

A Qualitative Analysis of Descriptors for the Organization Rating Scales in EFL Essay Writing

Kana Matsumura^{*1} and Akiko Takagi^{*2}

^{*1} College of Education, Psychology, and Human Studies, Aoyama Gakuin University

^{*2} College of Education, Psychology, and Human Studies, Aoyama Gakuin University

EFL エッセイライティングの評価尺度における構成に関する 記述子の質的分析

松村 香奈^{*1} 高木 亜希子^{*2}

^{*1} 青山学院大学教育人間科学部 ^{*2} 青山学院大学教育人間科学部

Abstract: Among the categories on the ESL/EFL writing scale, *organization* specifically has often been criticized for being vague, as its descriptors are represented frequently by abstract adjectives or adverbs, namely, indicators of the degree of quality. This inevitably leads to a heavy dependency on the judgment of the raters on scores. Moreover, the organizational skill in L2 writing covers multiple aspects, such as logical sequencing, topic development, and connective devices, where a different discourse pattern than in L1 is sometimes required. In addition, descriptors are represented in different wordings depending on the rating scale. These issues potentially make writing assessment in terms of *organization* more complex and difficult. This study qualitatively analyzes descriptors of rhetorical and ideational aspects in seven representative writing rating scales, both holistic and of multiple traits, which are considered suitable for argumentative writing by Barkaoui (2007). The descriptors in the rating scales were investigated in terms of the rater strategies for judgment making based on the categorization of Cumming, Kantor, and Powers (2001, 2002). The study suggests where the ambiguity in the *organization* writing subcategory lies and helps reconceptualize the components of EFL/ESL rhetorical and ideational writing skills.

Keywords: academic writing, organization, descriptors in rating scales

1. Introduction

Writing tests are usually evaluated directly by rater(s) on the performance of the examinees' products. Even when using the same rating scale, discrepancies can occur within and between raters. If avoiding such discrepancies is a top priority, the indirect test¹ could be a solution, where the questions that can be judged as having a correct or incorrect answer are predetermined in relation to the skill being measured. However, this type of writing test is much less common in the era of performance testing today from the perspective of face validity (Knoch, 2007). After all, it is concluded that "the best way to test people's writing ability is to get them to write," as Hughes claims (1989, p. 75). The direct type of writing assessment does not follow a simple correct or incorrect decision-making process, and the evaluation is left to the judgment of the rater based on the rater's decision-making behaviors. Previous research has often focused on investigating the perceptions of raters at the time of evaluation, but there has been little research on the descriptors of the rating scales themselves. It is important as well to explore the descriptors which raters rely on when evaluating. Descriptors in rubrics are a list of descriptions defining in advance the ability being measured (Bachman & Palmer, 1996; Weigle, 2011). Level descriptors in the scale are achievement level definitions linked to content standards of band levels. EFL writing descriptors can be categorized into two main assessment perspectives: rhetorical and ideational focus and language use focus (e.g., Barkaoui, 2007; Cumming et al., 2001, 2002). Rhetorical and ideational focus refers to content fulfillment and organizational effectiveness in the analytic evaluation category, while language focus refers to aspects of grammar, vocabulary use, and mechanics. While the rating descriptors of language use mainly require classification of errors, those of organizational aspects cover multiple subcomponents which often overlap and are not always clearly identified by raters. Considering the issues mentioned above, descriptors of rhetorical and ideational elements in the existing rating scales should be explored with qualitative text analysis and descriptive statistics to identify the possible causes of ambiguity in organization rating scales. This study was conducted with the following two research questions in an attempt to contribute to the intra- and inter-rater reliability of L2 writing assessment as well as to the teaching of the organizational aspects of academic writing classes by reconceptualizing the subcomponents of writing organization.

1. How can Rhetorical and Ideational strategies be defined in the level descriptors in existing rating scales?
2. To what extent and how do the level descriptors correspond to rhetorical and ideational strategies?

2. Literature review

Considering writing performance is scored according to the judgment standard of the raters, identifying and analyzing the decision-making process or the behavior of the raters is important from a reliability perspective. From this standpoint, previous research has been conducted to survey what raters are doing while evaluating essays using certain rating scales. Milanovic, Saville, and Shuhong (1996) investigated the decision-making behaviors and the essay aspects to which raters pay attention when using holistic rating scales, whereas Cumming (1990) and Lumley (2002) did so when using multiple-trait rating scales. Lumley (2002) presents findings that have some important implications in terms of the relation between raters' evaluations and descriptors in rating scales. His main findings can be summarized as a struggle in how raters reconcile the gap between their own intuitional perceptions, judgmental norms, or actual text samples and the wordings of the rating scale. He describes the situation, saying that raters "somehow managed to refer to the scale content" to make their decisions when scoring in a quest for justification (p. 263).

Unlike the above studies that used existing rating scales, Cumming, Kantor, and Powers (2001, 2002) investigated raters' behaviors and essay aspects with no specific rating guidelines. A series of studies was conducted to develop and verify a descriptive framework concerning the decision-making processes of raters as part of the development process of TOEFL 2000². The research consists of three studies. In the first study, a preliminary descriptive framework was developed based on the think-aloud protocols produced by 10 experienced ESL/EFL instructors/assessors of differing backgrounds. The verbal protocol was collected while they were rating 60 TOEFL essays. In the second study, the framework was refined by gathering additional think-aloud data from another seven experienced English-mother-tongue (EMT) composition raters with extensive experience assessing native-English compositions for various tests administered by English Testing Service (ETS®). The third study was conducted with the aim to refine the framework by analyzing think-aloud protocols from seven of the same ESL/EFL raters who rated different integrated types of writing tasks. In the descriptive framework, developed and refined through multiple stages of experimentation, the decision-making behaviors were categorized into three foci: Self-monitoring Focus, Rhetorical and Ideational Focus, and Language Focus. They include two strategies: Interpretation Strategies and Judgment Strategies, which were further categorized into multiple decision-making behavioral elements. The strategies were divided into 35 distinct and independent elements. Among them, nine elements of Rhetorical and Ideational Judgment Strategies are addressed in the current study. An overview of the framework can be found in Table 1 below.

Based on Cumming et al.'s (2001, 2002) framework, Barkaoui (2007) conducted further

research on how different rating scales affect rater performance when assessing L2 writing performance. One of his main findings was that raters employed a similar rating process, whether with holistic or analytic rating scales. For the purpose of his study, he consulted the other experienced raters and selected one representative rater of each type of rating scale in the earlier stage of the study. Prior to the selection, he identified four multiple-trait rating scales and four holistic scales from the literature that he believed, based on his teaching experience, were appropriate for marking EFL university students' argumentative essays.

Although these previous studies have focused on exploring the evaluator's perceptions and behaviors during evaluation, there is no specific indication of how the descriptors were interpreted and employed. Therefore, in order to shed light on the point, the authors investigated descriptors of multiple writing rating scales that Barkaoui (2007) selected. The rating scales covered in the study are shown in Tables 2 and 3.

Table 1 Descriptive framework of decision-making behaviors while rating TOEFL Writing Tasks (Adapted from Barkaoui, 2007, pp. 104–105; Cumming et al., 2001, p. 53; 2002, p. 88)

Self-monitoring Focus	Rhetorical and Ideational Focus	Language Focus
<i>Interpretation Strategies</i>		
Read or interpret prompt or task input or both	Discern rhetorical structure	Classify errors into types
Read or reread composition	Summarize ideas or propositions	Interpret or edit ambiguous or unclear phrases
Envision personal situation of the writer	Scan whole composition or observe layout	
<i>Judgment Strategies</i>		
Decide on macro strategy for reading and rating; compare with other compositions or summarize, distinguish, or tally judgments collectively	Assess reasoning, logic, or topic development	Assess quantity of total written production
Consider own personal response or biases	Assess task completion or relevance	Assess comprehensibility and fluency
Define or revise own criteria	Assess coherence and identify redundancies	Consider frequency and gravity of errors
Articulate general impression	Assess interest, originality, or creativity	Consider lexis
Articulate or revise scoring decision	Assess text organization, style, register, discourse functions, or genre	Consider syntax or morphology
	Consider use and understanding of source material	Consider spelling or punctuation
	Rate ideas or rhetoric	Rate language overall

(The items surrounded by a double line are the strategies of interest in this study.)

3. Method

In the current study, seven of the eight representative rating scales introduced by Barkaoui (2007) were adopted (shown in Table 2: four analytic / multiple-trait rating scales and Table 3: three holistic rating scales below), and the descriptors contained in the scales were qualitatively analyzed in terms of the judgment strategies of the raters' decision-making behaviors based on the framework categorization shown in the previous section (Cumming et al., 2001, 2002). Hugues' (1989/2003) holistic rating scale was excluded from the analysis because the descriptors in his holistic scale did not illustrate detailed descriptions. The analysis was conducted via the following steps: (1) Descriptors were extracted from the rating scales and segmented into sentence units. (2) In total, 204 sentence units were examined to see which of Barkaoui's nine judgment strategies the raters' decision-making behaviors fall under. (3) The item(s) that corresponded were assessed on a three-level grading of ++ (most applicable), + (somewhat applicable), and no mark (not or little applicable) depending on the degree of applicability. (4) In case of discrepancies in the authors' evaluations, final agreement was reached through discussion. (5) The results were tallied and the answers to the research questions were surmised.

Table 2 Analytic/multiple-trait rating scales for analysis

Author(s), Scale	Target (sub)categories
Brown, J. D., & Bailey, K. M. (1984). <i>Composition Grading Scale</i>	• Organization • Logical Development of Ideas: Content • Style and Quality of Expression
Hamp-Lyons, L., & Henning, G. (1991). <i>ECPS</i>	• Organization • Argumentation
Jacobs, H. L. et al. (1981). <i>ESL Composition Profile</i>	• Content • Organization
Weir, C. J. (1993). <i>TEEP Attribute Writing Scales</i>	• Relevance and Adequacy of Content • Compositional Organization • Cohesion

Table 3 Holistic rating scales for analysis

Author(s), Scale	Score Band Level
English Testing Service (2020). <i>TOEFL iBT® Writing Responses. Independent Writing Rubrics</i>	5 levels
Song, C. B., & Caruso, I. (1996). <i>Evaluation scale for the Writing Skills Assessment Test</i>	6 levels
Tyndall, B., & Kenyon, D. M. (1996). <i>EFL placement test, proposed holistic scoring guide cont.</i>	7 levels

4. Results

Prior to the analysis of the descriptors, the authors discussed and defined nine strategies based on the descriptive examples of the think-aloud protocols of Cumming et al. (2001, 2002) and other previous studies (e.g., Halliday & Hassan, 1976; Lautamatti, 1990; van Dijk, 1977). A list of definitions of the Rhetorical and Ideational judgment strategies compiled by the authors as a basis for descriptor analysis is shown in Table 4 below, which answers RQ 1.

In the next steps to answer RQ 2, 204 sentence units of the descriptors in the seven rating scales were coded by the authors individually concerning which of the judgment strategies they fall under. First, 66 descriptors were determined to be excluded from further analysis, as concerns were identified with content exclusively, not with organizational aspects. As a result, 138 descriptors in total were submitted for further analysis. Figure 1 shows the frequency of descriptors corresponding to Rhetorical and Ideational strategies. As the results suggest, the strategies are applicable to many descriptors, such as 1, 4, and 7, which require more comprehensive judgment. Presumably, strategies 1 and 7, with more double positive (++) labels specifically, are overlapping strategies located in a higher hierarchy. Lastly, the fact that the frequency of corresponding descriptors in Figure 1 ($N = 382$) is greater than the total number of descriptors ($N = 138$) suggests some correspond to more than one strategy, and this overlap is shown in Figure 2. The descriptive statistic of the number of strategy elements each descriptor contains is as follows: *Mean* = 2.8, *median* = 2.0, *mode* = 2.0, *SD* 1.4, *min.* = 1, *max.* = 7. The results suggest that each descriptor applies to multiple rhetorical strategies, and some even require reference to as many as seven of nine strategies.

Table 4 List of definitions of the strategies compiled by the authors as a basis for descriptor analysis

	Judgment Strategies	Key words and phrases for in definition
1	Assess reasoning, logic, or topic development	Elaboration on the thesis sentence or the topic; persuasiveness; fluency or stagnation in terms of time, space, and argument
2	Assess task completion	Finishing; task completion; conclusion; enough for evaluation; comprehensibility
3	Assess relevance	(Ir)relevance and off/on topic; answers to the question; relation to the topic
4	Assess coherence	Semantic property of discourses; propositional content of discourse based on the interpretation of other sentences; connection of adjacent sentences; convolution or certain focus
5	Assess interest, originality, or creativity	Uniqueness; intriguingness; empathy with readers
6	Identify redundancies	Mere repetition of the same arguments; no or little relation to the topic; not useful or unnecessary
7	Assess text organization	The whole guidance to the reader; reference to the larger parts of a piece of writing, such as a paragraph or an essay; integrity across paragraphs or throughout the essay; existence of organizational constituents in a paragraph or an essay: introduction, support(s), conclusion
8	Assess style, register, or genre	Bookish version of English; formulaic way of writing composition; appropriateness to the situation or type of writing
9	Rate ideas or rhetoric	Sophistication in rhetoric; comprehensive judgment when there is not enough material for a decision, or it is difficult to make a partial decision

5. Discussion

From the high frequency of occurrence in descriptors, it can be inferred that the existence of fundamental organizational elements and the topic development are likely to be the indispensable subcomponents in the nine organizational strategies. Furthermore, coherence and rhetorical sophistication, which can latently be included in any descriptor, have the potential to be a source of ambiguity, leading to different judgments by raters. As expected, it is now clear once again that a single descriptor contains multiple rhetorical strategies, and it goes without saying that this is the source of the ambiguity that may puzzle raters in the decision-making process.

Figure 1 Frequency of descriptors corresponding to Rhetorical and Ideational strategies

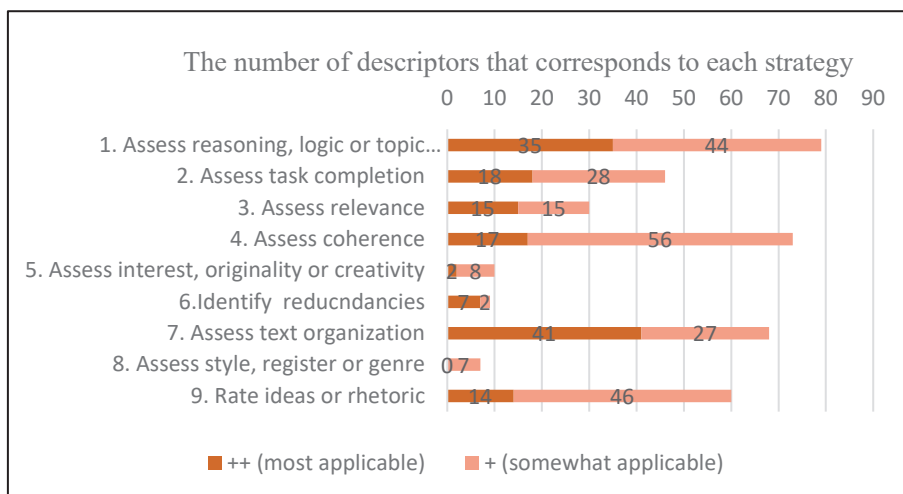
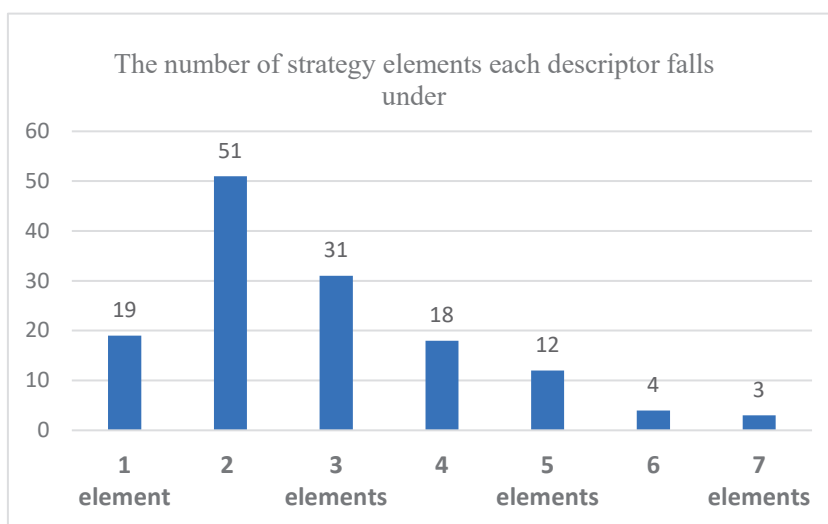


Figure 2 The number of strategy elements each descriptor contains



6. Concluding remarks

Judgments about rhetorical aspects of EFL/ESL writing using descriptors in rating scales require multiple assessment strategies. The interactive and overlapping nature among the subcomponents of the organization skill is perceived as an ambiguity in the descriptors. Hence, it is particularly important in writing tests to ensure a clear reconceptualization of the descriptors in the organization criteria among raters during rater training. Leaving the descriptors alone may cause inter-rater and intra-rater reliability issues. This is not only a matter of rater training, but also teaching practice in the classroom, where instructors should

be able clearly to articulate the component concepts of organizational skills to their students. Lastly, the limitation of this study is that the number of analysts was restricted. A more comprehensive and detailed analysis may then be needed.

Notes

- ¹ Indirect tests consist of such test items as multiple-choice questions, cloze items, paraphrasing, or sentence re-ordering. DIALANG is one example of indirect language diagnostic testing systems freely available online with five language skills or knowledge categories (Reading, Listening, Writing, Vocabulary, and Grammar, while Speaking is excluded for technical reasons) in 14 European languages. The test specification is based on the Common European Framework (CEFR; Council of Europe, 2001). <https://dialangweb.lancaster.ac.uk/>
- ² TOEFL 2000 is interrelated projects to develop a scoring scheme for a new Test of English as a Foreign Language (TOEFL®).

References

- Bachman, L. F. & Palmer, A. S. (1996). *Language testing in practice*. Oxford University Press.
- Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing*, 12, 86–107.
- Brown, J. D., & Bailey, K. M. (1984). A categorical instrument for scoring second language writing skills. *Language Learning*, 34(4), 21–42.
- Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing*, 7(1), 31–51.
- Cumming, A., Kantor, R., & Powers, D. E. (2001). *Scoring TOEFL essays and TOEFL 2000 prototype writing tasks: An investigation into raters' decision making and development of a preliminary analytic framework*. TOEFL® Monograph Series 22. Educational Testing Service.
- Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *Modern Language Journal*, 86(1), 67–96.
- Education Teal testing Service. (2020). *Scoring guides (rubrics) for TOEFL iBT® writing responses*. Independent Writing Rubrics.
- Halliday, M. A. K., & Hassan, R. (1976). *Cohesion in English*. Longman.
- Hamp-Lyons, L., & Henning, G. (1991). Communicative writing profiles: An investigation of the transferability of a multiple-trait scoring instrument across ESL writing assessment contexts. *Language Learning*, 41(3), 337–373.

- Hughes, A. (1989/2003). *Testing for language teachers*. Cambridge University Press.
- Jacobs, H. L., Zinkgraf, S. A., Wormuth, D. R., Hartfiel, V. F., & Hughey, J. B. (1981). *Testing ESL Composition: A practical approach*. Newbury House.
- Knoch, U. (2007). ‘Little coherence, considerable strain for reader’: A comparison between two rating scales for the assessment of coherence. *Assessing Writing*, 12(2), 108–128.
- Lautamatti, L. (1987). Observations on the development of the topic of simplified discourse. In U. Connor & R. B. Kaplan (Eds.), *Writing across languages: Analysis of L2 text* (pp. 87–114). Addison-Wesley.
- Lumley, T. (2002). Assessment criteria in a large-scale writing test: what do they really mean to the raters? *Language Testing*, 19(3), 246–276.
- Milanovic, M., Saville, N., & Shuhong, S. (1996). A study of the decision-making behavior of composition markers. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Colloquium, Cambridge and Arnhem* (pp. 92–114). Cambridge University Press.
- Song, C. B., & Caruso, I. (1996). Do English and ESL faculty differ in evaluating the essays of native English-speaking and ESL students? *Journal of Second Language Writing*, 5(2), 163–182.
- Tyndall, B., & Kenyon, D. M. (1996). Validation of a new holistic rating scale using Rasch multifaceted analysis. In A. Cumming & R. Berwick (Eds.), *Validation in language testing* (pp. 39–57). Multilingual Matters.
- van Dijk, T. A. (1972). *Some aspects of text grammar*. Mouton.
- Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press.
- Weir, C. J. (1993). *Understanding and developing language tests*. Prentice Hall.

(投稿受理日 2021 年 9 月 26 日)