THE JOURNAL OF ASIA TEFL
Vol. 16, No. 3, Fall 2019, 906-926
http://dx.doi.org/10.18823/asiatefl.2019.16.3.9.906

The Journal of Asia TEFL

http://journal.asiatefl.org/

e-ISSN 2466-1511

Developing and Validating Empirically-Derived Diagnostic Descriptors in ESL Academic Writing

Youn-Hee Kim
Daegu Catholic University, Korea

Despite the increasing interest in and need for developing empirical and diagnostic assessment schemes in L2 assessment and testing, very little research has been devoted to it. The literature in this area is scant, and only a handful of studies (e.g., Fulcher, 1993; Knoch, 2009; North & Schneider, 1998; Upshur & Turner, 1995) have attempted to explore such possibility. In response to this need for research, this two-phase study developed and validated empirically-derived diagnostic descriptors for use in small-scale classroom assessment in English as a second language (ESL) academic writing. In Phase 1, descriptors that are relevant to the construct of ESL academic writing were developed, and in Phase 2, the use of the descriptors was validated through the multiple analyses. The results indicated that 35 empirically-derived descriptors addressed all aspects of ESL writing skills, including content fulfillment, organizational effectiveness, grammatical knowledge, vocabulary use, and mechanics. In addition, teachers were able to use the descriptors in a consistent manner. These findings demonstrate the effectiveness of an empirical approach to assessment scheme development, and underscore the importance of its use for diagnostic purposes.

Keywords: ESL academic writing, ESL writing assessment, diagnostic assessment, empirical approach, scale development and validation

Introduction

Despite the extensive use of rating scales in second language (L2) assessment and testing, surprisingly little is known about their theoretical and empirical underpinnings. Most rating scales originate in intuitive or a priori methods of scale development, and little information is publicly known about their development procedures (Upshur & Turner, 1995).[1] A lack of empirical grounding keeps scale developers and assessors from knowing what evaluation elements should be included, resulting in low reliability and validity (Matthews, 1990). This problem becomes far more serious when a scale is used for diagnostic purposes. The identification of specific assessment elements is regarded as the most important procedure in implementing diagnostic assessment because these elements form the basis of detailed skill profiles (Jang, 2009; Lee & Sawaki, 2009a; Lee & Sawaki, 2009b; Sawaki, Kim, & Gentile, 2009). These criticisms will remain until different paradigms and approaches can be applied to the developmental process.

Acknowledging the limitations of intuition-based rating scales or a priori rating scales, Brindley (1998)

[1] An a priori method means developing rating scales based on experts' (e.g., experienced teachers, language testers, or language testing specialists in examination board) intuitive judgments concerning the development of language proficiency, a teaching syllabus, or a needs analysis (Fulcher, 2003).

suggested empirically-derived diagnosis-oriented rating scales:

> Rather than continuing to proliferate scales which use generalized and empirically unsubstantiated descriptors, therefore, it would perhaps be more profitable to draw on SLA and LT research to develop more specific empirically derived and diagnostically oriented scales [italics added] of task performance which are relevant to particular purposes of language use in particular contexts and to investigate the extent to which performance on these tasks taps common components of competence. (p. 134)

Recognizing the problems associated with intuitive or a priori methods in most rating scales, he placed particular emphasis on empirical sources, as well as the diagnostic functions that rating scales must have.

In a similar vein, Pollitt and Murray (1996) suggested diagnosis-oriented rating scales, pointing out the limited view of Alderson's (1991) trichotomous classification of rating scales (i.e., user-oriented, constructor-oriented, and assessor-oriented scales). The area of English for Specific Purposes (ESP) was no exception; Grove and Brown (2001) proposed a diagnostic assessment scheme that can help to assess medical students' oral communicative skills. Although she did not empirically develop or validate it, Luoma (2004) also put forward the idea of a diagnostic rating checklist for assessing L2 oral proficiency.

Despite the increasing interest in and need for developing empirical and diagnostic assessment schemes in L2 assessment and testing, very little research has been devoted to it. The literature in this area is scant, and only a handful of studies (e.g., Fulcher, 1993; Knoch, 2009; North & Schneider, 1998; Upshur & Turner, 1995) have attempted to explore such possibility. In response to this need for research, this study developed and validated empirically-derived diagnostic descriptors[2] for use in a small-scale classroom assessment in English as a second language (ESL) academic writing. Specifically, the following research questions were addressed:

1) What empirically-derived descriptors are relevant to the construct of ESL academic writing?
2) Are teachers able to use the empirically-derived diagnostic descriptors in an internally consistent manner?
3) To what extent do teachers agree with each other when using the empirically-derived diagnostic descriptors?

Literature Review

Empirically-based Rating Scales

Acknowledging the problems associated with intuitive or a priori methods in most scales, researchers turned their attention to empirical methods. Of particular interest were Fulcher's (1993) data-driven fluency rating scale, Upshur and Turner's (1995) empirically-derived, binary-choice, boundary-definition (EBB) rating scale, and the Common European Framework of Reference for languages (CEFR) scale (North & Schneider, 1998), and Knoch's (2009) theoretically-based and empirically-developed diagnostic rating scale. Unlike committee-based or authority-based scales, these are mostly noted for their attempt to incorporate real language performance into rating scale development.

Fulcher (1993) proposed a data-based or data-driven fluency rating scale based on observations of oral performance in order to define and measure the fluency of L2 learners. His data-based approach was built on the claim that observed learners' performance should be quantifiable, and that the development procedures of rating scales should reflect real linguistic performance. In contrast to a priori methods, this

[2] Descriptors differ from assessment criteria in that descriptors refer to more detailed and specific descriptions of the target ability to be assessed.

data-based procedure utilized a large database of speech samples, which were then used to create fluency rating descriptors. Fulcher's data-driven fluency scale demonstrates that discourse analysis can help to create detailed scale descriptors for L2 oral performance, thereby distinguishing itself from other traditional rating scales.

Upshur and Turner (1995) suggested the effectiveness of a series of empirical yes/no criteria questions in developing a scale. Known as empirically-derived, binary-choice, boundary-definition (EBB) scales, they are constructed based on performance samples in a particular task where raters were asked to make a sequence of yes/no choices about characteristics of test performance, which distinguishes boundaries between score levels (Upshur & Turner, 1995). Instead of generalizing to other contexts, EBB rating scales are usually developed within a particular context and with a particular task and learner group in mind. EBB rating scales are not constructed on theoretical models of language ability or learning, but on a set of hierarchical binary questions about small samples of the particular task being rated (Turner & Upshur, 1996; Upshur & Turner, 1995).

North's (1995, 1996) descriptor scaling method in the CEFR also demonstrates that a combination of theoretical and empirical approaches is useful in developing the framework of reference in which L2 performance levels are determined. North (1995, 1996) and North and Schneider (1998) introduced a new approach to scaling descriptors using the three steps: (a) a comprehensive pool of descriptors was created, (b) descriptors were qualitatively validated by consulting teacher workshops, and (c) descriptors were scaled using teacher assessment and the many-faceted Rasch model. They suggested that consistent scales can be constructed in a principled way using comprehensive surveys of existing scales, theoretical reviews and a priori validation of descriptors, descriptor scaling based on a measurement model, and replications of the scale (North & Schneider, 1998).

In a study that assessed L2 writing, Knoch (2009) developed a theoretically-based and empirically-developed rating scale for an L2 diagnostic writing test and evaluated its diagnostic function. In the first part of the two-phase study, Knoch examined the existing literature to identify objective discourse measures believed to best discriminate between writing samples at different proficiency levels. She then pilot-tested the measures and determined their discriminant functions based on descriptive statistics (e.g., means and standard deviations). To confirm that the measures that survived the pilot test had sufficient discriminant function, writing samples were evaluated and screened based on their descriptive statistics (i.e., histograms, box-plots, and means) and the ANOVA results. The resulting refined objective measures were finally used to construct a diagnostic L2 writing rating scale assessing accuracy, fluency, complexity, mechanics, coherence, cohesion, reader/writer interaction, and content. In the validation stage of the study, raters assessed writing samples using the rating scale, and the quality of the rating scale was evaluated using several statistics of the Rasch model: rater separation, reliability and fit statistics, and scale step calibration. Raters' reactions to the scale were also collected via questionnaires and interviews. After receiving satisfactory statistical results and positive comments from raters, Knoch concluded that the theoretically-based and empirically-developed rating scale was useful for a L2 diagnostic writing test.

This literature review suggests that assessment techniques built on empirical sources are promising in that they substantiate the construct to be measured and draw on concrete rationales and evidence of assessment criteria. Empirical assessments can also create a dialogue among stakeholders who might attach different philosophies, values, meanings, or purposes to the assessment. In that dialogue, assessors play an active role as generators of assessment criteria and interpreters of assessment outcomes. The nature of context-embeddedness also significantly enhances communication, emphasizing that no assessment can take place in isolation from its context and users. These features are particularly relevant to the underlying concepts of diagnostic assessment; in a diagnostic assessment framework, a dialogue with assessment users can help to create consensus about the elements to be evaluated, and can help to keep them better informed about learners' particular strengths and weaknesses.

A unified assessment framework could therefore integrate the empirical approach and diagnostic assessment; evaluation criteria would be identified from real language performance and confirmed by theoretical accounts, and would then be used to build a diagnostic assessment scheme. In that assessment scheme, each criterion would represent a single evaluation element. Raters could then concentrate on one element at a time, without the distraction of having to consider many evaluation criteria simultaneously. Such a scheme can be created using empirically-derived diagnostic descriptors and may have the potential to maximize the diagnostic benefit of assessment for various users.

Assessment Components in ESL Writing

One way to identify assessment components in L2 writing is to examine rater perceptions and rating scales. Most studies used think-aloud verbal protocols to explore the assessment criteria that raters focus on while assessing written compositions (Cumming, 1990; Cumming, Kantor, & Powers, 2001, 2002; Lumley, 2002, 2005; Milanovic, Saville, & Shuhong, 1996; Sakyi, 2000; Smith, 2000; Vaughn, 1991). Cumming (1990) identified 28 decision-making and assessment criteria used by experienced assessors to evaluate L2 written compositions. These were categorized into four foci (self-control, content, language, and organization) and two strategies (interpretation and judgment). Each focus contained subcriteria further specifying rater evaluation behaviors or criteria. For example, a focus on language was broken down into (a) classifying errors, (b) editing phrases, (c) establishing the level of comprehensibility, (d) establishing error frequency, (e) establishing command of syntactic complexity, (f) establishing appropriateness of lexis, and (g) rating overall language use. Similar criteria were found by Cumming et al. (2001, 2002). They documented 27 decision-making processes exhibited by experienced writing assessors on ESL/English as a foreign language (EFL) compositions; these were further characterized by three foci (self-monitoring, rhetorical and ideational, and language) and two strategies (interpretation and judgment).

Research interest has also been directed toward the ways in which evaluation criteria for a rating scale could interact with rater perceptions and judgments. In a pivotal study by Vaughn (1991), nine raters verbalized their thinking processes during the rating process using a six-point holistic scale. Raters' comments were categorized into 14 general evaluation criteria, and the six most frequently mentioned assessment elements were identified as (a) quality of content, (b) legibility of handwriting, (c) tense/verb problem, (d) punctuation/capitalization error, (e) quality of introduction, and (f) morphology/word form error. Of these, Vaughn found that raters most frequently focused on content problems.

In a large-scale EFL testing context involving two Cambridge examinations (First Certificate in English [FCE] and Certificate of Proficiency in English [CPE]), Milanovic et al. (1996) asked 16 raters from diverse backgrounds to report the evaluation components that they focused on when assessing EFL writing. A wide range of elements were identified, including (a) length, (b) legibility, (c) grammar, (d) structure, (e) communicative effectiveness, (f) tone, (g) vocabulary, (h) spelling, (i) content, (j) task realization, and (k) punctuation. They also found that raters focused more on vocabulary and content in high-level essays, and on communicative effectiveness and task realization in intermediate-level essays.

Similar findings were reported by Smith (2000), who examined the ways in which raters interpreted and applied evaluation criteria in the Certificates in Spoken and Written English (CSWE). Based upon six raters' verbal accounts, nine textual features were identified that described the examinees' writing performance: (a) grammar, (b) organization, (c) cohesion, (d) sentence structure, (e) punctuation/capitalization, (f) spelling, (g) handwriting, (h) length of text, and (i) lexical choice. Conversely, the study by Sakyi (2000) sought more global assessment criteria. Six raters were asked to describe their rating processes using a five-point scale, with their comments categorized as focusing on (a) content and organization, (b) grammatical and mechanical errors, and (c) sentence structure and vocabulary.

In a more recent study, Lumley (2002) examined the ways in which four experienced raters applied a rating scale to L2 written compositions. The scale provided to the raters was developed for the writing subtest of the Special Test of English Proficiency (STEP) and had four evaluation criteria: (a) task fulfillment and appropriateness, (b) conventions of presentation, (c) cohesion and organization, and (d) grammatical control. The findings indicated that even though the scale content seemed to accurately reflect what raters pay attention to, there were conflicts among the descriptors within the same criteria at the same level. The raters also focused on two additional evaluation criteria (quantity of ideas and explicit cohesive devices) that were not included in the STEP rating scale.

Methodology

Research Overview

This study was conducted in two phases; in Phase 1, descriptors that are relevant to the construct of ESL academic writing were developed. In Phase 2, the use of the descriptors was validated using multiple analyses.

Participants

ESL academic writing teachers

Thirteen experienced ESL teachers participated in the study, with nine participating in Phase 1 and seven participating in Phase 2. Three teachers participated in both phases. All ESL teachers were native English speakers with extensive experience (5 to 25 years; average 10.69 years) of teaching ESL writing to adult learners. Nine teachers held or were pursuing a graduate degree (mostly a master's degree) in areas related to linguistics and language education, and 10 held an ESL teaching certificate.

ESL academic writing experts

Four doctoral students in Second Language Education (hereafter referred to as ESL writing experts) participated in the study. All of them had extensive research experience in various areas such as teacher feedback on ESL writing, motivation, writing conferencing, and process and assessment of writing, as well as had varying experience (3 to 12 years) of teaching writing to ESL learners.

Instruments

The writing samples used in this study were 80 Test of English as a Foreign Language^{TM} Internet-based test (TOEFL® iBT) independent essays written on two different prompts (40 essays × 2 prompts). The writing samples were requested from the Educational Testing Service (ETS) in the United States for research purposes. The two essay prompts were about (a) choosing subjects according to one's own interest (hereafter referred to as the subject prompt) and (b) the importance of cooperation (hereafter referred to as the cooperation prompt). The test-takers ranged from 14 to 42 years of age (*M* = 23.61, *SD* = 6.38) and took the TOEFL® iBT in both U.S. and non-U.S. test centers. Those who spoke Chinese as a first language accounted for the largest number of test-takers, followed by Spanish, Japanese, and Arabic. The length of the essay that they wrote was also diverse, ranging from 135 to 519 words (*M* = 318.54, *SD* = 77.33), and the essay scores awarded by ETS were normally distributed, with few lowest or highest scores (*M* = 3.46 [*Min* = 0, *Max* = 5], *SD* = 0.87).

Data Collection Procedure

Think-aloud verbal protocol session

Two possible verbal reporting methods (i.e., concurrent and immediate retrospective [Ericsson &

Simon, 1993]) were introduced to and chosen by the teachers. The concurrent think-aloud method required teachers to verbalize their thought processes while reading and providing diagnostic feedback on essays with no time delay and was thought to be effective in minimizing memory loss (Ericsson & Simon, 1993). The immediate retrospective think-aloud method allowed teachers to read an essay first either silently or aloud, and then to speak their thoughts aloud. Although the retrospective method increased the potential for memory loss, it would improve concentration by allowing teachers to read the essays without interruption (Ericsson & Simon, 1993). Nine teachers were provided with an explanation of the two think-aloud methods, and were allowed to choose the method they thought would work best for them. After trying both methods, three ESL teachers selected the concurrent method, and six chose the retrospective method. The teachers who preferred the retrospective method reported that the concurrent method interfered with the reading process and did not effectively elicit their natural cognitive responses. The teachers provided their think-aloud reports on 10 essays (five subject essays and five cooperation essays) in the presence of the interviewer. Each time the teachers completed a think-aloud report, they were interviewed in order to clarify any unclear statements or ambiguous comments they had made. The role of the interviewer was minimized as much as possible so as not to unduly influence the feedback. Each think-aloud and follow-up interview session lasted two to three hours. With the teachers' permission, all verbal reports and interviews were tape recorded and immediately transcribed.

Transcripts of each teacher's verbal reports ranged from 5,422 to 8,504 words (average: 7,227 words). Based upon grounded theory (Glaser & Strauss, 1967), each transcript was analyzed. Ambiguous or hard-to-interpret evaluative comments were excluded from the analysis, while comments that were too general, such as "good language," "good introduction," and "I love the thought", were disregarded because the analysis focused on identifying fine-grained diagnostic evaluation themes. The transcripts were thus coded at micro-level ESL writing skills in order to come up with specific diagnostic descriptors. In assessing a writer's vocabulary knowledge, for example, several different aspects were identified and coded (e.g., word sophistication, word variety, word choice, collocation, etc.) instead of having one general evaluation criterion called "vocabulary".

The analysis identified a total of 1,715 segments representing 39 evaluation themes or descriptors. I coded all 1,715 segments, and a second coder independently coded the original, uncoded, segmented transcripts of each teacher's think-aloud reports on two essays (515 segments; approximately 30.03% of all segments) in order to examine inter-coder reliability. The second coder was a PhD student specializing in Second Language Education with substantial knowledge of ESL writing. She was provided with a coding scheme consisting of 39 descriptors and agreed with the meaning of the descriptors before beginning work. When inter-coder reliability was examined, satisfactory agreement (450/515 segments, 87.38%) was found. Before the descriptors were subjected to the ESL academic writing experts' substantive review and refinement process, I preliminarily reviewed each descriptor based upon theories of ESL writing skills and knowledge and a variety of existing ESL writing assessment schemes in order to justify their theoretical grounds.

Descriptor review and refinement

In a focus group meeting, four ESL writing experts reviewed the 39 descriptors derived from the teachers' think-aloud verbal reports. The experts discussed whether each descriptor was clear, non-redundant, useful or relevant to ESL academic writing. When the wording of a descriptor was not clear, the experts edited it. After examining each descriptor, they were also asked whether the descriptor pool was comprehensive enough to cover all aspects of ESL academic writing. If any missing theoretical aspects were found, the experts were asked to add the aspects to the descriptor pool based upon existing theories of ESL academic writing. The refined descriptors were then accompanied by a yes or no binary response option to function as an assessment scheme. The yes/no response option and the lack of a scale in the assessment scheme were chosen to facilitate the rating process as in Upshur and Turner's (1995) EBB rating scale, improving the simplicity and clarity of the evaluation.

ESL essay assessment

Seven ESL writing teachers assessed 80 TOEFL iBT independent essays written on two prompts (40 essays × 2 prompts) using the refined descriptors accompanied by a yes or no binary response option. Each teacher assessed 30 essays, with 10 essays assessed by all seven teachers and the remaining 70 essays assessed by two different teachers. The 30 essays assigned to each teacher were ordered by the prompt and counterbalanced. Three teachers assessed essays that were written on the subject prompt first, and the other four teachers assessed essays that were written on the cooperation prompt first. A rater training session was held before the teachers began their essay assessment. The purpose of the training was not to clone the teachers to achieve high inter-rater reliability or to make them professionally certified raters of ESL writing, but to orient the teachers to the empirically-derived diagnostic descriptor-based assessment scheme. During the training, not only was each descriptor explained using concrete examples but also how to use the binary scale (yes or no response option) attached to each descriptor (for more information about the rater training and the binary scale, see Kim, 2010, 2011). The teachers' yes responses were treated as "1s," and no responses were treated as "0s" when they were compiled in Microsoft® Excel spreadsheets.

Data Analysis Procedure

The Many-faceted Rasch Model (MFRM) computer software, FACETS version 3.66.0 (Linacre, 2009) analyzed the ratings awarded by the seven teachers using the descriptors on the 80 essays. Ten essays were assessed by all seven teachers, and the remaining 70 essays were assessed by two different teachers so that the data matrix was partially crossed. In the model specification, four facets were specified: student, prompt, teacher, and descriptor. While the student and descriptor facets were centered by anchoring logit means at zero, the teachers were allowed to float because the analysis of interest focused on the teacher's behavior in using the descriptors. The prompt facet was entered as a dummy facet and anchored at preset values of 0.03 logits (for the cooperation prompt) and -0.03 logits (for the subject prompt), respectively.[3] Anchoring was necessary in order to connect the two separate essay subsets in which each student wrote a single essay on one prompt only (Linacre, 2009). The preset values of 0.03 and -0.03 logits were derived from a preliminary analysis that showed the subject prompt (difficulty measure = 0.03 logits) was more difficult than the cooperation subject (difficulty measure = -0.03 logits). The teacher's behavior in employing the descriptors was analyzed using multiple methods. First, teacher internal consistency was examined based on infit and outfit mean square values. In addition, inter-teacher reliability was examined in order to explore the degree to which one teacher agreed with others when using the descriptors. Three reliability indices were computed: (a) the percentage of exact agreement, (b) the correlation between a single rater and the rest of the raters (SR/ROR), and (c) the percentage of the teachers' ratings that agreed on each descriptor.

Results

What Empirically-derived Descriptors are Relevant to the Construct of ESL Academic Writing?

Initial review of the teachers' think-aloud verbal protocols indicated that ESL teachers considered a variety of subcomponents of ESL writing skills and knowledge when determining the quality of an essay. The sheer amount of data they provided also confirmed the depth and comprehensiveness of the teachers' accounts. The coding of these protocols resulted in the identification of 39 recurrent writing subskills that

[3] Dummy facets are intended to investigate interactions without affecting main effects (Linacre, 2009).

formed the pool of the descriptors (see Appendix). The Appendix lists (a) all 39 descriptors of ESL academic writing, (b) a representative think-aloud verbal protocol that exemplifies each descriptor, and (c) the number of times/percentage that each descriptor occurred during the think-aloud verbalization. The total descriptor tally was 1,715, of which spelling (D34[4], 6.06%), essay structure (D09, 5.42%), verb tense (D22, 4.90%), tone and register (D39, 4.78%), and essay clarity (D02, 4.72%) were the five most frequently mentioned. By contrast, essay focus (D14, 0.87%), indentation (D37, 0.58%), use of conditional verbs (D28, 0.52%), syntactic variety (D16, 0.35%), and paraphrasing (D38, 0.29%) were the least frequently commented.

I preliminarily reviewed these 39 descriptors on the basis of theories of ESL writing and various existing ESL writing assessment schemes before subjecting them to the ESL academic writing experts' substantive review and refinement process. The descriptors showed that teachers paid considerable attention to the extent to which writers satisfactorily addressed the given topic (i.e., D01-D08). Although content fulfillment is not generally included in the research scope of traditional second language acquisition (SLA) studies examining the development of ESL writing (e.g., Wolfe-Quintero, Inagaki, & Kim, 1998), contemporary L1 and L2 writing theories do agree that it is a central component of good writing. For example, Grabe and Kaplan (1996) considered topical knowledge or knowledge of the world to be a parameter that determines writing performance. Similarly, research on rater perceptions and behaviours has verified content fulfillment as an important consideration. These empirical and theoretical accounts support the idea that an important criterion of written text assessment is the extent to which an essay fulfills the content requirement.

Teachers also felt that organizational effectiveness determined the quality of an essay (i.e., D09-D15). This finding was reasonable because the ability to coherently organize ideas has long had a place in writing instruction and research; Halliday and Hasan (1976) conceptualized cohesion and coherence as ways in which the textual structure is tied together in extended discourse. Similarly, Canale (1983) and Grabe and Kaplan (1996) suggested that unified text can be attained through cohesion in form and coherence in meaning. Most analytic rating scales have also highlighted the importance of cohesion and coherence in written discourse; Hamp-Lyons and Henning (1991) included organization as an independent evaluation criterion in their ESL writing scale, and the International English Language Testing System (IELTS) writing rating scale considers coherence and cohesion an important ESL writing subskill.

Teachers were also concerned about grammatical knowledge. Grammatical accuracy is one of the most-researched topics in SLA studies on writing development and a central theme in ESL writing instruction and research. The achievement of ESL writing skills has traditionally been defined as the mastery of discrete grammar knowledge and the ability to produce linguistically accurate written text (Kepner, 1991). Therefore, the presence of grammatical errors is the primary language-related factor affecting ESL composition teachers, suggesting that they are excessively concerned with eradicating grammatical errors in students' writing. This study reinforced that recurrent grammatical errors were teachers' primary concern in students' writing assessments. Specifically, teachers' attention was focused on more fine-grained, specific aspects of grammatical knowledge such as verb tense (D22), article use (D26), and preposition use (D25). This finding is noteworthy because most ESL writing scales measure learners' grammatical competence at a macro-level, obscuring students' performance on specific grammar components.

Teachers also showed considerable interest in various aspects of students' vocabulary use. Their attention to the quality of written vocabulary (sophistication [D29], diversity [D30], accuracy [D31], and collocation [D32]) echoed the idea that a good vocabulary leads to good writing. The importance of vocabulary in a written text is supported by theoretical frameworks of L1 and L2 writing (e.g., Grabe & Kaplan, 1996) and empirical SLA studies (e.g., Engber, 1995; Laufer, 1991; Laufer & Nation, 1995). Several ESL writing scales have also recognized this: Brown and Bailey (1984) and Jacobs, Zinkgraf, Wormuth, Hartfiel, and Hughey (1981) emphasized the close association between vocabulary and writing

[4] D34 indicates Descriptor 34. Hereafter, the notation, "D + number" will indicate "Descriptor + number."

performance in their ESL academic writing scales, as did the IELTS rating scale, which included lexical resource as one constituent subscale. Similarly, Mullen (1977) found that vocabulary appropriateness accounted for 84.4% of the variance in overall writing performance. These research findings suggest that vocabulary is indeed an indispensable factor in determining the quality of writing.

Writing mechanics constituted another area that drew teachers' attention; however, as Polio (2001) rightly pointed out, mechanics has not been a central concern for language researchers. Very little research has examined writers' mechanical proficiency in relation to their writing development. Indeed, Polio speculated as to whether mechanics should even be considered a part of writing construct, since the various aspects of mechanics (such as spelling [D34], punctuation [D35], capitalization [D36], and indentation [D37]) are not conceptually related to each other, making it difficult to form a unitary construct. Yet, mechanical knowledge does play a significant role in the writing process. Knowledge of written code is achieved through the mastery of orthography, spelling, punctuation, and formatting conventions (Grabe & Kaplan, 1996), and a writer's intended meaning would be obscured and lost without their appropriate use. The value of mechanics can also be found in existing writing rating scales. Brown and Bailey (1984) and Jacobs et al. (1981) considered mechanics a component of their academic writing scales.

The theoretical review of the descriptors suggested that the five writing skills appear to encompass all aspects of the 39 descriptors: (a) content fulfillment, (b) organizational effectiveness, (c) grammatical knowledge, (d) vocabulary use, and (e) mechanics. This skill configuration was consistent with the theoretical discussions and existing assessment schemes discussed earlier. The four ESL academic writing experts then reviewed and refined these descriptors focusing on whether each one was clear, non-redundant, useful, and relevant to ESL academic writing. The results showed that three descriptors were identified as problematic: Ideas reflect the central focus of the essay, without digressing (D14); Conditional verb forms are used appropriately (D28); and the essay prompt is well-paraphrased, and is not replicated verbatim (D38). The experts also pointed out that D14 overlapped with D11-D13, and that D28 and D38 addressed relevant, but too specific, aspects of ESL writing. Indeed, these descriptors were rarely mentioned in the ESL teachers' think-aloud verbal reports, with total comments accounting for less than 1% of all verbal protocols. The experts also suggested combining two descriptors (Syntactic variety is demonstrated in this essay [D16] and Complex sentences are used effectively [D17]) to form one new descriptor, such as "This essay demonstrates syntactic variety, including simple, compound, and complex sentence structures". The review and refinement process resulted in the elimination of three descriptors altogether, and the combination of two other descriptors into one, for a final total of 35 descriptors. The clarity of the descriptors was also reviewed. The experts read each descriptor iteratively and edited it to make an easy and clear descriptor for teachers to use. Twenty-two descriptors were edited in this manner, with most of the editing focused on specific wordings to minimize ambiguity. The experts then examined the descriptors for distinctiveness and comprehensiveness and confirmed each descriptor to be independent of the others and comprehensive enough to cover all aspects of ESL academic writing (see Table 1). No new descriptors were added to the descriptor pool. Finally, the refined 35 descriptors were then accompanied by a yes or no response option.

Are Teachers Able to Use the Empirically-derived Diagnostic Descriptors in an Internally Consistent Manner?

The extent to which the teachers were internally consistent in using the final 35 descriptors was examined based upon teacher fit statistics. Teacher fit statistics refers to the degree to which each teacher is internally consistent when awarding ratings. Different rules of thumb are applied for interpreting fit statistics and for setting upper and lower limits because they are more or less context-dependent (Myford & Wolfe, 2004). When a test of interest is used to make a high-stakes decision, tight quality control limits (such as mean squares of 0.8 to 1.2) are set; however, if the stakes are low, looser limits are acceptable.

TABLE 1
The Final 35 Descriptors

Descriptor
EDD01. This essay answers the question.
EDD02. This essay is written clearly enough to be read without having to guess what the writer is trying to say.
EDD03. This essay is concisely written and contains few redundant ideas or linguistic expressions.
EDD04. This essay contains a clear thesis statement.
EDD05. The main arguments of this essay are strong.
EDD06. There are enough supporting ideas and examples in this essay.
EDD07. The supporting ideas and examples in this essay are appropriate and logical.
EDD08. The supporting ideas and examples in this essay are specific and detailed.
EDD09. The ideas are organized into paragraphs and include an introduction, a body, and a conclusion.
EDD10. Each body paragraph has a clear topic sentence tied to supporting sentences.
EDD11. Each paragraph presents one distinct and unified idea.
EDD12. Each paragraph is connected to the rest of the essay.
EDD13. Ideas are developed or expanded well throughout each paragraph.
EDD14. Transition devices are used effectively.
EDD15. This essay demonstrates syntactic variety, including simple, compound, and complex sentence structures.
EDD16. This essay demonstrates an understanding of English word order.
EDD17. This essay contains few sentence fragments.
EDD18. This essay contains few run-on sentences or comma splices.
EDD19. Grammatical or linguistic errors in this essay do not impede comprehension.
EDD20. Verb tenses are used appropriately.
EDD21. There is consistent subject-verb agreement.
EDD22. Singular and plural nouns are used appropriately.
EDD23. Prepositions are used appropriately.
EDD24. Articles are used appropriately.
EDD25. Pronouns agree with referents.
EDD26. Sophisticated or advanced vocabulary is used.
EDD27. A wide range of vocabulary is used.
EDD28. Vocabulary choices are appropriate for conveying the intended meaning.
EDD29. This essay demonstrates facility with appropriate collocations.
EDD30. Word forms (noun, verb, adjective, adverb, etc.) are used appropriately.
EDD31. Words are spelled correctly.
EDD32. Punctuation marks are used appropriately.
EDD33. Capital letters are used appropriately.
EDD34. This essay contains appropriate indentation.
EDD35. Appropriate tone and register are used throughout the essay.

Note. EDD01 indicates the finalized empirically-derived diagnostic (EDD) descriptor 01. Hereafter, “EDD + number” indicates “the finalized EDD descriptor + number.”

In this study, the mean square values of 0.5 and 1.5 were set as the lower and upper quality control limits, respectively (Lunz & Stahl, 1990), since this study examines the rating patterns of teachers in a small-scale classroom setting rather than in a high-stakes test setting. A mean square value less than 0.5 indicates a lack of variability in their rating, while a mean square value greater than 1.5 indicates a high degree of inconsistency in the ratings. Table 2 presents several of the statistics associated with the teacher facet; in particular, the fifth and sixth columns display the infit and outfit mean squares for each teacher. All infit and outfit mean squares were within the range of 0.5 and 1.5, indicating that all of the teachers were internally consistent in their ratings.

TABLE 2
Teacher Measure Statistics

Teacher	Obsvd Average	Measure (logits)	Model S.E.	Infit MnSq	Outfit MnSq	Corr. PtBis	Exact Obs %	Agree. Exp %
Angelina	0.6	-0.35	0.07	1.02	1.02	0.20	65.9	59.7
Ann	0.6	-0.64	0.07	1.01	1.05	0.22	64.1	60.5
Beth	0.5	0.15	0.07	0.89	0.84	0.29	65.9	57.8
Brad	0.6	-0.37	0.07	1.01	1.01	0.25	63.7	60.9
Esther	0.6	-0.93	0.07	1.04	1.15	0.20	63.9	60.4
Susan	0.7	-0.71	0.07	1.02	1.07	0.23	64.7	61.4
Tom	0.6	-0.39	0.07	1.00	0.97	0.25	64.6	60.1
M	0.6	-0.46	0.07	1.00	1.01	0.23		
SD	0.1	0.32	0.00	0.05	0.09	0.03		
RMSE (Model) = 0.07				Adj. *SD* = 0.31				
Separation = 4.38				Separation (not inter-rater) Reliability = 0.95				
Fixed (all same) chi-square = 143.2				d.f. = 6				
Significance (probability) = 0.00				Inter-Rater agreement opportunities: 9,701				
Exact agreements: 6,275 = 64.7%				Expected: 5832.1 = 60.1%				

To What Extent do Teachers Agree with Each Other When Using the Empirically-derived Diagnostic Descriptors?

Three approaches were used in order to examine the degree of agreement between teacher assessments. The first used a percentage of exact agreement, which indicated the percentage of times that each teacher provided exactly the same ratings as another teacher under identical circumstances. The agreement statistics and expected values were provided by FACETS. As the eighth column of Table 2 shows, the exact observed agreement of the teachers ranged from 63.7% to 65.9% (*M* = 64.7%). Although this range does not seem to support the idea of substantial agreement among teachers, it is reasonable considering that the teachers were not trained as professional raters of ESL writing. A similar agreement pattern has been found in other writing assessment research. Barkaoui (2008) reported that teachers' agreement reached 22.4% when they used a nine-point holistic rating scale and that their agreement was 23.1% when they used a nine-point analytic rating scale. When his teacher group was examined further, novice teachers showed 20.0% agreement, while experienced teachers exhibited 26.3% agreement. His findings seem to confirm the difficulty of achieving high agreement among teachers who are not trained as professional assessment raters, echoing this study's finding.

However, when well-trained certified raters are involved in a high-stakes ESL writing assessment, a fair amount of agreement can be achieved. Knoch (2009) examined the functionality of two analytic ESL writing scales: the Diagnostic English Language Needs Assessment (DELNA) and a newly developed diagnostic scale, and reported somewhat fair, but still unsubstantial, agreement. The two rating scales were developed to assess student writing skills and consisted of six levels (in the case of the DELNA) and four to six levels (in the case of the new diagnostic scale). When raters used the DELNA rating scale, their agreement ranged from 33% to 41.7% (*M* = 37.92, *SD* = 2.49); when the new diagnostic scale was used, agreement ranged from 36.1% (for a six-level scale) to 61.9% (for a four-level scale) (*M* = 51.15, *SD* = 7.94). That the raters were well-trained certified professionals must have contributed to this fair or moderate agreement, but it still indicates that it is extremely difficult for raters to achieve substantial agreement on writing assessments, possibly because of the inherently subjective nature of the task.

The second approach to examining inter-rater reliability was a correlation between a single rater and the rest of the raters (SR/ROR). SR/ROR correlation indicates the degree to which one particular rater (i.e., the single rater) rank-orders examinees in a manner consistent with all other raters. According to Myford and Wolfe (2004), SR/ROR correlations greater than 0.7 are considered high for an assessment in which a multiple-level rating scale is involved, whereas SR/ROR correlations less than 0.3 are thought to be somewhat low. Still, they caution that the control limit must be relaxed as the number of scale categories decreases: for example, they report that SR/ROR correlations as low as 0.2 are common in dichotomous ratings.[5] As the seventh column of Table 2 illustrates, teachers' SR/ROR correlations in this study ranged from 0.20 to 0.29 (*M* = 0.23, *SD* = 0.03), suggesting that each teacher rank-ordered students in a manner similar to that of the other teachers. [6]

Further analysis was conducted in order to examine the extent to which the teachers agreed on each individual descriptor. The percentage of teachers' ratings that agreed on each descriptor per essay was calculated, and the mean and standard deviations of the agreements on 10 essays were examined. Ratings were derived from the 10 essays that were assessed by all of the teachers. As Table 3 shows, teachers had the highest agreement on EDD16 (word order; agreement = 90%) and exhibited the lowest agreement on EDD13 (idea development; agreement = 61.43%). When the descriptors that elicited high agreement (> 85%) were examined, most were related to discrete grammatical and mechanical knowledge (e.g., EDD16, EDD23, EDD31, and EDD32). On the other hand, when the descriptors that elicited low agreement (< 70%) were examined, they were found to be associated with global content and organizational skills (e.g., EDD01, EDD05, EDD06, EDD07, and EDD13). These results are consistent with Milanovic et al.'s (1996) findings suggesting that essay content is the most subjective component element because raters' personal reactions might significantly affect their rating.

TABLE 3
Teacher Agreement on Descriptors

Descriptor	Agreement (%)	*SD*	Descriptor	Agreement (%)	*SD*
EDD01	65.71	13.80	EDD19	77.14	13.80
EDD02	80.00	12.05	EDD20	82.86	17.56
EDD03	74.29	17.56	EDD21	80.00	21.51
EDD04	70.00	18.38	EDD22	77.14	18.07
EDD05	68.57	14.75	EDD23	85.71	15.06
EDD06	65.71	13.80	EDD24	78.57	15.43
EDD07	68.57	14.75	EDD25	84.29	18.38
EDD08	74.29	13.13	EDD26	77.14	13.80
EDD09	80.00	15.36	EDD27	81.43	15.13
EDD10	71.43	15.06	EDD28	77.14	15.36
EDD11	70.00	14.21	EDD29	82.86	14.75
EDD12	75.71	16.56	EDD30	78.57	13.88
EDD13	61.43	6.90	EDD31	87.14	18.38
EDD14	77.14	16.77	EDD32	85.71	15.06
EDD15	78.57	15.43	EDD33	84.29	15.72
EDD16	90.00	15.13	EDD34	72.86	18.38
EDD17	77.14	18.07	EDD35	85.71	9.52
EDD18	77.14	19.28			

Discussion

The first research question discussed the development of descriptors that can diagnose ESL academic writing ability, with a focus on empirically identifying evaluation elements that reflect knowledge, processes, and strategies consistent with the construct of ESL writing in an academic context. It was thus critical to empirically identify descriptors operationalizing the ESL writing skills required in an academic context. Theoretical analysis was also of great importance to justify and confirm the identified descriptors. Considering that the construct of ESL writing is multi-faceted and complicated (Cumming, 2001; Cumming et al., 2000), it was necessary to identify fine-grained and separable assessment elements so

[5] An SR/ROR correlation near or less than 0 indicates low inter-rater reliability.

[6] SR/ROR correlations are referred to as point-biserial correlations in FACETS analysis.

that it would be possible to implement diagnostic assessment.

To this end, multiple empirical sources were sought from diverse perspectives. Not only was the incidence of writing performance observed using real student writing samples, but also assessment elements were elicited from teachers' think-aloud verbalization on ESL essays. These verbal accounts provided rich descriptions of ESL academic writing ability, resulting in 39 descriptors. The descriptors were then theoretically examined and then reviewed by four ESL academic writing experts. The review and refinement process further confirmed the soundness of the descriptors; three descriptors were eliminated, and two descriptors were merged into one. The final 35 descriptors were empirically-derived, concrete, fine-grained, and addressed all aspects of ESL writing skills, including content fulfillment, organizational effectiveness, grammatical knowledge, vocabulary use, and mechanics. The greatest number of descriptors associated with grammatical knowledge was also reasonable considering that students greatly desire specific feedback on grammatical problems in their writing (Cohen & Cavalcanti, 1990; Ferris, 1995; Hedgcock & Lefkowitz, 1994; Leki, 1991) and teachers are concerned with eradicating grammatical errors in student writing.

The evidence gathered throughout the descriptor development procedure suggests that the descriptors accurately represent the multidimensional construct of ESL academic writing. The teachers' think-aloud verbal data were a valuable empirical source that substantiated the construct being measured and provided concrete rationales and evidence justifying the selected assessment criteria. The theoretical analysis further confirmed that the descriptors were neither atheoretical nor free of theory. These approaches were particularly well-aligned with the concepts of diagnostic assessment because it enabled teachers to be active generators of assessment elements and interpreters of assessment outcomes rather than passive listeners. In a diagnostic assessment framework, a dialogue with assessment users and developers can help to create consensus about the elements to be evaluated, and can help to keep diverse educational clientele better informed about the assessment outcomes.

The second and third research questions validated the use of the descriptors by examining the potential impact of the variability associated with sampling conditions of observation. Teacher facet is the primary source of variability suspected of preventing accurate inferences about student ESL academic writing ability. If the student writing scores obtained from a sample of teachers cannot be generalized beyond that specific set of teachers, it will undermine valid score interpretations. Two approaches were used to explore this suspected variability. First, teacher internal consistency was examined based on infit and outfit mean square values. In addition, inter-teacher reliability was examined in order to explore the degree to which one teacher agreed with others when using the descriptors. Three reliability indices were computed: (a) the percentage of exact agreement, (b) the correlation between a single rater and the rest of the raters (SR/ROR), and (c) the percentage of the teachers' ratings that agreed on each descriptor.

The teacher fit statistics of the MFRM analysis indicated that all of the teachers exhibited internally consistent rating patterns when using the descriptors. These results suggest that teachers are able to use the descriptors in an internally consistent manner. However, a mixed result was found when agreement rates among teachers were investigated. While the correlation between a single rater and the rest of the raters (SR/ROR) indicated that each teacher might have rank-ordered students in a manner similar to that of the other teachers, teacher agreement statistics reported that for a somewhat low or moderate percentage of times, each teacher provided exactly the same ratings as another teacher under identical circumstances. In addition, when teacher agreement rates were examined at the descriptor level, teachers showed high agreement (> 85%) on descriptors assessing discrete grammatical and mechanical knowledge, but low agreement (< 70%) on descriptors assessing global content and organizational skills. These results indicate that it might be difficult to claim that a particular teacher's assessment of student writing performance is generalizable beyond that specific teacher. However, as discussed earlier, the reported reliability indices must be interpreted carefully because the teachers were not well-trained certified professional writing assessment raters, and dichotomous ratings (rather than polytomous ratings) were used in the assessment.

Overall, the findings of the third research questions present a somewhat fuzzy picture of the generalizability of the scores derived from the descriptors. Unlike traditional fixed-response assessments (such as multiple-choice tests), the presence of raters and tasks in performance assessment adds a new dimension of interaction, making it even more crucial to monitor reliability and validity. A greater number of raters and tasks in an assessment would be desirable in order to improve consistency from one performance sample to another, but this is not always possible due to limited resources. The problem becomes more serious when one considers that the descriptors were developed to be used for diagnostic assessment purposes in a small-scale classroom, where relatively few resources are allocated. One way of resolving this problem would be to standardize essay prompts by providing clear specifications. Another way would be to train teachers on a continuous basis, since effective training would help teachers to use the descriptors consistently and reliably. Care must be taken, however, because high inter-teacher reliability could counter the contextual validity gained from using the descriptors. As the descriptors were developed to be used in small-scale classroom assessment, which is typically provided by just one teacher, high inter-teacher reliability would not be crucial in such learning-oriented assessment cases where consistency of judgment is less important than richness of judgment and could even threaten the valid use of the descriptors.

Conclusion and Implications

This study developed and validated the 35 empirically-derived diagnostic descriptors for use in a small-scale classroom assessment in ESL academic writing. These descriptors were fine-grained, addressing all aspects of ESL writing skills, including content fulfillment, organizational effectiveness, grammatical knowledge, vocabulary use, and mechanics. Multiple psychometric analyses also validated the usefulness of the descriptors regarding sources of score variability.

The results of this study therefore support the idea that an empirical approach is useful when developing an assessment scheme (Brindley, 1998; Fulcher, 1993; Upshur & Turner, 1995). As many researchers have pointed out, the most serious problem with intuition-based or a priori rating scales is that it is not always clear how the scale descriptors were created (or assembled) and calibrated (e.g., Brindley, 1998; Chalhoub-Deville, 1997; de Jong, 1988; Lantolf & Frawley, 1985; North, 1993; Pienemann, Johnson, & Brindley, 1988; Upshur & Turner, 1995). In light of these problems, this study aimed to demonstrate the potential of an assessment scheme developed using an empirical approach. Not only did teachers' think-aloud verbal protocols provide rich verbal descriptions of the assessment criteria to be assessed, but a theoretical analysis also confirmed the assessment criteria. These findings demonstrate the effectiveness of the empirical approach to assessment scheme development, and underscore the importance of its use for diagnostic purposes, as the identification of specific assessment elements is the most important procedure in implementing diagnostic assessment (Jang, 2009; Lee & Sawaki, 2009a; Lee & Sawaki, 2009b; Sawaki, Kim, & Gentile, 2009).

This study further reconceptualized current L2 writing scale classifications. Despite an increasing need for diagnostic assessment, very few scales (e.g., Knoch's [2009] diagnostic ESL academic writing scale) have been developed to offer such assessment in L2 academic writing. In the L2 writing assessment literature, rating scales are classified primarily as holistic, analytic, or primary trait scales (based on scoring methods) or as user-oriented, assessor-oriented, or constructor-oriented (based on assessment purpose), with little consideration of their diagnostic nature. In response, this study contributed to the current L2 writing scale literature by developing and validating diagnostic ESL writing descriptors.

Several methodological limitations and suggestions should be noted. First, the results should be limited to the specific contexts in which the study was conducted. Only TOEFL® iBT independent essays were included in the ESL academic writing sample. Therefore, generalizations of the research outcomes to other contexts will not support more valid interpretations of the study. The use of other validation approaches is also recommended. In this study, the use of the descriptors was empirically validated solely through the examination of the reliability indices, which failed to offer a full account of the usefulness of the descriptors. The descriptors could be further investigated using qualitative research methods such as verbal protocols and in-depth interviews in order to explore the ways in which teachers evaluate the effectiveness of the descriptors. Finally, it is suggested that the use of the descriptors be compared to other assessment tools, focusing on whether they are more accurate, easier to use, and more highly favored by teachers.

Acknowledgements

This work was supported by research grants from Daegu Catholic University in 2018. The responsibility for any errors that remain is solely mine, and any opinions expressed in this article are not necessarily those of Educational Testing Service.

The Author

Youn-Hee Kim is an associate professor in the Department of English Education at Daegu Catholic University, South Korea. Her primary research interests are second language assessment and testing. She has published numerous papers in *Applied Linguistics*, *Language Learning*, and *Language Testing*, and has been working as a reviewer of international journals such as *Applied Linguistics*, *Language Assessment Quarterly*, *Language Testing*, and *Language Learning.*

Department of English Education
Daegu Catholic University
13-13 Hayang-ro, Hayang-eup
Kyungsan-si, Kyungpook
South Korea, 38430
Tel: +82-53-850-3133
Email: younkim@cu.ac.kr

References

Alderson, J. C. (1991). Bands and scores. In J. C. Alderson & B. North (Ed.), *Language testing in the 1990s* (pp. 71-86). London: Macmillan.

Barkaoui, K. (2008). *Effects of scoring method and rater experience on ESL essay rating processes and outcomes* (Unpublished doctoral dissertation). University of Toronto, Canada.

Brindley, G. (1998). Describing language development? Rating scales and SLA. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 112-140). Cambridge: Cambridge University Press.

Brown, J. D., & Bailey, K. (1984). A categorical instrument for scoring second language writing skills. *Language Learning, 34,* 21-42.

Canale, M. (1983). From communicative competence to communicative language pedagogy. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication* (pp. 2-27). London: Longman.

Chalhoub-Deville, M. (1997). Theoretical models, assessment frameworks and test construction. *Language Testing, 14,* 16-33.

Cohen, A. D., & Cavalcanti, M. (1990). Feedback on compositions: Teacher and student verbal reports. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 155-177). Cambridge: Cambridge University Press.

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

Cumming, A. (2001). The difficulty of standards, for example in L2 writing. In T. Silva & P. Matsuda (Eds.), *On second language writing* (pp. 209-229). Mahwah, NJ: Lawrence Erlbaum.

Cumming, A., Kantor, R., Powers, D. Santos, T., & Taylor, C. (2000). *TOEFL 2000 writing framework: A working paper.* TOEFL Monograph Series, Report No. 18. Princeton, NJ: Educational Testing Service.

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. Princeton, New Jersey: Educational Testing Service.

Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal, 86,* 67-96.

de Jong, J. (1988). Rating scales and listening comprehension. *Australian Review of Applied Linguistics, 11,* 73-87.

Engber, C. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing, 4,* 139-155.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data.* Cambridge, MA: MIT Press.

Ferris, D. (1995). Student reactions to teacher response in multiple-draft composition classrooms. *TESOL Quarterly, 29*, 3-53.

Fulcher, G. (1993). *The construction and validation of rating scales for oral tests in English as a foreign language* (Unpublished doctoral dissertation). University of Lancaster, UK.

Fulcher, G. (2003). *Testing second language speaking.* London: Pearson Longman.

Glaser, B., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research.* Chicago; IL: Aldine.

Grabe, W., & Kaplan, R. (1996). *Theory and practice of writing.* New York: Longman.

Grove, E., & Brown, A. (2001). Tasks and criteria in a test of oral communication skills for first-year health science students. *Melbourne Papers in Language Testing, 10,* 37-47.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English.* London: 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,* 337-373.

Hedgcock, J., & Lefkowitz, N. (1994). Feedback on feedback: Assessing learner receptivity to teacher response in L2 composing. *Journal of Second Language Writing, 3,* 141-163.

Jacobs, H., Zinkgraf, S., Wormuth, D., Hartfiel, V., & Hughey, J. (1981). *Testing ESL composition: A practical approach.* Rowley, MA: Newbury House.

Jang, E. E. (2009). Demystifying a Q-Matrix for making diagnostic inferences about L2 reading skills. *Language Assessment Quarterly, 6,* 210-238.

Kepner, C. (1991). An experiment in the relationship of types of written feedback to the development of second-language writing skills. *The Modern Language Journal, 75,* 305-313.

Kim, Y-H. (2010). *An argument-based validity inquiry into the Empirically-derived Descriptor-based Diagnostic (EDD) assessment in ESL academic writing* (Unpublished doctoral dissertation). University of Toronto, Canada.

Kim, Y-H. (2011). Diagnosing EAP writing ability using the Reduced Reparameterized Unified Model. *Language Testing, 28,* 509-541.

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

Lantolf, J. P., & Frawley, W. (1985). Oral proficiency testing: A critical analysis. *The Modern Language Journal, 69,* 337-345.

Laufer, B. (1991). The development of L2 lexis in the expression of the advanced learner. *The Modern Language Journal, 75,* 440-448.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics, 16,* 307-322.

Lee, Y-W., & Sawaki, Y. (2009a). Cognitive diagnosis and Q-matrices in language assessment. *Language Assessment Quarterly, 6,* 169-171.

Lee, Y-W., & Sawaki, Y. (2009b). Cognitive diagnosis approaches to language assessment: An overview. *Language Assessment Quarterly, 6,* 172-189.

Leki, I. (1991). The preferences of ESL students for error correction in college level writing classes. *Foreign Language Annals*, *24*, 203-218.

Linacre, J. M. (2009). *A user's guide to Facets: Rasch-model computer programs.* Version3.66.0 [Computer software and manual]. Retrieved October 21, 2009, from www.winsteps.com.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing, 19,* 246-276.

Lumley, T. (2005). *Assessing second language writing: The raters' perspective.* Frankfurt: Peter Lang.

Lunz, M. E., & Stahl, J. A. (1990). Judge severity and consistency across grading periods. *Evaluation and the Health Professions, 13,* 425-444.

Luoma, S. (2004). *Assessing speaking.* Cambridge: Cambridge University Press.

Matthews, M. (1990). The measurement of productive skills: Doubts concerning the assessment criteria of certain public examinations. *ELT Journal, 44,* 117-121.

Milanovic, M., Saville, N., & Shuhong, S. (1996). A study of the decision making behavior of composition markers. In M. Milanovic & N. Saville (Eds.), *Studies in language testing 3: Performance testing, cognition and assessment* (pp. 92-111). Cambridge: Cambridge University Press.

Mullen, K. A. (1977). Using rater judgements in the evaluation of writing proficiency for nonnative speakers of English. In H. D. Brown, C. A. Yorio, & R. H. Crymes (Eds.), *On TESOL 77: Teaching and learning English as a second language: Trends in research and practice* (pp. 309-320). Washington, D.C.: TESOL.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. In Smith, Jr., E. V. & Smith, R. M. (Ed.), *Introduction to Rasch measurement* (pp. 460-517). Maple Grove, MN: JAM Press.

North, B. (1993). *The development of descriptors on scales of language proficiency.* Washington, DC: National Foreign Language Center.

North, B. (1995). The development of a common framework scale of descriptors of language proficiency based on a theory of measurement. *System, 23,* 445-465.

North, B. (1996). *The development of a common framework scale of descriptors of language proficiency based on a theory of measurement* (Unpublished doctoral dissertation). Thames Valley University.

North, B., & Schneider, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing, 15,* 217-263.

Pienemann, M., Johnson, M., & Brindley, G. (1988). Constructing an acquisition-based procedure for second language assessment. *Studies in Second Language Acquisition, 10,* 217-243.

Polio, C. (2001). Research methodology in second language writing: The case of text-based studies. In T. Silva & P. Matsuda. (Eds.), *On second language writing* (p. 91-116). Mahwah, NJ: Erlbaum.

Pollitt, A., & Murray, N. L. (1996). What raters really pay attention to? In M. Milanovic, & N. Saville (Ed.), *Studies in language testing 3: Performance testing, cognition and assessment* (pp. 74-91). Cambridge: Cambridge University Press.

Sakyi, A. A. (2000). Validation of holistic scoring for ESL writing assessment: How raters evaluate compositions. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium* (pp. 129-152), Orlando, Florida. Cambridge: Cambridge University Press.

Sawaki, Y., Kim, H-J., & Gentile. C. (2009). Q-matrix construction: Defining the link between constructs and test items in large-scale reading and listening comprehension assessments. *Language*

Assessment Quarterly, 6, 172-189.

Smith, D. (2000). Rater judgments in the direct assessment of competency-based second language writing ability. In G. Brindley (Ed.), *Studies in immigrant English language assessment* (pp. 159-189). Sydney, Australia: National Centre for English Language Teaching and Research, Macquarie University.

Turner, C. E., & Upshur, J. A. (1996). Developing rating scales for the assessment of second language performance. In G. Wigglesworth & C. Elder (Eds.), *The language testing cycle: From inception to washback* (pp. 55-79). Melbourne: Australian Review of Applied Linguistics.

Upshur, J. A., & Turner, C. E. (1995). Constructing rating scales for second language tests. *ELT Journal, 49*, 3-12.

Vaughan, C. (1991). Holistic assessment: What goes on in the rater's mind? In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 111-125). Norwood, New Jersey: Ablex Publishing Corporation.

Wolfe-Quintero, K., Inagaki, S., & Kim, H-Y. (1998). *Second language development in writing: Measures of fluency, accuracy and complexity.* Technical Report No. 17. Honolulu, HI: University of Hawai'i Press.

Appendix

39 Descriptors of ESL Academic Writing Skills

Descriptor	Example of the teachers' think-aloud verbal protocols	*f*	%
D01. This essay demonstrates an understanding of the topic and answers a specific question.	Again that's off topic, you've told him to decide and support it with reasons and examples, and he's gone off into changing the topic. (Ann)	67	3.91
D02. This essay is written clearly enough to be read without inferring or interpreting the meaning.	A lot of kind of summaries of what this person thinks, people think, but no clear sense of what the writer thinks. I mean in the conclusion… um… it becomes clearer. So, you have to guess as you read what the writer's real opinion is in the sense of what it might be. (Shelley)	81	4.72
D03. This essay is concise, containing few redundant ideas or linguistic expressions.	*But when it comes to this question I think it is hard to say which one is important, people should consider 'both' these 'two' things carefully and make their own choose,* redundant. (Judy)	21	1.22
D04. The beginning of the essay contains a clear thesis statement.	I don't see any sort of overriding thesis statement or no main, um, statement, outlining his or her argument as to what he's going to say, so I see that as a bit of a weakness in this introductory paragraph. (George)	34	1.98
D05. The main arguments in this essay are strong.	I think they're trying here to, um, they're trying to hedge their bets. It's not a great argument, it's a bit wish-wash. It's not a great argument, not a sophisticated argument. (Esther)	47	2.74
D06. There are sufficient supporting ideas and examples in this essay.	And just also giving more support behind your ideas because it's very minimal so it's very brief. More support behind the ideas would be important. (George)	24	1.40
D07. The supporting ideas and examples in this essay are logical and appropriate.	My biggest problem with it for me is that it's very illogical. They've also supported what they've said with some really interesting details, but it isn't logical to say that individual means that we're alone and we don't require other people. That's not what individual means. So, a basic lack of logic. (Shelley)	64	3.73
D08. The supporting ideas and examples in this essay are specific and detailed.	Again, he hasn't given specific reasons and examples. (Ann)	21	1.22
D09. The ideas are organized into paragraphs and include an introduction, a body, and a conclusion.	My feeling is they have the ideas, but they haven't been able to sort of organize and have a beginning, a middle, and an end. (Beth)	93	5.42
D10. Each paragraph is complete, with a clear topic sentence tied to its supporting sentences.	Okay, so again, hard to follow, lack of topic sentences. Very… weak introduction, um, I think my greatest problem with this one is lack of organization, I don't have topic sentence. I can't determine the supporting sentences. (Beth)	34	1.98
D11. Each paragraph presents one distinct and unified idea in a coherent way.	Again he's…, there's no cohesion. Within this paragraph, the sentences started off with the sister, not getting a job, now he's into being useful to your country, and pay your life…, referring back to killing in the first paragraph. (Ann)	28	1.63
D12. Each paragraph links well to the rest of the essay.	His…there's no links between paragraphs, they're discrete, first second third… (Ann)	18	1.05
D13. Ideas are developed or expanded throughout each paragraph.	He needs to go. I can see the gems of it, *people do not need to talk to each other*, you can see where he's going, but he just hasn't expanded it enough. (Ann)	58	3.38
D14. Ideas reflect the central focus of the essay, without digressing.	All the sudden he's talking about insuring successful business, so we're kind of losing	15	0.87

	focus. (Beth)		
D15. Transition devices are used effectively.	As I said, he starts the last *finally*, where he’s going to summarize and he’s tried to link it back to the points he’s made. (Ann)	56	3.27
D16. Syntactic variety is demonstrated in this essay.	I’m not seeing a lot of demonstration of a variety of syntax and grammar. (Beth)	6	0.35
D17. Complex sentences are used effectively.	*I have a sister. She older than me. She finish her school now.* Should make that a relative clause, and could’ve made one sentence out of those first three. (Tim)	53	3.09
D18. Normal word order is followed except in cases of special emphasis.	Some of the word order is okay, like the subject verb is in order. There’s a subject and a predicate, but not always. Not always. (Esther)	19	1.11
D19. Sentences are well-formed and complete, and are not missing necessary components.	One recurrent error seems to be the lack of a subject in phrases. *‘It’ is important*, um, *my country, ‘we’ do not have courses.* (Ann)	62	3.62
D20. Independent clauses are joined properly, using a conjunction and punctuation, with no run-on sentences or comma splices.	Run-on sentences. Um, comma splices… they lack the ability to cut things short, to get to the point. They’re stringing too many thoughts together so it’s really hard to figure out what they’re saying. (Tim)	41	2.39
D21. Major grammatical or linguistic errors impede comprehension.	Unfortunately, the grammar errors are obscuring the comprehension of that. (Tim)	42	2.45
D22. Verb tenses are used appropriately.	Um, then there’s some grammar things, sometimes a present, past tense, *when I was a child, my dream ‘is’ to be a soccer player*. (James)	84	4.90
D23. There is agreement between subject and verb.	*That is why technology exist‘s’*, subject verb agreement, but from a content perspective, I know what his opinion is and I’m assuming what’s going to follow. (Ann)	64	3.73
D24. Singular and plural nouns are used appropriately.	*… finished the bachelor degree with very good grade‘s’,* plural for *grades* (Judy)	40	2.33
D25. Prepositions are used appropriately.	Problem with preposition. It’s *‘in’ the labor market*, not *‘on’ the labor market.* Prepositions are often difficult. It makes a huge difference because as soon as you use the wrong one, you know something is quite off there. (Tim)	44	2.57
D26. Articles are used appropriately.	He stated his argument again, problem with articles, both the definite and indefinite, but he stayed in his premise, he said. Then he goes on to say *I will extend my article in three points*. Again, articles. (Ann)	52	3.03
D27. Anaphora (i.e., pronouns) reflects appropriate referents.	*Because I regret it*, we don’t know what *it* is and we don’t know of course if they regret answering the question or whether they regret that they didn’t study subjects they were interested in, so there’s no reference. (Esther)	51	2.97
D28. Conditional verb forms are used appropriately.	And again his verb sequencing there, he needs some sort of conditional, *they would, they could, they should.* (Ann)	9	0.52
D29. Sophisticated or advanced vocabulary is used.	I like the attempt at using more sophisticated vocabulary, so, *utterly, delved, dazzling*, um, *in this regard*, um, *transparency*. (Beth)	48	2.80
D30. A wide-range of vocabulary is used, with minimal repetition.	He’s not using any substitutions, he’s not saying *‘work together’* or *‘participate,’* or anything to replace *‘cooperate.’* (Ann)	16	0.93
D31. The meaning of vocabulary is understood correctly and used in the appropriate context.	*Hence it makes little sense to study extravagant subjects*, I mean it starts off good but *extravagant*, it is an inappropriate adjective. Doesn’t tell me anything? It’s totally out of place. *Extravagant*, we talk about material things. (Tim)	53	3.09

D32. The essay demonstrates facility with collocations, and does not contain unnatural word-by-word translations.	I always point them out but I don't um… word choice errors, *I did the best choice* rather than *made the best choice*, they need to understand it's like a collocation, you *make a choice* not *do a choice*, or verb phrases, I often put them as verb phrases. (Shelley)	25	1.46
D33. Words change their forms where necessary and appropriate.	*Capitalism, capitalist*, he knows his word groups. (Ann)	59	3.44
D34. Words are spelled correctly.	Um, the one word I don't understand is *dein or dyin themself.* I think they mean *deny…, that is simply because in modern society, people deny themself…,* oh, sorry, it's *define*. The *'f'* is missing. Then *'self'* should be *'selves,' to a great extend* is wrong, it should be *'extent.'* (Judy)	104	6.06
D35. Punctuation marks are used correctly.	Yeah, *with the help of the internet*, period, *we can communicate with each other easily*, period, *but before the internet*, period, *people contact with others only by letters.* So, inability to use punctuation properly which then does impact meaning. In my opinion, quite basic and leads to misunderstanding, improper use of punctuation… (Beth)	66	3.85
D36. Capital letters are used appropriately.	*the average american, american,* capitalization, *had left the of rest*. (George)	19	1.11
D37. The essay contains appropriate indentation.	I'll call this paragraph one, the first sentence. They haven't indented paragraphs so it's hard to tell where paragraphs are. (Shelley)	10	0.58
D38. The essay prompt is well-paraphrased, and is not replicated verbatim.	I think also *everywhere in this society. To be specific…,* okay, here we go, again they're repeating what's in the prompt. (Tim)	5	0.29
D39. Appropriate tone and register are used throughout the essay.	Um… too many *'I'*s, it's, um, every sentence is 'I,' 'I,' 'I,' so again tone and register, totally inappropriate. (Ann)	82	4.78
Total		1,715	100.00

Note. Transcribed text from each tape-recorded think-aloud session that was read directly from an essay is italicized.