument with weak or no refutation; problems of organization interfere<br>*1. Poor:* Shaky or minimally recognizable introduction; organization<br>can barely be seen; severe problems of ordering ideas<br>*0. Very Poor:* Writer has not made any effort to organize the composition (could not be outlined by reader) |
| *Grammar*<br>*4. Very Good*: Proficient in English grammar with almost no grammar errors; all/majority of simple, compound, and complex sentences are correct and appropriate; mostly accurate and appropriate use of prepositions, modals, articles, verb forms, and tense sequencing; no fragments or run-on sentences<br>*3. Good:* Rather proficient in English grammar; few grammar problems which don't distort meaning; no fragments or run-on sentences<br>*2. Moderate*: Somewhat proficient in English grammar; grammar problems distort meaning; run-on sentences or fragments present<br>*1. Poor*: Limited proficiency in English grammar; grammar problems hinder meaning; difficult to understand sentences |

| |
|---|
| *0. Very Poor* Unintelligible sentence structure |
| *Vocabulary*<br>*4. Very Good:* All/majority of vocabulary use is appropriate, accurate, and varied<br>*3. Good*: Appropriate and accurate use of vocabulary with few errors; attempts variety<br>*2. Moderate*: Some vocabulary misused; repetitive use of vocabulary with little variety<br>*1. Poor*: Serious problems of vocabulary; lacks variety<br>*0. Very Poor:* Inappropriate and inaccurate use of vocabulary; no concept of variety |
| *Punctuation, Spelling, & Mechanics*<br>*4. Very Good*: Correct use of English writing conventions; all needed capitals, paragraphs indented, almost no punctuation and spelling errors; very neat<br>*3. Good*: Some problems with writing conventions or punctuation; occasional spelling errors; paper is neat<br>*2. Moderate*: Uses general writing conventions but has errors; spelling problems distract reader; punctuation errors interfere with ideas<br>*1. Poor*: Serious problems with format of paper; parts of essay not legible; errors in sentence-final punctuation; unacceptable to educated readers<br>*0. Very Poor*: Complete disregard of English writing conventions; paper illegible; obvious capitals missing, no margins, severe spelling problems |

APPENDIX-11. The Fourth Draft (Refinements of the Second Expert)

| |
|---|
| *Content*<br>*4. Exceeds Expectations:* Essay addresses the assigned topic; ideas are concrete and thoroughly developed with supporting evidence provided; counterargument with a sound refutation; no extraneous material; essay reflects carefully planned thought<br>*3. Meets Expectations:* Essay addresses the assigned topic but misses some points; ideas could be more fully developed as some evidence may be lacking; counterargument with an acceptable refutation; some extraneous material is present<br>*2. Approaches Expectations:* Development of ideas not complete or essay is somewhat off the topic; the main ideas may not be fully supported by the evidence given; counterargument with weak or no refutation<br>*1. Needs Development:* Ideas incomplete due to lack of supporting data; essay does not reflect careful thinking or was hurriedly written; inadequate effort in area of content<br>*0. Off Topic/Did Not Try:* Essay is completely inadequate; no apparent effort to consider the topic carefully |
| *Organization*<br>*4. Exceeds Expectations:* An informative title that covers the topic; effective introductory paragraph, topic is stated in a clear thesis statement, leads to body with three full paragraphs; appropriate transitional expressions used; conclusion logical & complete<br>*3. Meets Expectations:* Adequate title, introduction & conclusion; body of essay is acceptable with three paragraphs that may not be fully developed; sequence is logical but few transitional expressions may be absent, misused, or overused<br>*2. Approaches Expectations:* Mediocre or scant introduction, body, or conclusion; paragraphs aren't divided exactly right; problems with the order of ideas in body; problems of organization interfere<br>*1. Needs Development:* Shaky or minimally recognizable introduction; organization can barely be seen; severe problems of ordering ideas<br>*0. Off Topic/Did Not Try:* Writer has not made any effort to organize the composition (could not be outlined by reader) |
| *Grammar*<br>*4. Exceeds Expectations:* Proficient in English grammar with almost no grammar errors; all/majority of simple, compound, and complex sentences are accurate and appropriate; mostly accurate and appropriate use of prepositions, modals, articles, verb forms, and tense sequencing; no fragments or run-on sentences<br>*3. Meets Expectations:* Rather proficient in English grammar; few grammar problems which don't distort meaning; almost no fragments or run-on sentences |

| |
|---|
| *2. Approaches Expectations:* Somewhat proficient in English grammar; grammar problems distort meaning; fragments or run-on sentences present<br>*1. Needs Development:* Limited proficiency in English grammar; grammar problems hinder meaning; difficult to understand sentences<br>*0. Off Topic/Did Not Try:* Unintelligible sentence structure |
| *Vocabulary*<br>*4. Exceeds Expectations:* All/majority of vocabulary use is appropriate, accurate, and varied<br>*3. Meets Expectations:* Appropriate and accurate use of vocabulary with few errors; attempts variety<br>*2. Approaches Expectations:* Some vocabulary misused; repetitive use of vocabulary with little variety<br>*1. Needs Development:* Serious problems of vocabulary; lacks variety<br>*0. Off Topic/Did Not Try:* Inappropriate and inaccurate use of vocabulary; no concept of variety |
| *Punctuation, Spelling, & Mechanics*<br>*4. Exceeds Expectations:* Accurate use of English writing conventions; all needed capitals, paragraphs indented, almost no punctuation and spelling errors; very neat<br>*3. Meets Expectations:* Some problems with writing conventions or punctuation; occasional spelling errors; paper is neat<br>*2. Approaches Expectations:* Uses general writing conventions but has errors; spelling problems distract reader; punctuation errors interfere with ideas<br>*1. Needs Development:* Serious problems with format of paper; errors in sentence-final punctuation; unacceptable to educated readers<br>*0. Off Topic/Did Not Try:* Complete disregard of English writing conventions; obvious capitals missing, no margins, severe spelling problems |

APPENDIX-12. Demographic Information on the Participants in Phase 3 of the Study

| Instructor | Gender | Native/Nonnative | Age | Education | Active Period of Teaching |
|---|---|---|---|---|---|
| Participant 1 | Female | Non-native | 30-39 | MA in ELT | 12 |
| Participant 2 | Female | Non-native | 30-39 | MA in Gender Studies | 11 |
| Participant 3 | Female | Non-native | 30-39 | MA in ELT | 17 |
| Participant 4 | Female | Non-native | 40-49 | MA in ELT | 18 |
| Participant 5 | Female | Non-native | 50+ | BA in ELT | 28 |

APPENDIX-13. Assessor Guide for the Draft Rubric

CONTENTS



## SECTION 1 PILOTING PROCESS

Aims of the piloting process:

Four aims of the piloting process are listed as follows:

- to trial the new rubric through an initial pilot implementation,
- to identify possible weaknesses in the use of it, and refine the rubric based on the feedback that will be received from the raters as carried out in Knoch (2007) and Hattingh (2009)
- to pilot the open-ended questionnaire that will be used to explore the perceptions of the raters on the efficacy of the draft rubric and
- to discover the potential problems that may exist in the open-ended questionnaire, such as the clarity of the items. (Zacharias, 2012: 71).

-

SECTION 2 ASSESSOR GUIDE FOR THE DRAFT RUBRIC

Strands of the rubric design:

Four sources of data has been consulted during the rubric design process:

- contextual needs of BUU-SFL-IEP,
- expectations of the raters (that are the instructors at BUU-SFL-IEP) from a writing rubric,
- relevant literature, and
- expert guidance (outside and in-house).

Areas of the rubric design:

Four areas of rubric desiat are considered:

- categories (content and number),
- wording of descriptors,
- labels and number of band levels, and
- weighting.

1. Categories

Five categories in the new rubric are:

- content,
- organization,
- grammar,
- vocabulary, and
- punctuation, spelling, and mechanics.

Content:

Aspects that are considered under *content*:

- topicality,
- development of ideas,
- counterargument and refutation, and
- thoughtfulness/effort.

*Content*

*4. Exceeds Expectations:* Essay addresses the assigned topic; ideas are concrete and thoroughly developed with supporting evidence provided; counterargument with a sound refutation; no extraneous material; essay reflects carefully planned thought

*3. Meets Expectations:* Essay addresses the assigned topic but misses some points; ideas could be more fully developed as some evidence may be lacking; counterargument with an acceptable refutation; some extraneous material is present

*2. Approaches Expectations:* Development of ideas not complete or essay is somewhat off the topic; the main ideas may not be fully supported by the evidence given; counterargument with weak or no refutation

*1. Needs Development:* Ideas incomplete due to lack of supporting data; essay does not reflect careful thinking or was hurriedly written; inadequate effort in area of content

*0. Off Topic/Did Not Try:* Essay is completely inadequate; no apparent effort to consider the topic carefully

Organization:

Aspects that are considered under *organization*:

- title,
- paragraph structure,
- thesis statement,
- transitions,
- support for arguments, and
- conclusion.

*Organization*

*4. Exceeds Expectations:* An informative title that covers the topic; effective introductory paragraph, topic is stated in a clear thesis statement, leads to body with three full paragraphs; appropriate transitional expressions used; conclusion logical & complete

*3. Meets Expectations:* Adequate title, introduction & conclusion; body of essay is acceptable with three paragraphs that may not be fully developed; sequence is logical but few transitional expressions may be absent, misused, or overused

*2. Approaches Expectations:* Mediocre or scant introduction, body, or conclusion; paragraphs aren't divided exactly right; problems with the order of ideas in body; problems of organization interfere

*1. Needs Development:* Shaky or minimally recognizable introduction; organization can barely be seen; severe problems of ordering ideas

*0. Off Topic/Did Not Try:* Writer has not made any effort to organize the composition (could not be outlined by reader)

Grammar:

Aspects that are considered under *grammar*:

- accurate and appropriate use of sentence structure (simple, compound, and complex),

- accurate and appropriate use of prepositions, modals, articles, verb forms, and tense sequences, and
- absence/presence of fragments or run-on sentences.

*Grammar*

*4. Exceeds Expectations:* Proficient in English grammar with almost no grammar errors; all/majority of simple, compound, and complex sentences are accurate and appropriate; mostly accurate and appropriate use of prepositions, modals, articles, verb forms, and tense sequencing; no fragments or run-on sentences

*3. Meets Expectations:* Rather proficient in English grammar; few grammar problems which don't distort meaning; almost no fragments or run-on sentences

*2. Approaches Expectations:* Somewhat proficient in English grammar; grammar problems distort meaning; run-on sentences or fragments present

*1. Needs Development:* Limited proficiency in English grammar; grammar problems hinder meaning; difficult to understand sentences

*0. Off Topic/Did Not Try:* Unintelligible sentence structure

Vocabulary:

Aspects that are considered under *vocabulary*:

- appropriate,
- accurate, and
- varied use of words (form and meaning) and collocations.

*Vocabulary*

*4. Exceeds Expectations:* All/majority of vocabulary use is appropriate, accurate, and varied

*3. Meets Expectations:* Appropriate and accurate use of vocabulary with few errors; attempts variety

*2. Approaches Expectations:* Some vocabulary misused; repetitive use of vocabulary with little variety

*1. Needs Development:* Serious problems of vocabulary; lacks variety

*0. Off Topic/Did Not Try:* Inappropriate and inaccurate use of vocabulary; no concept of variety

Punctuation, Spelling, and Mechanics:

Aspects that are considered under *punctuation, spelling, and mechanics:* general English writing conventions including

- punctuation
- capitalization,

- indentation,
- spelling, and
- neatness.
-

| |
|---|
| *Punctuation, Spelling, & Mechanics*<br>*4. Exceeds Expectations:* Correct use of English writing conventions; all needed capitals, paragraphs indented, almost no punctuation and spelling errors; very neat<br>*3. Meets Expectations:* Some problems with writing conventions or punctuation; occasional spelling errors; paper is neat<br>*2. Approaches Expectations:* Uses general writing conventions but has errors; spelling problems distract reader; punctuation errors interfere with ideas<br>*1. Needs Development:* Serious problems with format of paper; errors in sentence-final punctuation; unacceptable to educated readers<br>*0. Off Topic/Did Not Try:* Complete disregard of English writing conventions; obvious capitals missing, no margins, severe spelling problems |

2. Wording of descriptors:

Descriptors with *concrete* and *objective* formulation style were aimed at.

3. Labels and number of band levels:

Five band levels with new labels are as follows:

4. Exceeds expectations

3. Meets expectations

2. Approaches expectations

1. Needs development

0. Off topic/Did not try

4. Weighting:

The 20-point is distributed equally for weighing:

- Content: 4,
- Organization: 4,
- Vocabulary: 4,
- Grammar: 4, and
- Punctuation, Spelling, and Mechanics: 4.

APPENDIX-14. Consent for Students whose Essays to Be Used

ARAŞTIRMA GÖNÜLLÜ KATILIM FORMU

Değerli Öğrenciler,

"Bursa Uludağ Üniversitesi Yabancı Diller Yüksekokulu İngilizce Hazırlık Okulu'nda İngilizce Yazılı Anlatım Becerisini Ölçmek Üzere Geliştirilen Çok Boyutlu Notlandırma Ölçeğinin Güvenirlik ve Geçerliliği" başlıklı bu doktora tez araştırma çalışması, Bursa Uludağ Üniversitesi Yabancı Diller Yüksekokulu İngilizce Hazırlık Okulu'nda her akademik yıl sonunda uygulanan İngilizce yeterlik sınavının yazılı anlatım bölümünde öğrencilerden yazmaları istenen kompozisyonun değerlendirilmesi için kuramsal ve deneysel olarak geçerliği ve güvenirliği sağlanacak çok boyutlu bir notlandırma ölçeği geliştirme amacını taşımaktadır. Çalışma, Prof. Dr. Fatma Hülya Özcan danışmanlığında Öğretim Görevlisi Aliye Evin Yörüdü tarafından yürütülmekte ve sonuçları ile Bursa Uludağ Üniversitesi Yabancı Diller Yüksekokulu İngilizce Yeterlik sınavının önemli bileşenlerinden birini oluşturan yazılı anlatım bölümünün geçerlik ve güvenirliğini artırmayı amaçlamaktadır.

- Bu çalışmaya katılımınız gönüllülük esasına dayanmaktadır.
- Çalışmanın amacı doğrultusunda, 2017-2018 akademik yılı sonunda uygulanacak olan Bursa Uludağ Üniversitesi Yabancı Diller Yüksekokulu İngilizce Yeterlik sınavı yazılı anlatım bölümünde sizden yazmanızı beklenen kompozisyonun kullanımı için izniniz istenmektedir.
- İsminizi yazmak ya da kimliğinizi açığa çıkaracak bir bilgi vermek zorunda değilsiniz/araştırmada katılımcıların isimleri gizli tutulacaktır.
- Araştırma kapsamında toplanan veriler, sadece bilimsel amaçlar doğrultusunda kullanılacak, araştırmanın amacı dışında ya da bir başka araştırmada kullanılmayacak ve gerekmesi halinde, sizin (yazılı) izniniz olmadan başkalarıyla paylaşılmayacaktır.
- İstemeniz halinde sizden toplanan verileri inceleme hakkınız bulunmaktadır.
- Sizden toplanan veriler güvenli bir ortamda korunacak ve araştırma bitiminde arşivlenecek veya imha edilecektir.

- Veri toplama sürecinde/süreçlerinde size rahatsızlık verebilecek herhangi bir soru/talep olmayacaktır. Yine de katılımınız sırasında herhangi bir sebepten rahatsızlık hissederseniz çalışmadan istediğiniz zamanda ayrılabileceksiniz. Çalışmadan ayrılmanız durumunda sizden toplanan veriler çalışmadan çıkarılacak ve imha edilecektir.

Gönüllü katılım formunu okumak ve değerlendirmek üzere ayırdığınız zaman için teşekkür ederim. Çalışma hakkındaki sorularınızı Bursa Uludağ Üniversitesi Yabancı Diller Yüksekokulu İngilizce Hazırlık Programı'ndan Aliye Evin Yörüdü' ye yöneltebilirsiniz.

Araştırmacı Adı :
Adres :
İş Tel :
Cep Tel :

Bu çalışmaya tamamen kendi rızamla, istediğim takdirde çalışmadan ayrılabileceğimi bilerek verdiğim bilgilerin bilimsel amaçlarla kullanılmasını kabul ediyorum.

*(Lütfen bu formu doldurup imzaladıktan sonra veri toplayan kişiye veriniz.)*

Katılımcı Ad ve Soyadı:
İmza:
Tarih:

APPENDIX-15. Sample Rating Sheet

*The scores assigned by each rater to each category and total score assigned to each essay*

*Rater:* _____

| Essay No. | Content Score | Organization Score | Grammar Score | Vocabulary Score | Punctuation, Spelling, & Mechanics Score | Total Score |
|---|---|---|---|---|---|---|
| 1. | | | | | | |
| 2. | | | | | | |
| 3. | | | | | | |
| 4. | | | | | | |
| 5. | | | | | | |
| 6. | | | | | | |
| 7. | | | | | | |
| 8. | | | | | | |
| 9. | | | | | | |
| 10. | | | | | | |

APPENDIX-16. Open-Ended Questionnaire (Phase 3)

RATERS' PERSPECTIVES ON THE NEW WRITING RUBRIC DESIGNED TO BE USED IN THE PROFICIENCY EXAMINATION

PART A

*Please, put a tick (√) inside the suitable box.*

1. Male □ Female □

2. Native Speaker □ Non-native Speaker □

3. Age: □ 20-29 □ 30-39 □40-49 □ 50+

4. Educational Background:

| Degree | Name of the University | Department |
|---|---|---|
| Bachelor (Lisans) □ | | |
| Master (Yuksek Lisans) □ | | |
| Doctor (Doktora) □ | | |

5. Active teaching period: _____ years

6. Active teaching period at Bursa Uludağ University, Preparatory School: _____ years

PART B

I. *Please, answer the following questions.*

2. What are the strengths and weaknesses of the new writing rubric designed to be used in the proficiency examination?

*Strengths:*

_____

_____

_____

_____

*Weaknesses:*

2. There are five levels in each category of the writing rubric used currently (Exceeds Expectations, Meets Expectations, Approaches Expectations, and Off Topic/Did Not Try). *These five levels in each category help me assess the students' writings effectively.*

To what extent do you agree with this statement? Dou you think these five levels are adequate? Please explain your reason(s).

4. Are there any categories in which the wording of the descriptors needs to be changed? If yes, please explain your reason(s).

*Content*

*4. Exceeds Expectations:* Essay addresses the assigned topic; ideas are concrete and thoroughly developed with supporting evidence provided; counterargument with a sound refutation; no extraneous material; essay reflects carefully planned thought

*3. Meets Expectations:* Essay addresses the assigned topic but misses some points; ideas could be more fully developed as some evidence may be lacking; counterargument with an acceptable refutation; some extraneous material is present

*2. Approaches Expectations:* Development of ideas not complete or essay is somewhat off the topic; the main ideas may not be fully supported by the evidence given; counterargument with weak or no refutation

*1. Needs Development:* Ideas incomplete due to lack of supporting data; essay does not reflect careful thinking or was hurriedly written; inadequate effort in area of content

*0. Off Topic/Did Not Try:* Essay is completely inadequate; no apparent effort to consider the topic carefully

*Organization*

*4. Exceeds Expectations:* An informative title that covers the topic; effective introductory paragraph, topic is stated in a clear thesis statement, leads to body with three full paragraphs; appropriate transitional expressions used; conclusion logical & complete

*3. Meets Expectations:* Adequate title, introduction & conclusion; body of essay is acceptable with three paragraphs that may not be fully developed; sequence is logical but few transitional expressions may be absent, misused, or overused

*2. Approaches Expectations:* Mediocre or scant introduction, body, or conclusion; paragraphs aren't divided exactly right; problems with the order of ideas in body; problems of organization interfere

*1. Needs Development:* Shaky or minimally recognizable introduction; organization can barely be seen; severe problems of ordering ideas

*0. Off Topic/Did Not Try:* Writer has not made any effort to organize the composition (could not be outlined by reader)

*Grammar*

*4. Exceeds Expectations:* Proficient in English grammar with almost no grammar errors; all/majority of simple, compound, and complex sentences are accurate and appropriate; mostly accurate and appropriate use of prepositions, modals, articles, verb forms, and tense sequencing; no fragments or run-on sentences

*3. Meets Expectations:* Rather proficient in English grammar; few grammar problems which don't distort meaning; almost no fragments or run-on sentences

*2. Approaches Expectations:* Somewhat proficient in English grammar; grammar problems distort meaning; run-on sentences or fragments present

*1. Needs Development:* Limited proficiency in English grammar; grammar problems hinder meaning; difficult to understand sentences

*0. Off Topic/Did Not Try:* Unintelligible sentence structure

*Vocabulary*

*4. Exceeds Expectations:* All/majority of vocabulary use is appropriate, accurate, and varied

*3. Meets Expectations:* Appropriate and accurate use of vocabulary with few errors; attempts variety

*2. Approaches Expectations:* Some vocabulary misused; repetitive use of vocabulary with little variety

*1. Needs Development:* Serious problems of vocabulary; lacks variety

*0. Off Topic/Did Not Try:* Inappropriate and inaccurate use of vocabulary; no concept of variety

*Punctuation, Spelling, & Mechanics*

*4. Exceeds Expectations:* Accurate use of English writing conventions; all needed capitals, paragraphs indented, almost no punctuation and spelling errors; very neat

*3. Meets Expectations:* Some problems with writing conventions or punctuation; occasional spelling errors; paper is neat

*2. Approaches Expectations:* Uses general writing conventions but has errors; spelling problems distract reader; punctuation errors interfere with ideas

*1. Needs Development:* Serious problems with format of paper; errors in sentence-final punctuation; unacceptable to educated readers

*0. Off Topic/Did Not Try:* Complete disregard of English writing conventions; obvious capitals missing, no margins, severe spelling problems

4. The 20-point is distributed equally for weighing: Content: 4, Organization: 4, Vocabulary: 4, Grammar: 4, Punctuation, Spelling, and Mechanics: 4.

What do you think about this even weighting?

_____

_____

_____

_____

_____

_____

_____

_____

6. Are there any categories that you have found difficult to apply? If yes, please explain your reason(s).

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

II. *Please, answer the following questions.*

3. To what extent does the new writing rubric designed to be used in the proficiency examination facilitate fair assessment of students' written work? Please, rate it from 1 to 5 (1 being the lowest, 5 being the highest).

   2 ___ 2 ___ 3 ___ 4 ___ 5 ___

Please briefly explain your answer to item 1.

_____

_____

_____

4. How confident do you feel in applying the new writing rubric? Please, rate your confidence level from 1 to 5 (1 being the lowest, 5 being the highest).

1 ___ 2 ___ 3 ___ 4 ___ 5 ___

Please briefly explain your answer to item 2.

_____

_____

_____

_____

APPENDIX-17. Sample Data File

| Examinee | Rater | Content | Organization | Grammar | Vocabulary | Mechanics | Total |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 3 | 3 | 3 | 2 | 4 | 15 |
| 1 | 2 | 3 | 2 | 3 | 3 | 3 | 14 |
| 1 | 3 | 4 | 3 | 3 | 3 | 4 | 17 |
| 1 | 4 | 3 | 3 | 2 | 2 | 3 | 13 |
| 1 | 5 | 3 | 3 | 2 | 3 | 3 | 14 |
| 1 | 6 | 2 | 2 | 3 | 4 | 4 | 15 |
| 1 | 7 | 3 | 4 | 3 | 3 | 4 | 17 |
| 1 | 8 | 3 | 4 | 2 | 3 | 4 | 16 |
| 1 | 9 | 3 | 3 | 3 | 3 | 3 | 15 |
| 1 | 10 | 3 | 4 | 3 | 3 | 3 | 16 |
| 1 | 11 | 3 | 3 | 2 | 3 | 3 | 14 |
| 1 | 12 | 2 | 3 | 3 | 3 | 4 | 15 |
| 1 | 13 | 3 | 3 | 2 | 3 | 3 | 14 |
| 2 | 1 | 2 | 2 | 1 | 2 | 2 | 9 |
| 2 | 2 | 2 | 2 | 2 | 2 | 4 | 12 |
| 2 | 3 | 3 | 3 | 2 | 2 | 2 | 12 |
| 2 | 4 | 2 | 3 | 2 | 2 | 2 | 11 |
| 2 | 5 | 2 | 2 | 2 | 2 | 2 | 10 |
| 2 | 6 | 2 | 3 | 3 | 3 | 3 | 14 |
| 2 | 7 | 2 | 3 | 2 | 3 | 3 | 13 |
| 2 | 8 | 3 | 3 | 2 | 2 | 2 | 12 |
| 2 | 9 | 3 | 2 | 2 | 2 | 2 | 11 |
| 2 | 10 | 2 | 3 | 3 | 3 | 2 | 13 |
| 2 | 11 | 3 | 3 | 2 | 2 | 3 | 13 |
| 2 | 12 | 2 | 3 | 3 | 3 | 3 | 14 |
| 2 | 13 | 2 | 2 | 2 | 2 | 3 | 11 |

APPENDIX-18. Demographic Information on the Participants in Phases 4 and 5 of the Study

| Instructor | Gender | Native/Nonnative | Age | Education | Active Period of Teaching |
|---|---|---|---|---|---|
| Participant 1 | Male | Non-native | 20-29 | MA in ELT | 7 |
| Participant 2 | Female | Non-native | 30-39 | MA in ELT | 15 |
| Participant 3 | Female | Non-native | 30-39 | BA in Translation and Interpretation | 8 |
| Participant 4 | Female | Non-native | 40-49 | BA in ELT | 18 |
| Participant 5 | Male | Non-native | 40-49 | MA in ELT | 20 |
| Participant 6 | Female | Non-native | 40-49 | MA in ELT | 18 |
| Participant 7 | Female | Non-native | 30-39 | MA in ELT | 10 |
| Participant 8 | Female | Non-native | 30-39 | BA in ELT | 12 |
| Participant 9 | Male | Non-native | 40-49 | BA in ELT | 18 |
| Participant10 | Female | Non-native | 30-39 | MA in ELT | 17 |
| Participant11 | Female | Native | 20-29 | MA in Linguistics | 5 |
| Participant12 | Female | Non-native | 40-49 | MA in ELT | 17 |
| Participant13 | Female | Non-native | 30-39 | BA in ELT | 12 |

APPENDIX-19. Assessor Guide for the New Rubric

1. Categories

Five categories in the new rubric are:

- content,
- organization,
- grammar,
- vocabulary, and
- punctuation, spelling, and mechanics.

2. Wording of descriptors:

For the ease of use, a detailed description of the concepts represented by the descriptors in the new rubric at the *Meets Expectations* mastery level is presented below. The other three levels of competence are to be thought of as varying degrees, with the primary distinguishing factor being the degree to which the writer's intended meaning is successfully delivered to the reader or is diminished or completely lost by insufficient mastery of the criteria for meeting expectations.

Content

Aspects that are considered under *content*:

- topicality,
- development of ideas,
- counterargument and refutation, and
- thoughtfulness/effort.

| |
|---|
| *Content* |
| *4. Meets Expectations*: Essay addresses the assigned topic; ideas are concrete and thoroughly developed with supporting evidence provided; counterargument with a sound refutation; no extraneous material; essay reflects carefully planned thought |
| *3. Approaches Expectations:* Essay addresses the assigned topic but misses some points; ideas could be more fully developed as some evidence may be lacking; counterargument with an acceptable refutation; some extraneous material is present |
| *2. Needs Development:* Development of ideas not complete or essay is somewhat off the topic; the main ideas may not be fully supported by the evidence given; counterargument with weak or no refutation |

| |
|---|
| *1. Inadequate:* Ideas incomplete due to lack of supporting data; essay does not reflect careful thinking or was hurriedly written; inadequate effort in area of content |

Descriptor and Criteria for Content

*4. Meets Expectations*

*Essay addresses the assigned topic:*

- Is there understanding of the topic?
- Is information clearly related to the topic?

*Ideas are concrete and thoroughly developed:*

- Are facts or other relevant information used?
- Is the topic expanded enough to convey a sense of completeness?
- Are several main points discussed?
- Is there sufficient detail?

*Counterargument with sound refutation:*

- Is an opposing point of view included in a separate paragraph?
- Is this opposing point of view disproved logically?

*No extraneous material:*

- Is the information discussed without unnecessary repetition?

Organization:

Aspects that are considered under *organization*:

- title,
- paragraph structure,
- thesis statement,
- transitions,
- support for arguments, and
- conclusion.

| |
|---|
| *Organization*<br>*4. Meets Expectations*: An informative title that covers the topic; effective introductory paragraph, topic is stated in a clear thesis statement, leads to body with three full paragraphs; appropriate transitional expressions used; conclusion logical & complete<br>*3. Approaches Expectations:* Adequate title, introduction & conclusion; body of essay is acceptable with three paragraphs that may not be fully developed; sequence is logical but few transitional expressions may be absent, misused, or overused<br>*2. Needs Development:* Mediocre or scant introduction, body, or conclusion; paragraphs aren't divided exactly right; problems with the order of ideas in body; problems of organization interfere<br>*1. Inadequate:* Shaky or minimally recognizable introduction; organization can barely be seen; severe problems of ordering ideas |

Descriptor and Criteria for Organization

*4. Meets Expectations*

*An informative title that covers the topic:*

- Is there a beginning, a middle, and an end to the paper starting from the title until the concluding paragraph?

*Effective introductory paragraph:*

- Is there an introduction with appropriate background information?
- Does the introduction clearly lead to the thesis statement?

*Topic is stated in a clear thesis statement, leads to body:*

- Is there a clearly stated controlling idea or central focus that leads to the reader into the topic?

*Body with three full paragraphs including the paragraph with counterargument and refutation:*

- Does each paragraph have a topic sentence that supports, limits, and direct the thesis?
- Is the overall relationship of ideas within and between paragraphs clearly indicated?

*Appropriate transitional expressions used:*

- Do the ideas flow cohesively due to the effective use of linking words and referencing?

*Conclusion logical and complete*:

- Does the conclusion have a strong summary and/or a final comment?

Grammar

Aspects that are considered under *grammar*:

- accurate and appropriate use of sentence structure (simple, compound, and complex),
- accurate and appropriate use of prepositions, modals, articles, verb forms, and tense sequences, and
- absence/presence of fragments or run-on sentences.

| *Grammar* |
| --- |
| *4. Meets Expectations*: Proficient in English grammar with almost no grammar errors; all/majority of simple, compound, and complex sentences are correct and appropriate; mostly accurate and appropriate use of prepositions, modals, articles, verb forms, and tense sequencing; no fragments or run-on sentences |
| *3. Approaches Expectations:* Rather proficient in English grammar; few grammar problems which don't distort meaning; almost no fragments or run-on sentences |
| *2. Needs Development:* Somewhat proficient in English grammar; grammar problems distort meaning; fragments or run-on sentences present |
| *1. Inadequate:* Limited proficiency in English grammar; grammar problems hinder meaning; difficult to understand sentences |

Descriptor and Criteria for Grammar

*4. Meets Expectations*

*Proficient in English grammar with almost no grammar errors:*

- Is there basic agreement between sentence elements? (auxiliary-verb; subject-verb; pronoun-antecedent; adjective-noun; noun-quantifier)
- Are sentences well-formed and complete, with appropriate complements?
- Are coordinate and subordinate elements linked to other elements with appropriate conjunctions, adverbials, and relative pronouns?
- Are sentence types and length varied?
- Are verb tenses correct and properly sequenced?
- Do modals convey intended meaning and time?
- Are articles (*a, an,* and *the*) used correctly?
- Do pronouns reflect appropriate person, gender, and number?
- Are prepositions chosen carefully to introduce modifying elements?

Vocabulary

Aspects that are considered under *vocabulary*:

- appropriate,
- accurate, and
- varied use of words (form and meaning) and collocations.

| *Vocabulary* |
| --- |
| *4. Meets Expectations::* All/majority of vocabulary use is appropriate, accurate, and varied |
| *3. Approaches Expectations:* Appropriate and accurate use of vocabulary with few errors; attempts variety |
| *2. Needs Development:* Some vocabulary misused; repetitive use of vocabulary with little variety |
| *1. Inadequate:* Serious problems of vocabulary; lacks variety |

Descriptor and Criteria for Vocabulary

*4. Meets Expectations*

*Appropriacy:*

- Is the vocabulary proper for the topic, audience, and tone of the paper?
- Does the vocabulary make the intended impression?

*Accuracy:*

- In the context which it is used, is the choice of vocabulary correct, effective, and concise?
- Are words correctly distinguished as to their function (adjective, adverb, noun, and verb?
- Are phrasal and prepositional verbs correct? Do they convey the intended meaning?
- Are prefixes, suffixes, roots, and compounds used correctly and effectively?

*Variety:*

- Is there facility with words to convey differences of meaning, intended information, attitudes, and feelings?
- Is the arrangement and interrelationship of words sufficiently varied?

Mechanics

Aspects that are considered under *mechanics:* general English writing conventions including

- punctuation

- capitalization,
- indentation,
- spelling, and
- neatness.

| |
|---|
| *Mechanics*<br>*4. Meets Expectations::* Correct use of English writing conventions; all needed capitals, paragraphs indented, almost no punctuation and spelling errors; very neat<br>*3. Approaches Expectations:* Some problems with writing conventions or punctuation; occasional spelling errors; paper is neat<br>*2. Needs Development:* Uses general writing conventions but has errors; spelling problems distract reader; punctuation errors interfere with ideas<br>*1. Inadequate:* Complete disregard of English writing conventions ; severe spelling problems; errors in sentence-final punctuation; obvious capitals missing |

Descriptor and Criteria for Mechanics

*4. Meets Expectations*

- Are words spelled correctly?
- Are periods, commas, semicolons, dashes, question marks used correctly?
- Are capital letters used where necessary and appropriate?
- Are paragraphs indented to indicate when one sequence of thought ends and another begins?
- Is handwriting easy to read, without impeding communication?

APPENDIX-20. The Final Draft of the New Rubric

| |
|---|
| *Content*<br>*4. Meets Expectations*: Essay addresses the assigned topic; ideas are concrete and thoroughly developed with supporting evidence provided; counterargument with a sound refutation; no extraneous material; essay reflects carefully planned thought<br>*3. Approaches Expectations:* Essay addresses the assigned topic but misses some points; ideas could be more fully developed as some evidence may be lacking; counterargument with an acceptable refutation; some extraneous material is present<br>*2. Needs Development:* Development of ideas not complete or essay is somewhat off the topic; the main ideas may not be fully supported by the evidence given; counterargument with weak or no refutation<br>*1. Inadequate:* Ideas incomplete due to lack of supporting data; essay does not reflect careful thinking or was hurriedly written; inadequate effort in area of content |
| *Organization*<br>*4. Meets Expectations*: An informative title that covers the topic; effective introductory paragraph, topic is stated in a clear thesis statement, leads to body with three full paragraphs; appropriate transitional expressions used; conclusion logical & complete<br>*3. Approaches Expectations:* Adequate title, introduction & conclusion; body of essay is acceptable with three paragraphs that may not be fully developed; sequence is logical but few transitional expressions may be absent, misused, or overused<br>*2. Needs Development:* Mediocre or scant introduction, body, or conclusion; paragraphs aren't divided exactly right; problems with the order of ideas in body; problems of organization interfere<br>*1. Inadequate:* Shaky or minimally recognizable introduction; organization can barely be seen; severe problems of ordering ideas |
| *Grammar*<br>*4. Meets Expectations*: Proficient in English grammar with almost no grammar errors; all/majority of simple, compound, and complex sentences are correct and appropriate; mostly accurate and appropriate use of prepositions, modals, articles, verb forms, and tense sequencing; no fragments or run-on sentences<br>*3. Approaches Expectations:* Rather proficient in English grammar; few grammar problems which don't distort meaning; almost no fragments or run-on sentences<br>*2. Needs Development:* Somewhat proficient in English grammar; grammar problems distort meaning; fragments or run-on sentences present<br>*1. Inadequate:* Limited proficiency in English grammar; grammar problems hinder meaning; difficult to understand sentences |
| *Vocabulary*<br>*4. Meets Expectations:*: All/majority of vocabulary use is appropriate, accurate, and varied<br>*3. Approaches Expectations:* Appropriate and accurate use of vocabulary with few errors; attempts variety |

| |
|---|
| *2. Needs Development:* Some vocabulary misused; repetitive use of vocabulary with little variety<br>*1. Inadequate:* Serious problems of vocabulary; lacks variety |
| *Mechanics*<br>*4. Meets Expectations*:: Correct use of English writing conventions; all needed capitals, paragraphs indented, almost no punctuation and spelling errors; very neat<br>*3. Approaches Expectations:* Some problems with writing conventions or punctuation; occasional spelling errors; paper is neat<br>*2. Needs Development:* Uses general writing conventions but has errors; spelling problems distract reader; punctuation errors interfere with ideas<br>*1. Inadequate:* Complete disregard of English writing conventions ; severe spelling problems; errors in sentence-final punctuation; obvious capitals missing |

APPENDIX-21. Final Draft of The Open-Ended Questionnaire

RATERS' PERSPECTIVES ON THE NEW WRITING RUBRIC DESIGNED TO BE USED IN THE PROFICIENCY EXAMINATION

PART A

*Please, put a tick (√) inside the suitable box.*

1. Male □ Female □

2. Native Speaker □ Non-native Speaker □

3. Age: □ 20-29 □ 30-39 □40-49 □ 50+

4. Educational Background:

| Degree | Name of the University | Department |
|---|---|---|
| Bachelor (Lisans) □ | | |
| Master (Yuksek Lisans) □ | | |
| Doctor (Doktora) □ | | |

5. Active teaching period: _____ years

6. Active teaching period at Bursa Uludağ University, Preparatory School: _____ years

PART B

I. *Please, answer the following questions.*

3. What are the strengths and weaknesses of the new writing rubric designed to be used in the proficiency examination?

*Strengths:*

_____

_____

_____

_____

*Weaknesses:*

2. There are five levels in each category of the writing rubric used currently (Exceeds Expectations, Meets Expectations, Approaches Expectations, and Off Topic/Did Not Try). *These five levels in each category help me assess the students' writings effectively.*

To what extent do you agree with this statement? Dou you think these five levels are adequate? Please explain your reason(s).

4. Are there any categories in which the wording of the descriptors needs to be changed? If yes, please explain your reason(s).

| |
|---|
| *Content*<br>*4. Meets Expectations*: Essay addresses the assigned topic; ideas are concrete and thoroughly developed with supporting evidence provided; counterargument with a sound refutation; no extraneous material; essay reflects carefully planned thought<br>*3. Approaches Expectations:* Essay addresses the assigned topic but misses some points; ideas could be more fully developed as some evidence may be lacking; counterargument with an acceptable refutation; some extraneous material is present<br>*2. Needs Development:* Development of ideas not complete or essay is somewhat off the topic; the main ideas may not be fully supported by the evidence given; counterargument with weak or no refutation<br>*1. Inadequate:* Ideas incomplete due to lack of supporting data; essay does not reflect careful thinking or was hurriedly written; inadequate effort in area of content |

*Organization*

*4. Meets Expectations*: An informative title that covers the topic; effective introductory paragraph, topic is stated in a clear thesis statement, leads to body with three full paragraphs; appropriate transitional expressions used; conclusion logical & complete

*3. Approaches Expectations:* Adequate title, introduction & conclusion; body of essay is acceptable with three paragraphs that may not be fully developed; sequence is logical but few transitional expressions may be absent, misused, or overused

*2. Needs Development:* Mediocre or scant introduction, body, or conclusion; paragraphs aren't divided exactly right; problems with the order of ideas in body; problems of organization interfere

*1. Inadequate:* Shaky or minimally recognizable introduction; organization can barely be seen; severe problems of ordering ideas

*Grammar*

*4. Meets Expectations*: Proficient in English grammar with almost no grammar errors; all/majority of simple, compound, and complex sentences are correct and appropriate; mostly accurate and appropriate use of prepositions, modals, articles, verb forms, and tense sequencing; no fragments or run-on sentences

*3. Approaches Expectations:* Rather proficient in English grammar; few grammar problems which don't distort meaning; almost no fragments or run-on sentences

*2. Needs Development:* Somewhat proficient in English grammar; grammar problems distort meaning; fragments or run-on sentences present

*1. Inadequate:* Limited proficiency in English grammar; grammar problems hinder meaning; difficult to understand sentences

*Vocabulary*

*4. Meets Expectations::* All/majority of vocabulary use is appropriate, accurate, and varied

*3. Approaches Expectations:* Appropriate and accurate use of vocabulary with few errors; attempts variety

*2. Needs Development:* Some vocabulary misused; repetitive use of vocabulary with little variety

*1. Inadequate:* Serious problems of vocabulary; lacks variety

*Mechanics*

*4. Meets Expectations::* Correct use of English writing conventions; all needed capitals, paragraphs indented, almost no punctuation and spelling errors; very neat

*3. Approaches Expectations:* Some problems with writing conventions or punctuation; occasional spelling errors; paper is neat

*2. Needs Development:* Uses general writing conventions but has errors; spelling problems distract reader; punctuation errors interfere with ideas

*1. Inadequate:* Complete disregard of English writing conventions ; severe spelling problems; errors in sentence-final punctuation; obvious capitals missing

5. The 20-point is distributed equally for weighing: Content: 4, Organization: 4, Vocabulary: 4, Grammar: 4, and Mechanics: 4.

What do you think about this even weighting?

6. Are there any categories that you have found difficult to apply? If yes, please explain your reason(s).

_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____

II. *Please, answer the following questions.*

5. To what extent does the new writing rubric designed to be used in the proficiency examination facilitate fair assessment of students' written work? Please, rate it from 1 to 5 (1 being the lowest, 5 being the highest).

   3 ___ 2 ___ 3 ___ 4 ___ 5 ___

Please briefly explain your answer to item 1.

_____
_____
_____

6. How confident do you feel in applying the new writing rubric? Please, rate your confidence level from 1 to 5 (1 being the lowest, 5 being the highest).

1 ___ 2 ___ 3 ___ 4 ___ 5 ___

Please briefly explain your answer to item 2.

_____

_____

_____

_____

_____

APPENDIX-22. Unexpected Responses for the Draft Rubric in the MFRM Analysis (Phase 3)

| Cat | Score | Exp. | Resd | StRes | Nu | ex | N | Re | N | Criteria |
|---|---|---|---|---|---|---|---|---|---|---|
| 4 | 4 | 2.5 | 1.5 | 2.2 | 9 | 9 | 1 | R1 | 1 | content |
| 1 | 1 | 2.4 | -1.4 | -2.1 | 9 | 9 | 1 | R1 | 3 | grammar |
| 4 | 4 | 2.5 | 1.5 | 2.1 | 5 | 5 | 2 | R2 | 5 | punctuation |
| 3 | 3 | 1.5 | 1.5 | 2.2 | 9 | 9 | 3 | R3 | 1 | content |
| 1 | 1 | 2.4 | -1.4 | -2.0 | 5 | 5 | 3 | R3 | 2 | organization |
| 3 | 3 | 1.6 | 1.4 | 2.0 | 9 | 9 | 3 | R3 | 2 | organization |
| 3 | 3 | 1.4 | 1.6 | 2.4 | 1 | 1 | 3 | R3 | 5 | punctuation |
| 2 | 2 | 3.4 | -1.4 | -2.3 | 4 | 4 | 4 | R4 | 5 | punctuation |
| 0 | 0 | 2.1 | -2.1 | -3.1 | 1 | 1 | 5 | R5 | 1 | content |
| 2 | 2 | 3.4 | -1.4 | -2.2 | 5 | 5 | 5 | R5 | 2 | organization |
| 1 | 1 | 2.6 | -1.6 | -2.4 | 6 | 6 | 5 | R5 | 3 | grammar |
| Cat | Score | Exp. | Resd | StRes | Nu | ex | N | Re | N | Criteria |

APPENDIX-23. Unexpected Responses for the New Rubric in the MFRM Analysis (Phase 4)

| Cat | Score | Exp. | Resd | StRes | Nu | ex Nu | Rat | N | Criteria |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 2.9 | -1.9 | -3.0 | 4 | 4 | 1 | R1 | 1 content |
| 1 | 1 | 3.1 | -2.1 | -3.4 | 6 | 6 | 1 | R1 | 1 content |
| 2 | 2 | 3.6 | -1.6 | -3.0 | 42 | 42 | 1 | R1 | 1 content |
| 4 | 4 | 2.6 | 1.4 | 2.2 | 7 | 7 | 1 | R1 | 2 organization |
| 2 | 2 | 3.6 | -1.6 | -3.1 | 42 | 42 | 1 | R1 | 2 organization |
| 1 | 1 | 2.7 | -1.7 | -2.7 | 6 | 6 | 1 | R1 | 3 grammar |
| 4 | 4 | 2.5 | 1.5 | 2.4 | 24 | 24 | 1 | R1 | 3 grammar |
| 4 | 4 | 2.4 | 1.6 | 2.5 | 37 | 37 | 1 | R1 | 3 grammar |
| 4 | 4 | 2.3 | 1.7 | 2.8 | 11 | 11 | 1 | R1 | 4 vocabulary |
| 2 | 2 | 3.4 | -1.4 | -2.5 | 25 | 25 | 1 | R1 | 5 mechanics |
| 4 | 4 | 2.4 | 1.6 | 2.5 | 34 | 34 | 1 | R1 | 5 mechanics |
| 1 | 1 | 2.7 | -1.7 | -2.7 | 11 | 11 | 2 | R2 | 1 content |
| 2 | 2 | 3.5 | -1.5 | -2.8 | 43 | 43 | 2 | R2 | 1 content |
| 1 | 1 | 2.3 | -1.3 | -2.1 | 13 | 13 | 2 | R2 | 2 organization |
| 1 | 1 | 3.3 | -2.3 | -3.8 | 17 | 17 | 2 | R2 | 2 organization |
| 2 | 2 | 3.3 | -1.3 | -2.1 | 49 | 49 | 2 | R2 | 2 organization |
| 1 | 1 | 2.6 | -1.6 | -2.6 | 21 | 21 | 2 | R2 | 3 grammar |
| 4 | 4 | 2.5 | 1.5 | 2.3 | 46 | 46 | 2 | R2 | 3 grammar |
| 4 | 4 | 2.6 | 1.4 | 2.2 | 32 | 32 | 2 | R2 | 4 vocabulary |
| 4 | 4 | 2.6 | 1.4 | 2.3 | 35 | 35 | 2 | R2 | 4 vocabulary |
| 4 | 4 | 2.7 | 1.3 | 2.1 | 46 | 46 | 2 | R2 | 4 vocabulary |
| 4 | 4 | 2.6 | 1.4 | 2.2 | 2 | 2 | 2 | R2 | 5 mechanics |
| 4 | 4 | 2.7 | 1.3 | 2.1 | 7 | 7 | 3 | R3 | 2 organization |
| 4 | 4 | 2.4 | 1.6 | 2.7 | 23 | 23 | 3 | R3 | 2 organization |
| 1 | 1 | 2.5 | -1.5 | -2.5 | 20 | 20 | 3 | R3 | 4 vocabulary |
| 4 | 4 | 2.6 | 1.4 | 2.2 | 24 | 24 | 3 | R3 | 4 vocabulary |

| 1 1 2.5 -1.5 -2.3 | 23 23 3 R3 5 mechanics |
| --- | --- |
| 2 2 3.5 -1.5 -2.6 | 25 25 3 R3 5 mechanics |
| 4 4 2.5 1.5 2.3 | 33 33 3 R3 5 mechanics |
| 2 2 3.3 -1.3 -2.1 | 45 45 3 R3 5 mechanics |
| 1 1 2.5 -1.5 -2.5 | 18 18 4 R4 1 content |
| 1 1 2.3 -1.3 -2.2 | 37 37 4 R4 3 grammar |
| 2 2 3.3 -1.3 -2.1 | 36 36 4 R4 4 vocabulary |
| 1 1 2.6 -1.6 -2.6 | 8 8 5 R5 1 content |
| 2 2 3.3 -1.3 -2.2 | 36 36 5 R5 2 organization |
| 1 1 2.3 -1.3 -2.1 | 8 8 5 R5 3 grammar |
| 1 1 2.7 -1.7 -2.7 | 11 11 6 R6 1 content |
| 1 1 2.3 -1.3 -2.1 | 13 13 6 R6 1 content |
| 1 1 2.5 -1.5 -2.4 | 34 34 6 R6 1 content |
| 1 1 2.3 -1.3 -2.1 | 13 13 6 R6 2 organization |
| 1 1 2.6 -1.6 -2.5 | 33 33 6 R6 2 organization |
| 2 2 3.5 -1.5 -2.8 | 40 40 6 R6 2 organization |
| 2 2 3.5 -1.5 -2.6 | 41 41 6 R6 2 organization |
| 4 4 2.6 1.4 2.3 | 35 35 6 R6 4 vocabulary |
| 2 2 3.4 -1.4 -2.3 | 38 38 7 R7 2 organization |
| 2 2 3.7 -1.7 -3.6 | 40 40 7 R7 2 organization |
| 2 2 3.7 -1.7 -3.3 | 41 41 7 R7 2 organization |
| 2 2 3.5 -1.5 -2.7 | 45 45 7 R7 2 organization |
| 2 2 3.5 -1.5 -2.6 | 49 49 7 R7 2 organization |
| 1 1 2.4 -1.4 -2.3 | 33 33 7 R7 3 grammar |
| 2 2 3.4 -1.4 -2.5 | 40 40 7 R7 3 grammar |
| 2 2 3.4 -1.4 -2.5 | 43 43 7 R7 3 grammar |
| 1 1 3.2 -2.2 -3.5 | 45 45 7 R7 3 grammar |
| 1 1 3.0 -2.0 -3.2 | 47 47 7 R7 3 grammar |
| 4 4 2.6 1.4 2.2 | 30 30 7 R7 4 vocabulary |
| 4 4 2.5 1.5 2.5 | 34 34 7 R7 4 vocabulary |
| 2 2 3.5 -1.5 -2.6 | 41 41 7 R7 4 vocabulary |
| 2 2 3.3 -1.3 -2.1 | 45 45 7 R7 4 vocabulary |

| 2 2 3.3 -1.3 -2.1 | 37 37 7 R7 5 mechanics |
| --- | --- |
| 3 3 3.8 -.8 -2.3 | 42 42 7 R7 5 mechanics |
| 2 2 3.6 -1.6 -2.9 | 45 45 7 R7 5 mechanics |
| 4 4 2.6 1.4 2.2 | 11 11 8 R8 2 organization |
| 3 3 3.9 -.9 -2.7 | 9 9 9 R9 2 organization |
| 1 1 3.5 -2.5 -4.3 | 24 24 9 R9 2 organization |
| 2 2 3.3 -1.3 -2.2 | 32 32 9 R9 2 organization |
| 3 3 3.9 -.9 -3.0 | 9 9 9 R9 5 mechanics |
| 2 2 3.6 -1.6 -3.0 | 10 10 9 R9 5 mechanics |
| 3 3 3.8 -.8 -2.1 | 25 25 9 R9 5 mechanics |
| 3 3 3.9 -.9 -3.2 | 42 42 9 R9 5 mechanics |
| 3 3 3.9 -.9 -2.4 | 43 43 9 R9 5 mechanics |
| 1 1 2.5 -1.5 -2.5 | 5 5 10 R10 1 content |
| 1 1 2.3 -1.3 -2.1 | 7 7 10 R10 1 content |
| 1 1 2.4 -1.4 -2.3 | 39 39 10 R10 1 content |
| 4 4 2.6 1.4 2.3 | 27 27 10 R10 2 organization |
| 1 1 2.4 -1.4 -2.2 | 35 35 10 R10 2 organization |
| 4 4 2.7 1.3 2.1 | 47 47 10 R10 2 organization |
| 3 3 1.8 1.2 2.1 | 2 2 10 R10 3 grammar |
| 4 4 2.7 1.3 2.1 | 41 41 10 R10 3 grammar |
| 2 2 3.4 -1.4 -2.4 | 25 25 11 R11 5 mechanics |
| 2 2 3.3 -1.3 -2.1 | 1 1 12 R12 1 content |
| 2 2 3.3 -1.3 -2.2 | 6 6 12 R12 1 content |
| 1 1 2.8 -1.8 -2.9 | 11 11 12 R12 1 content |
| 1 1 2.8 -1.8 -2.9 | 11 11 12 R12 2 organization |
| 1 1 2.6 -1.6 -2.5 | 34 34 12 R12 2 organization |
| 4 4 2.5 1.5 2.4 | 13 13 12 R12 5 mechanics |
| 1 1 2.9 -1.9 -3.1 | 5 5 13 R13 1 content |
| 1 1 2.6 -1.6 -2.6 | 11 11 13 R13 1 content |
| 1 1 2.8 -1.8 -2.9 | 20 20 13 R13 1 content |
| 1 1 2.5 -1.5 -2.4 | 33 33 13 R13 1 content |
| 1 1 2.4 -1.4 -2.3 | 34 34 13 R13 1 content |

| 1 1 2.9 -1.9 -3.1 | 5 5 13 R13 2 organization |
| 1 1 2.6 -1.6 -2.6 | 11 11 13 R13 2 organization |
| 1 1 2.9 -1.9 -3.1 | 24 24 13 R13 2 organization |
| 1 1 2.8 -1.8 -2.8 | 32 32 13 R13 2 organization |
| 1 1 2.9 -1.9 -2.9 | 46 46 13 R13 2 organization |
| 1 1 3.1 -2.1 -3.3 | 47 47 13 R13 2 organization |
| 1 1 2.9 -1.9 -3.0 | 48 48 13 R13 2 organization |
| 2 2 3.3 -1.3 -2.2 | 17 17 13 R13 5 mechanics |
| 1 1 3.2 -2.2 -3.6 | 26 26 13 R13 5 mechanics |
| 4 4 2.7 1.3 2.1 | 30 30 13 R13 5 mechanics |
|---|---|
| Cat Score Exp. Resd StRes| Nu ex Nu Rat N Criteria |
+---+

# ETİK KURUL İZNİ

Evrak Kayıt Tarihi: 30.01.2018 | Protokol No: 12643

Tarih: 26.02.2018

ANADOLU ÜNİVERSİTESİ
SOSYAL VE BEŞERÎ BİLİMLER BİLİMSEL ARAŞTIRMA VE YAYIN ETİĞİ KURULU
KARAR BELGESİ

| | |
|---|---|
| **ÇALIŞMANIN TÜRÜ:** | Doktora Tez Çalışması |
| **KONU:** | Eğitim Bilimleri |
| **BAŞLIK:** | Uludağ Üniversitesi Yabancı Diller Yüksekokulu İngilizce Hazırlık Okulunda İngilizce Yazılı Anlatım Becerisini Ölçmek Üzere Geliştirilen Çok Boyutlu Notlandırma Ölçeğinin Güvenirlik ve Geçerliliği |
| **~~PROJE~~/TEZ YÜRÜTÜCÜSÜ:** | Doç. Dr. Fatma Hülya ÖZCAN |
| **TEZ YAZARI:** | Aliye Evin YÖRÜDÜ |
| **ALT KOMİSYON GÖRÜŞÜ:** | - |
| **KARAR:** | Olumlu |

| | |
|---|---|
| **Prof.Dr. Coşkun BAYRAK**<br>*(Başkan-Eğitim Fak.)* | |
| **Prof.Dr. T. Volkan YÜZER**<br>*(Başkan Yardımcısı-Açıköğretim Fak.)* | **Prof.Dr. Esra CEYHAN**<br>*(Eğitim Fak.)* |
| **Prof.Dr. Münevver ÇAKI**<br>*(Güzel Sanatlar Fak.)* | **Prof.Dr. M. Erkan ÜYÜMEZ**<br>*(İkt. ve İdari Bil. Fak.)* |
| **Prof.Dr. Handan DEVECİ**<br>*(Eğitim Fak.)* | **Prof.Dr. Emel ŞIKLAR**<br>*(İkt. ve İdari Bil. Fak.)* |