Language Learning 47:1, March 1997, pp. 101–143

Measures of Linguistic Accuracy in Second Language Writing Research

Charlene G. Polio
Michigan State University

Because a literature review revealed that the descriptions of measures of linguistic accuracy in research on second language writing are often inadequate and their reliabilities often not reported, I completed an empirical study comparing 3 measures. The study used a holistic scale, error-free T-units, and an error classification system on the essays of English as a second language (ESL) students. I present detailed discussion of how each measure was implemented, give intra- and interrater reliabilities and discuss why disagreements arose within a rater and between raters. The study will provide others doing research in the area of L2 writing with a comprehensive description that will help them select and use a measure of linguistic accuracy.

Studies of second language (L2) learner writing (and sometimes speech) have used various measures of linguistic accuracy (which can include morphological, syntactic and lexical accuracy) to answer a variety of research questions. With perhaps one excep-

I would like to thank David Breher for his assistance rating essays and Susan Gass and Alison Mackey for their helpful comments on earlier drafts.

Correspondence concerning this article may be addressed to Charlene Polio, English Language Center, Center for International Programs, Michigan State University, East Lansing, Michigan 48824-1035, U.S.A.. Internet: polio@pilot.msu.edu

tion (Ishikawa, 1995), researchers have not discussed these measures in great detail, making replication of a study or use of a particular measure in a different context difficult. Furthermore, they have rarely reported intra- and interrater reliabilities, which can call into question the conclusions based on the measures. The purpose of this article is to examine the various measures of linguistic accuracy to provide guidance to other researchers wanting to use such a measure.

I first review various measures of linguistic accuracy that studies of L2 learner writing have used, explaining not only the context in which each measure was used, but also how the authors described each measure and whether or not they reported its reliability.

First, why should we be concerned with the construct of linguistic accuracy at all, particularly with more emphasis now being placed on other areas in L2 writing pedagogy? Even if one ignores important concepts such as coherence and content, many factors other than the number of linguistic errors determine good writing: for example, sentence complexity and variety. However, linguistic accuracy is an interesting, relevant construct for research in three (not mutually exclusive) areas: second language acquisition (SLA), L2 writing assessment, and L2 writing pedagogy.

SLA research often asks questions about learners' interlanguage under different conditions. Is a learner more accurate in some conditions than others, and if so, what causes that difference? For example, if a learner is paying more attention in one condition and produces language with fewer errors, that might inform us about some of the cognitive processes in L2 speech production. Not only are such questions important for issues of learning, but also, they help us devise methods of eliciting language for research. Similarly, those involved in language testing must elicit samples of language for evaluation. Do certain tests or testing conditions have an effect on a learner's linguistic accuracy? Crookes (1989), for example, examined English as a second language (ESL) learners' speech under 2 conditions: time for planning and no time for planning. He hypothesized that the learners' speech would be more accurate, but it was not.

Researchers studying writing have asked similar questions. Does a L2 writer's accuracy change under certain conditions? Kobayashi and Rinnert (1992), for example, examined ESL students' writing under 2 conditions: translation from their L1 and direct composition. Kroll (1990) examined ESL students' writing on timed essays and at-home essays. These studies give us information not only about how ESL students write, but also about assessment measures. If, for example, there is no difference in students' timed and untimed writing, we may want to use timed writing for assessment because it is faster. And again, even though other factors are related to good writing, linguistic accuracy is usually a concern in writing assessment.

The issue of the importance of linguistic accuracy to pedagogy is more complex. Writing pedagogy currently emphasizes the writing process and idea generation; it has placed less emphasis on getting students to write error-free sentences. However, the trend toward a more process-oriented approach in teaching writing to L2 learners simply insists that editing wait until the final drafts. Even though students are often taught to wait until the later stages to edit, editing is not necessarily less important. Indeed, research on sentence-level errors continues. Several studies have looked at different pedagogical techniques for improving linguistic accuracy. Robb, Ross, and Shortreed (1986) examined the effect of different methods of feedback on essays. More recently, Ishikawa (1995) looked at different teaching techniques and Frantzen (1995) studied the effect of supplemental grammar work.

In sum, several researchers have studied the construct of linguistic accuracy for a variety of reasons and have used different techniques to measure it.[1] The present study arose out of an attempt to find a measure of linguistic accuracy for a study on ESL students' essay revisions (Polio, Fleck & Leder, 1996). Initial coding schemes measuring both the quality and quantity of writing errors were problematic. Thus, I decided that as a priority one

should compare and examine more closely different measures of linguistic accuracy. The research questions for this study were:

1. What measures of linguistic accuracy are used in L2 writing research?
2. What are the reported reliabilities of these measures?
3. Can intra- and interrater reliability be obtained on the various measures?
4. When raters do not agree, what is the source of those disagreements?

Review of Previous Studies

The data set used to answer questions 1 and 2 consisted of studies from 7 journals[2] (from 1984 to 1995) that I expected to have studies using measures of linguistic accuracy. Among those studies that reported measuring linguistic or grammatical accuracy, I found 3 different types of measures: holistic scales, number of error-free units, and number of errors (with or without error classification). A summary of these studies appears in Table 1, which provides the following information about each study: the independent variable(s), a description of the accuracy measure, the participants' L1 and L2, their reported proficiency level, intra- and interrater reliabilities, the type of writing sample, and whether or not the study obtained significant results. I report significance because unreliable measures may cause nonsignificant results and hence nonsignificant findings; lack of reliability does not, however, invalidate significant findings.[3]

Holistic Scales

The first set of studies used a holistic scale to assess linguistic or grammatical accuracy as one component among others in a composition rating scale. Hamp-Lyons and Henning (1991) tested a composition scale designed to assess communicative writing ability across different writing tasks. They wanted to ascertain the reliability and validity of various traits. They rated essays on 7

Table 1

Studies Using Measures of Linguistic Accuracy

Study	Independent variable	Accuracy measure	Subjects			Reliability		Writing sample	Significance
			L1	L2	Level	Intrarater	Interrater		
Holistic measures									
Hamp-Lyons & Henning (1991)	correlational study of multitrait scoring instrument	linguistic accuracy as one of 7 components	varied	English	varied	none	.33–.79 between pairs of raters; averages were .61 on one sample and .91 on the other sample	Test of Written English, Michigan Writing Assessment	correlations with all subscores on all samples were significant
Hedgcock & Lefkowitz (1992)	type of feedback (instructor vs. peer)	grammar, vocabulary, mechanics as 3 of 5 components	English	French	"basic" accelerated first year university	none	.88 average among 4 raters on total composition score; none given for subscores	descriptive and persuasive essays	yes

Table 1 (continued)

Studies Using Measures of Linguistic Accuracy

Study	Independent variable	Accuracy measure	Subjects			Reliability		Writing sample	Significance
			L1	L2	Level	Intrarater	Interrater		
Tarone et al. (1993)	grade level, ESL vs. mainstream, age of arrival, years in US	Accuracy as one of 4 components	Cambodian Laotian Hmong Vietnamese	English	8th, 10th, 12 graders and univer- sity students	none	"excellent"	in-class narratives	yes, in some cases
Wesche (1987)	test development project	language use as one of 3 components for writing section	varied	English	post-secondary, high proficiency	none	high KR-20 for entire test; none given for writing section	giving and supporting opinion	significant correlations with other exams
Error-free units									
Casanave (1994)	time	percent of EFTs words per EFT	Japanese	English	intermediate (420–500 TOEFL) advanced (>500 TOEFL)	none	none	journals	not tested
Ishikawa (1995)	teaching task (guided - answering questions vs. free -picture description)	percent of EFTs percent of EFCs words per EFT words per EFC (and others)	Japanese	English	College freshman, "low proficiency"	.92 (total words in EFCs) .96 (number of EFCs on sample	none	30-minute picture- story description	yes

Robb, Ross & Shortreed (1986)	type of feedback	ratio of EFT/total T-units ratio of EFT.total clauses words in EFTs/total word (and others)	Japanese	English	university freshman	none	.87 on sample (average?)	in-class narratives	no
Number of errors without classification									
Carlisle (1989)	type of program (bilingual vs. submersion)	average number of errors per T-unit (mechanical, lexical, morphological, syntactic errors)	Spanish	English	4th and 6th graders	none	"high" on sample	five tasks, three rhetorical modes	no

Table 1 (continued)

Studies Using Measures of Linguistic Accuracy

Study	Independent variable	Accuracy measure	Subjects			Reliability		Writing sample	Significance
			L1	L2	Level	Intrarater	Interrater		
Fischer (1984)	correlational study of: communicative value, clarity of expression and level of syntactic complexity, and grammar	ratio of total number of errors in structures studied in class to total number of clauses	English	French	first year university	none	.73 for total exam (none given for error measure)	letter written for a given context	significant correlation with other subscores
Kepner (1991)	type of written feedback (message related vs. surface error corrections) verbal ability	surface- level error count (mechanical, grammatical, vocabulary, syntax)	English	Spanish	second year university	none	.97 (on sample or whole set?)	journals	no
Zhang (1987)	cognitive complexity of question/ response	number of errors per 100 words	varied (mostly Asian)	English	university undergraduate and graduate students	none	.85 on sample	answers to questions about a picture	no

Number of errors with classification									
Bardovi-Harlig & Bofman (1989)	L1, university placement exam results	ratio of syntactic, lexical-idiomatic, and morphological errors to total errors	Arabic, Chinese, Korean, Malay, Spanish	English	university TOEFL (543–567)	none	"88%"	45-minute placement exam on nontechnical topic	L1 -no; exam results - yes on lexical only
Chastain (1990)	grading	ratio of errors to total number of words (also ratio of vocabulary, morphological, syntactical error to total number of errors)	English	Spanish	3rd and 4th year university	none	none	argumentative, compare/ contrast	no
Frantzen (1995)	supplemental grammar instruction vs. none	ratio of 12 different errors to total number of obligatory contexts	English	Spanish	university 2nd year Spanish	none	none	in-class, memorable experience	no on most measures; yes on a few

Table 1 (continued)

Studies Using Measures of Linguistic Accuracy

Study	Independent variable	Accuracy measure	Subjects			Reliability		Writing sample	Significance
			L1	L2	Level	Intrarater	Interrater		
Kobayashi & Rinnert (1992)	translation vs. direct composition	number of lexical choice, awkward forms, transitional words per 100 words	Japanese	English	university English comp I and II	none	none	choice of four comparison topics completed in class	yes for higher level students on two error types; no for lower level
Kroll (1990)	in-class vs. at-home writing	Ratio of words to number of errors (33 error types)	Arabic, Chinese, Japanese, Persian, Spanish	English	advanced undergraduate ESL composition students	none	none	in-class and at-home	no for accuracy ratio; high correlation for error distribution

traits on a scale of 0 to 9 in each category. (The descriptors of the "linguistic accuracy" category appear in Appendix A.) They gave raters no formal training in using the scales. The reliability between pairs of raters varied from .70 to .79 on essays from the Test of Written English (TWE) and from .33 to .35 on essays from the Michigan Writing Assessment (MWA). When the authors averaged the correlations, using the Spearman-Brown formula, the reliability was .91 for the TWE and .61 for the MWA.

Hedgcock and Lefkowitz (1992) compared 2 different techniques for giving feedback on essays (oral feedback from peers and written feedback from the teacher). They found significant differences between the experimental and control groups with regard to accuracy. They used a writing scale adapted from the well-known scale in Jacobs, Zinkgraf, Wormuth, Hartfiel, and Hughey (1981). Three components of the scale (grammar, vocabulary, mechanics) relate to accuracy; they appear in Appendix A. Hedgcock and Lefkowitz reported interrater reliability on the entire composition score at .87 as the average of pair-wise correlations among 4 raters. They gave no reliability for any of the individual components.

Tarone et al. (1993) examined the writing of Southeast Asian students in secondary school and university. They compared students on the basis of grade level as well as age of arrival and time in the United States; they found significant differences among some of the groups on linguistic accuracy. They used a 4-component scale, of which one component was "accuracy syntax" (see Appendix A). The study used 3 raters for each essay and reported interrater reliability only as "excellent" (p. 156). The authors do not state whether this was the case for only the entire score or for the subscores as well.

Wesche (1987) reported on the construction of a new performance test for ESL students entering university in Ontario. The test had several parts, including writing. Wesche graded the writing part of the exam on 3 traits, one of which was "language use." This scale also appears in Appendix A. Wesche gave no reliability rating for the writing portion of the exam, although she reported a high reliability for the test as a whole.

In sum, the various scales include descriptors related to vocabulary, spelling, punctuation, syntax, morphology, idiom use, paragraph indentation, and word form. Some of the scales attempt to quantify the number of errors, using words such as "frequent" and "occasional." Others try to characterize the quality of the language with terms such as "significant," "meaning disrupted," "effective," and "sophisticated." Thus, the holistic scales can go beyond counting the number of errors and allow the rater to consider the severity of the errors as well.

With regard to reliability, only *one* of the studies (Hamp-Lyons and Henning, 1991) reported reliability on the linguistic accuracy subscores. They were able to obtain a reliability of .91 on one set of essays without training raters. They also pointed out that the scale used was intended for a wider range of proficiency levels and that one set of essays fell within a restricted range. Similarly, Ishikawa (1995) pointed out:

> [B]oth holistic and analytic scoring protocols are usually aimed at placement. This means they are suitable for a wide range of proficiencies, but less suitable for discrimination at a single proficiency level. (p. 56)

Because all the studies published the scales, any researcher wanting to use one of the measures or replicate the studies should not have any difficulty. Future studies, however, should report subscore reliabilities if they investigate an individual component, as opposed to general writing proficiency.

Error-free Units

The next set of studies evaluated accuracy by counting the number of error-free T-units (EFTs) and/or error-free clauses (EFCs). Such studies have used a more objective measure than those discussed above. Furthermore, error-free units are more clearly a measure of accuracy as distinct from complexity; an essay can be full of error-free T-units but contain very simple sentences. This measure does not, however, take into account the severity of the error nor the number of errors within one T-unit. A

T-unit is defined as an independent clause and its dependent clauses (Hunt, 1965). To use a measure such as EFT or EFC, one must define both the unit (clause or T-unit) and what "error-free" means. Discrepancies identifying units are probably insignificant (as will be shown later in this paper) whereas identifying an *error-free* unit is much more problematic. How these studies dealt with such a problem is addressed in the discussion below.

Robb, et al. (1986) examined the effects of 4 different kinds of feedback on EFL students' essays and found no significant difference on accuracy among the 4 groups of students receiving different kinds of feedback. They used 19 objective measures; through factor analysis they concluded that 3 of the measures, ratio of EFTs/total T-units, ratio of EFTs/total clauses, and ratio of words in EFTs/total words, measured accuracy. They did not discuss in any detail how they identified an error-free unit. With regard to reliability, they said:

> Interrater reliability estimates (Kendall's coefficient of concordance) calculated at the start of the study were sufficient at .87 for the objective scoring . . . (p. 87)

It seems that .87 was an average of the 19 objective measures (which included measures like number of words, number of clauses and others). Thus, we do not know the reliability of the actual coding of the accuracy measures, but it was probably below .87; it is undoubtedly easier to get a high reliability on measures, such as number of words or number of clauses, that do not involve judgements of error.

Casanave (1994) wanted to find measures that could document change in ESL students' journal writing over 3 semesters. With regard to accuracy, she chose to examine the ratio of EFTs and the length of EFTs. She did not report her accuracy measures separately, but combined the scores with measures of length and complexity. Some students' individual scores showed an increase and some a decrease in accuracy, but Casanave did not test significance. She gave no reliability scores; her only discussion of what constituted an error was as follows:

> I did not count spelling or typing mistakes as errors, but did count word endings, articles, prepositions, word usage, and tense. In a few cases it was difficult to determine whether the writer had made an error or not. (pp. 199–200)

Ishikawa's (1995) study investigated how 2 different types of writing practice tasks affected writing proficiency for low-proficiency EFL students. She was also concerned with finding a measure that would document change in students at this level. She found significant changes on 9 measures and also a significant change on 1 teaching task (writing out picture stories as opposed to answering questions about them). Those measures related to accuracy involved both EFCs and EFTs. Unfortunately, Ishikawa did not report interrater reliability on these measures. She did, however, report a high *intra*rater reliability on 2 measures (.92 for total words in EFCs and .96 for number of EFCs per composition[4]). Though she also acknowledged that determining correctness can be difficult, Ishikawa gave far more detail than most on how she coded her data. For example, she said specifically that she did not count punctuation except at sentence boundaries and disregarded spelling unless it involved a grammatical marker. She explained that when a student used more than one tense, she considered the most common one correct and that in cases of ambiguity, she gave students the benefit of the doubt. Most important, she stated that correctness was determined "with respect to discourse, vocabulary, grammar, and style, and strictly interpreted" (p. 59), and that she considered a sentence or clause in context; she considered its correctness not in isolation but as part of the discourse. Ishikawa went into even further detail; though one may not agree with all of her decisions, the relevant point is that anyone reading her study has a good sense of how she handled correctness.

Reviewing the above studies, we see that EFTs or EFCs are a way to get at the quantity of errors but not the quality. Defining an error may be problematic and most studies do not discuss it in great detail. Ishikawa (1995) also noted that most studies do not define the term "error-free." Furthermore, we have no idea how easy it is to obtain interrater reliability on these measures; given

that "error" is not well-defined, interrater reliability may be difficult to obtain.

Error Counts Without Classification

Four studies measured accuracy by counting the number of errors as opposed to counting the number of error-free units. Fischer (1984) discussed the development of a test of written communicative competence for learners of French. He set up a social situation that called for a written response. He then had the responses rated for Degree of Pertinence and Communicative Value, Clarity of Expression and Level of Syntactic Complexity, and Grammar. This last measure is relevant here. In the pilot study, Fischer used a holistic scale, but for reasons that are not clear, replaced it by a measure that involved counting the number of errors. The measure used was a ratio of number of errors to the number of clauses.

With regard to explicitness, Fischer defined a clause as "a syntactic unit which contains a finite verb" (1984, p. 15). Errors included both grammar and vocabulary problems. One puzzling part of the description of the measure is Fischer's statement that errors were "mistakes made in structures previously studied in class" (p. 16). Because he did not elaborate on this point, it is not clear what kinds of errors he counted. The interrater reliability of the entire test among teachers who were not formally trained in rating was .73 using Kendall's Coefficient of Concordance. Fischer gave no reliability for the Grammar portion.

Zhang (1987) examined the relationship between the cognitive complexity of questions (as prompts), and the length, syntactic complexity, and linguistic accuracy of written responses. He found no change in linguistic accuracy related to question type. Linguistic accuracy was determined "by the number of errors, whether in spelling, punctuation, semantics or grammar per 100 words" (p. 473). About half of the written responses were coded by 2 raters and the Pearson correlation was .85 for the accuracy measure.

Carlisle (1989) studied elementary school students in 2 types of programs, bilingual and submersion. To compare the writing of students in these programs, Carlisle collected samples of writing on 5 different tasks. Carlisle measured 5 dependant variables for each essay: rhetorical effectiveness, overall quality, productivity, syntactic maturity, and error frequency and found all differed significantly between the students in the 2 programs. The error-frequency measure Carlisle defined as the average number of errors per T-unit, elaborating as follows:

> In the current study, error was defined as any deviation from the written standard, Edited American English. Six types of errors were scored: mechanical errors (punctuation and capitalization), spelling errors, word choice errors, agreement errors, syntactic errors, and tense shifts across T-unit boundaries. (p. 264)

Reliability on the subjective measures was high, particularly after essays on which there were disagreements went to a third rater. For the objective measures (productivity: total number of words; syntactic maturity: average number of words per T-unit; error frequency: average number of errors per T-unit) Carlisle provided the following discussion of reliability:

> After the original researcher had identified and coded these measures in the 434 essays written in English, a second researcher, who had become completely familiar with the coding procedures, went over a sample of 62 essays, the entire group of "Kangaroo" papers, to check for any possible mistakes on the part of the original researcher in identifying and coding T-units, mechanical errors, spelling errors, word choice errors, agreement errors, syntactic errors, and switches in tense across T-unit boundaries. For all measures, the agreement between the two researchers was exceptionally high, even on switches in tense across T-units, a measure for which no strict guidelines were available. Because the method used to check the reliability of identifying and coding the objective measures in this study was less than ideal, no attempt was made to calculate reliability coefficients between the coders. From the information given

> above, the coefficients would have been very high, and probably artificially so. (p. 267)

It seems that the second rater simply checked the first rater's coding; that is, the coding was not done blindly. It is not clear if the "less-than ideal" measure to check reliability refers to this procedure or to the method of calculation.

Kepner (1991) studied second-year university Spanish students. Types of feedback on journals (message-related and surface-error correction) as well as verbal ability were the independent variables. Kepner examined students' journals for higher-level propositions and surface-level errors. Students receiving message-related feedback had significantly more higher-level propositions, but there was no difference between the groups in terms of surface-level errors. The errors included "all incidences of sentence-level mechanical errors of grammar, vocabulary and syntax" (p. 308). An interrater reliability of .97 was obtained for the error-count measure.

Counting the number of errors gets at the quantity of errors better than a measure, such as EFT, that does not distinguish between 1 and more than 1 error per T-unit. In cases of homogeneous populations, a more fine-grained measure of accuracy such as an error-count may be a better option. The studies above did not discuss problems in disagreement regarding error identification, nor did they say how they handled ambiguous cases of an error that could be counted as 1 or more errors.[5] Two of the 4 studies reported interrater reliability on this measure, achieving .85 and .97.

Error Count With Classification

The remaining studies tallied not only individual errors, as in the 4 studies above, but also classified the errors. Bardovi-Harlig and Bofman (1989) examined differences in syntactic complexity, and error distribution and type, between ESL students who had passed a university placement exam and those who had not. They also compared 6 native language groups. To determine accuracy, they classified each error into one of 3 superordinate categories

(syntactic, morphological, and lexical-idiomatic) and then classified it further within the superordinate category. They found a significant difference in errors per clause between the pass and nonpass groups for lexical errors but not for syntactic or morphological errors.[6] They found no significant difference in number of errors across language groups, and the distribution of the 3 error types seemed to be the same for the pass/no-pass groups.

Bardovi-Harlig and Bofman (1989) described in more detail than other studies how they identified errors, giving examples and explaining that they had not counted spelling and punctuation. Regarding reliability they said, "errors were identified by the authors with an interrater reliability of 88%" (p. 21). What they meant by this is not clear. It could mean that once an error was identified, they agreed on its classification 88% of the time. But probably there were cases that both authors did not agree were errors. In fact, they coded only those errors that both agreed to be errors and they agreed on a classification of 88% of those errors. (Bardovi-Harlig, personal communication, June, 1996)

Chastain (1990) compared 2 essays written by U.S. university students studying Spanish. The teacher graded 1 of the essays but not the other. Chastain compared the essays for accuracy using 3 measures: ratio of errors to total number of words, ratio of vocabulary errors to total number of words, and ratio of morphological errors to total number or words. There were no significant differences on these 3 measures.

Frantzen (1995) examined the effects of supplemental grammar instruction on grammatical accuracy in the compositions of U.S. university Spanish students. To measure grammatical accuracy, Frantzen used 12 categories and scored essays for the correct use of a particular structure divided by the total number of obligatory contexts for that structure. To examine the difference between the 2 groups, Frantzen compared 20 scores including the original 12 categories, 2 composite scores, and 2 categories subdivided, from the pre-to posttest. There was a significant difference from pre-to posttest on 4 of the 20 measures and a significant difference between the 2 groups on 2 of the 20 measures.

Frantzen's study differs from the others mentioned here in that she determined an accuracy score not by dividing the number of errors by the number of words or T-units, but by the number of obligatory contexts. Thus, she was coding correct uses of each of the structures examined as well. She divided the number of correct uses by the sum of the correct uses plus the number of errors. She stated that most of the errors were coded except for those few that were "infrequent and difficult to categorize" (p. 333).

Kobayashi and Rinnert (1992) studied differences in essays written by Japanese EFL students in their L1 and translated into their L2, and essays written directly in the L2. To compare the 2 kinds of writing with regard to accuracy, the authors counted 3 kinds of errors "likely to interfere with the communication of a writer's intended meaning" (p. 190). These included errors of lexical choice, awkward form, and transitional problems. They gave examples of each type of error. The lexical and transitional errors are fairly straightforward. "Awkward form" seems a little more difficult to operationalize but consisted of:

> grammatically and/or semantically deviant phrases or sentences that interfered with naturalness of a writer's expression and/or obscured the writer's intended meaning. (p. 191)

The researchers counted all the errors and resolved differences by discussion. Regarding reliability, they stated:

> Because the overall frequency count tallied quite well, an interrater reliability check was not conducted on these more objective measures. (p. 191)

They found significant differences between the direct compositions and the translations for the high-proficiency group on awkward phrases and transitional problems.

Kroll (1990) examined differences between students' writing in class under time constraints and writing done at home (i.e., without time constraints). Kroll coded 33 different error types, giving the following information on error coding:

> In closely examining each sentence in the corpus of essays, the criterion for deciding whether or not an error had been committed and, if so, what type of error, was to determine what "syntactic reconstruction" could most easily and economically render the sentence into acceptable English given the context. For example, a singular subject with a plural verb was labeled a "subject-verb agreement" violation, while a correctly formed past tense had to be labeled "incorrect tense" if the context show a present-tense orientation. (p. 143)

Kroll gave accuracy scores on the basis of total words/total number of errors, finding no significant differences in terms of error ratios. There was, however, a high correlation between in-class and at-home essays with regard to distribution of errors. Kroll gave no further information on coding or interrater reliability of the error coding scheme.

The studies in this group went a step further by classifying the type of error a learner makes and not simply the number. This is obviously potentially useful information. But again, the studies gave only a few guidelines for how to determine an error or how to deal with cases that could be considered more than one kind of error. With the exception of Bardovi-Harlig and Bofman (1989), none reported any reliability scores.

Examining the 16 studies above provided a starting point for considering different measures of linguistic accuracy. Many questions, however, remained. Furthermore, one cannot be certain about which measures resulted in reliable scores. Thus, I conducted this study to examine 3 of these measures more closely; that is, to determine what problems one encounters in their implementation and how high an interrater reliability one could achieve.

It is not my intention to determine the most appropriate measure for all populations on which one may do writing research, but rather to describe the problems involved in implementing and obtaining reliability on the various measures.

Method

Participants. To test the 3 accuracy measures, I used 38 one-hour essays. The participants were 38 undergraduate and graduate university (about 50% of each) ESL students, most of whom were already taking other university courses. Their English proficiency was deemed high enough by the university to take other academic courses but they were deficient on the writing portion of a university placement exam.

Procedure. To test the 3 accuracy measures, I used a one-hour essay written by each student. I used the same 38 essays for each measure. I used the most general method, a holistic scale, first, followed by EFT identification, and then by the most specific measure, error classification. Each essay was rated twice by myself (the author) and once by a graduate-student assistant. Below is a description of each method and the reliability results.

Holistic scale. I developed the holistic scale in an attempt to find a quick and reliable method of measuring accuracy without having to count and identify errors. It appears in Appendix B. I adapted it from one currently used to place students into ESL courses. I modified the original so that it omitted references to complexity, because we were concerned only with accuracy. The scale describes the use of syntax, morphology, vocabulary, word form, and punctuation. The reason for using this scale, as opposed to one of the scales from the other studies, is that we were already familiar with a version of it; it was not our impression that any of the other scales were inherently better (or worse) than ours. This scale represents a second attempt; the original resulted in interrater reliability so low as to be not even significant. I revised the scale and did more norming with my assistant.

Error-free units. For this measure, each rater tabulated the number of T-units, the number of clauses, the number of words, and the number of error-free T-units. After we had coded several practice essays, problems regarding each of these counts arose. Most of the problems or disagreements at this stage related to structures not addressed in any of the studies discussed above.

For example, several sentences were grammatical in British English, but not American. There were also errors of prescriptive English that native speakers could have made. As a result of the preliminary coding, we developed some guidelines (Appendix C). Included in them are rules for determining T-units, clauses, and words. Problems such as how to deal with sentence fragments and tag-questions are included. After the initial ratings, we compared the counts for any that were far apart. These we double-checked and changed only if the difference was due to a *counting* error; that is, if one rater, for example, had marked 15 T-units as error-free but recorded 5. Similarly, if a word count was off by more than 20, we rechecked it. We made no changes based on *judgements* of unit or error. The three measures calculated were: EFT/TT, EFT/TC, and EFT/TW (following Robb et al., 1986).

Error count and classification. We classified errors using a system modified from Kroll (1990) (Appendix D). I made several changes to Kroll's system, adding categories that she did not include and deleting categories that seemed to be covered by other errors. For example, I added other categories such as wrong case, wrong comparative form, and genitive. Another category, "awkward phrasing," I included under lexical/phrasal choice. I included other guidelines such as: "Don't double penalize for subject-verb agreement errors when the number of the noun is wrong." Thus, a sentence such as "Visitor are pleased with the sight," counted as only a number error and not a subject-verb agreement error too. If the sentence had been "Visitor is pleased . . ." it still would have been counted as only 1 error. Kroll stated that if more than 1 error was possible, she counted the error that was least different from a correct usage. Another guideline I added was: "if there is more that one change to be made of the same magnitude, the *first* error should be counted." We classified each error and tabulated a count of error/number of words.

Results and Discussion

Holistic Scale

An initial scoring of all 38 essays resulted in the intra-and interrater reliabilities in Table 2.[7] The low reliabilities are fairly respectable considering the homogeneous population, particularly in comparison to Hamp-Lyons and Henning's (1991) study, the *only* study to report reliability on the linguistic accuracy component. Their pairwise correlations were between .33 and .35 for *the set of essays falling within a restricted range of proficiency*. (They were able to obtain a higher interrater reliability of .61 by using more than two raters.) The time taken to rate essays was far quicker than for the other two methods. The problem was, however, that the reliability was too low and the raters felt that the scale could not be modified to make it any more reliable; the scale could not be constructed so as to distinguish differences in linguistic accuracy among a group of homogeneous students, (i.e., students placed into the same ESL class). This does not mean that it is impossible to construct a holistic scale aimed at a homogeneous group of students, we simply felt that we did not have the ability to do it.

Table 2

Reliabilities of Holistic Scoring

Rater/Time	Rater 1/Time 1	Rater 1/Time 2	Rater 2
Rater 1/Time 1	—	.77	.44
Rater 1/Time 2		—	.53
Rater 2			—

Error-free Units

The reliability of these measures (Table 3) was better, with intrarater reliabilities above .90 and interrater reliabilities at .80 or higher (on 2 of the 3 measures). To achieve these reliabilities, I had to write the guidelines (Appendix C). As seen in the survey of

previous research, other studies did not report problems in identifying the error-free units. Furthermore, even with detailed guidelines, disagreements arose.

Table 3

Reliabilities of Error-free T-unit Measures

Error-free T-units / Total T-units

Rater/Time	Rater 1/Time 1	Rater 1/Time 2	Rater 2
Rater 1/Time 1	—	.91	.80
Rater 1/Time 2		—	.80
Rater 2			—

Error-free T-units / Total clauses

Rater/Time	Rater 1/Time 1	Rater 1/Time 2	Rater 2
Rater 1/Time 1	—	.93	.80
Rater 1/Time 2		—	.85
Rater 2			—

Error-free T-units / Total words

Rater/Time	Rater 1/Time 1	Rater 1/Time 2	Rater 2
Rater 1/Time 1	—	.93	.76
Rater 1/Time 2		—	.78
Rater 2			—

Determining a T-unit was generally not a problem; the intra- and interrater reliabilities for number of T-units were .99 and higher. The guidelines may have helped achieve such a high agreement on T-unit identification. Nevertheless, some disagreements, did occur as in the example below:

(1) "All in all, I would like to say that my previous home was the place where I spent my childhood in / and it is now in the middle of many new houses shining like something precious."

One rater divided the sentence at the break indicated and the other rater counted it as 1 T-unit because it is unclear whether the last clause is a second dependent clause or a new independent clause.

A far greater problem was determining what counted as an error. To examine these problems more closely, I recorded each case of disagreement both within and between raters. I classified these cases with regard to the type of error (e.g., lexical, tense/aspect, punctuation) and the reason for the disagreement, based on a discussion between us two raters. I determined 20 possible categories of errors that caused the disagreements. These, with examples, appear in Appendix E. In addition to the various grammatical structures, I included the category "unknown" for cases where a rater did not remember why a T-unit was not marked as error-free. Also included was the category "T-unit." This was for cases of disagreement caused by 2 different T-unit divisions

There were 5 possible reasons for disagreement. They were: legibility; questionable prescriptive rule; questionable native-like usage; intended meaning not clear; and a mistake on the part of the rater.

Legibility.

(2) "We small kids always *w(a)nt* to swim in the sea."

One could have read "went" as "want" because of the way the vowel was written. In the context of the essay, if "want" was intended, the writer should have used the verb in the past tense. One rater thought the writer intended "want," the other "went."

Questionable prescriptive rule.

(3) "It's weird to my friends, even to *myself* too."

Despite a trend in spoken English to use the reflexive in place of the object form of a pronoun, this sentence is prescriptively incorrect. One rater did not notice that it was incorrect.

Questionable native-like usage.

(4) "Like in many other countries, the *happy* 20's didn't bring anything happy with them."

(5) "He was always busy *at* his work."

The two raters disagreed over whether sentences (4) and (5) would be written by a native speaker.

Intended meaning not clear.

(6) "Finally, I will talk about my *sister*."

The sentence above appeared in an essay that referred at times to one sister and at times to more than one. It was not clear whether the writer meant he had one or more than one sister. One rater counted it as error-free, thinking the writer had only one sister; the other rater counted it as containing an error of number.

Table 4

Differences within Rater on Error-free Units

	Reason for error					
	Legibility	Prescript.	NL usage	Meaning unclear	Error	Total
Structure						
Lex	0	0	12	2	0	14 (.18)
Art	0	0	7	0	1	8 (.10)
T/A	0	0	4	2	3	9 (.11)
Prep	1	1	3	0	1	6 (.08)
Un	0	0	0	0	8	8 (.10)
Ref	0	0	4	0	0	4 (.05)
Pun	2	0	1	0	7	10 (.13)
Num	0	0	1	0	5	6 (.08)
WF	0	0	0	0	0	0
SV	0	0	2	0	2	4 (.05)
Det	0	0	0	0	0	0
Mod	0	0	0	1	0	0
MW	0	1	0	0	0	0
Refl	0	1	0	0	0	0
Coor	0	0	0	0	0	0
Parr	0	0	0	0	1	1 (.01)
VF	0	0	0	0	1	2 (.03)
Case	0	0	1	0	0	1 (.01)
WO	0	0	0	0	0	0 (.00)
T-uni 4 (both type and reason)						4 (.05)
Total 4	4	3	35	4	29	79
	(.05)	(.05)	(.04)	(.44)	(.05)	(.37)

See Appendix D for examples of each category.
un=Unknown

Mistake.

(7) "Why *this day is* so important for us?"

This category includes sentences that were clearly ungrammatical or grammatical but which the raters, upon looking again, conceded that they had coded wrong. This sentence is clearly ungrammatical, yet one of the raters failed to notice the error.

Tables 4 and 5 classify the disagreements within and between the raters. In both cases, the greatest cause of disagree-

Table 5

Differences between Two Raters on Error-free Units

	Reason for error						
	Legibility	Prescript.	NL usage	Meaning unclear	Error	Total	
Structure							
Lex	2	0	29	2	0	33	(.19)
Art	0	0	19	2	3	24	(.14)
T/A	0	0	17	3	0	20	(.12)
Prep	1	1	10	1	0	13	(.08)
Un	0	0	0	0	12	12	(.07)
Ref	0	2	8	1	0	11	(.06)
Pun	4	0	1	0	5	10	(.06)
Num	0	0	6	0	3	9	(.05)
WF	0	0	5	0	1	6	(.04)
SV	0	0	1	0	2	3	(.02)
Det	0	0	3	0	0	3	(.02)
Mod	0	0	2	1	0	3	(.02)
MW	0	0	0	0	2	2	(.01)
Refl	0	1	1	0	0	2	(.01)
Coor	0	0	2	0	0	2	(.01)
Parr	0	0	0	0	1	1	(.01)
VF	1	0	0	0	0	1	(.01)
Case	0	0	1	0	0	1	(.01)
WO	0	0	0	0	0	1	(.01)
T-uni 13 (both type and reason)						13	(.08)
Total 13	8	4	106	10	30	170	
(.08)	(.05)	(.02)	(.62)	(.06)	(.18)		

See Appendix D for examples of each category.
Un=Unknown

ment was the nativeness of a given sentence, more so for the disagreements between raters. The second greatest cause of disagreement was a coding mistake. Disagreements were rarely caused by confusion about the writers' intended meaning, probably because of their relatively high proficiency level.

Another issue worth mentioning is the extent to which the raters agreed on the quality of *individual* T-units. That is, raters could have a high correlation of EFT/TT but still not agree on the same T-units as being error-free. Thus, I made an additional calculation, counting the number of T-units that the raters did not agree on as error-free. This, divided by the total number of T-units (an average of each raters' count), gave a median rate of agreement at time 1 and time 2 of .94 for rater 1. The agreement between the two raters was .86.

Table 6

Reliabilities of Errors / Words

Rater/Time	Rater 1/Time 1	Rater 1/Time 2	Rater 2
Rater 1/Time 1	—	.89	.94
Rater 1/Time 2		—	.89
Rater 2			—

Error Count and Classification

The reliabilities for error counts appear in Table 6. The correlation for the number of errors was quite high. But, just as with EFT, it is important to look beyond the correlation. The percent of errors determined by *both* raters to be errors shows something different. We calculated this rate by counting the number of errors coded by only 1 rater (and not both) and dividing by the total number of errors (an average of each rater's count). The median rate of agreement between time 1 and 2 for rater 1 was .84; between raters 1 and 2 it was .81. Another question was what the agreement rate on error *type* was. Taking all the errors agreed by both raters to be errors and checking the rate of agreement on error type, the median for one rater was .79 and between raters was .74.

In addition to the problems of raters not agreeing on occurrence of errors listed above for the EFT measure, in many instances the raters did not agree on classification of errors. Below are some examples.

(8) "I wish I can see them soon."

This was coded as both a lexical error in that "wish" should be "hope" and a modal error in that "can" should be "could." One rater was following the first error rule and the other the minimal change rule.

(9) "I redecorated the whole apartment with blue tone color."

One rater called this a preposition error and two extraneous words, "tone" and "color." The other rater coded it as a lexical/phrasal misuse.

(10) "In this day evening, if the weather is fine, almost every family will go outdoors to the parks."

One rater assumed the target was "On this evening," coding it as a preposition and extraneous word error. The other rater assumed the target was "In the evening," coding it as a deixis and extraneous word error. Again, the first error and the minimal change rules seemed to conflict.

Implications

After reviewing the published studies and comparing measures of linguistic accuracy on a set of essays, I would like to make the following general conclusions. First, except for studies that used holistic scoring, the surveyed studies provide too little information for other researchers to use the measures or to replicate the studies. This does not mean that the studies were poorly done or that the results are unreliable. However, providing more information helps other researchers anticipate problems when using similar methods. Also, the *Publication Manual of the American Psychological Society* (APA, 1994) requires that authors provide enough information on their methods for others to replicate their studies. If one tried to replicate some of the studies discussed

above, one might interpret the measures differently and achieve different results. Researchers should provide more detailed information on measures of linguistic accuracy, not only for replication, but also to prevent other researchers from having to reinvent the wheel.[8]

Second, studies should more consistently report interrater reliability, even if only on a portion of the data. When nonsignificant results are obtained, we do not know whether such results are real or an artifact of an unreliable measure.

With regard to specific measures, holistic measures may not be suitable for homogeneous populations, unless one can come up with a better measure than that used in the present study. Both EFTs and error counts were more reliable measures for the range of proficiency examined here. The error classification scheme, however, resulted in agreement rates of below .80. Only one of the studies reviewed that used an error classification system (Bardovi-Harlig & Bofman, 1989) reported agreement rates, so we do not know whether one can obtain a high agreement rate on other coding schemes.[9]

Whether one decides to use EFTs or error-counts, one must consider that the discrepancies described here arose most often because of raters' disagreement on nativelike usage. By using 2 raters, one can average the results; thus, a T-unit marked error-free by one rater and not by another will be scored somewhere between error-free and correct. This is probably valid, because we considered many of the T-units in this category borderline correct usage. Thus, having 2 raters allows these T-units to, in effect, be given a score that is halfway between ungrammatical and correct. Furthermore, with 2 raters, errors missed by 1 rater will be counted at least once and not missed completely.

This paper has reviewed the various measures of linguistic accuracy. These measures need to be considered individually for other populations. One measure may show change for one population but not another. Furthermore, not only the population will affect the reliability, but also the length of the writing samples. Most likely, the longer the piece of writing, to a certain extent, the

more reliable the measure will be. I did not attempt to validate any of the measures nor to determine whether they are measuring the same construct.[10] I did provide a detailed description of what is involved in using the measures and what issues other researchers need to consider when choosing a measure of linguistic accuracy.

Revised version accepted 10 October 1996

Notes

[1]All techniques claiming to measure linguistic accuracy are not necessarily measuring exactly the same thing. This paper is no way an attempt to validate any of the measures nor to determine if they are measuring the same construct. Validity is an important issue but beyond this paper's scope. Furthermore, I am not attempting to define linguistic accuracy, but rather to present how other researchers have suggested it be measured.

[2]The journals were *Applied Linguistics*, *Journal of Second Language Writing*, *Language Learning*, *Modern Language Journal*, *Studies in Second Language Acquisition*, *Language Testing*, and *TESOL Quarterly*. I included Kroll's (1990) study in the list as well. Database searches were not helpful in finding relevant studies. A keyword such as "accuracy" did not lead to the studies reviewed here.

[3]Schils, van der Poel, and Weltens (1991) elaborated this point. They discuss the issue of test reliability in applied linguistics research. One can extend their conclusions to interrater reliability.

[4]Note that the measure is the *number* of EFCs and not the ratio of EFCs/total clauses. Thus, the measure is also affected by the length of the essay.

[5]Consider a sentence such as "Every day my parent tells me that I should study hard." This could be counted as only a number error or as both a number and subject-verb agreement error.

[6]The probability level of significance for morphological errors was $p<.066$. Bardovi-Harlig and Bofman (1989) said that this is "weakly significant" (p. 24).

[7]All the correlations reported in this study are Spearman-Brown rank order. One could argue that some of the measures do represent an interval scale; however, I used Spearman-Brown because the data for each measure were not normally distributed and because what is really of interest is how the essays compared to one another. I also calculated Pearson correlations on the 3 measures, obtaining differences of +/- .02.

[8]One reviewer suggested that the replication standard may be unrealistic; journals simply do not have enough space for researchers to adhere to it. This is a point worthy of further discussion and is elaborated by Polio and Gass (1996). The reviewer stated that if one wants more information, one should simply contact the researchers. Thus, one solution is for authors to say in a footnote that one may obtain guidelines for the measure from the authors.

[9]Cumming and Mellow (1996) examined the extent to which accuracy on certain grammatical morphemes in written English can predict general L2 proficiency. They too noted problems in coding for accuracy and discussed the fact that most of the studies they reviewed on article use did not report intercoder reliability. By limiting the types of errors coded to lexical articles, the plural -s on regular nouns, and the third person singular -s on verbs, they were able to obtain a high intercoder reliability. The studies reviewed here attempted to code all or a much wider range of errors.

[10]Polio, et al. (1996) used the ratio of EFT measure for a study on students of the same proficiency level as in this study. The measure did indicate significant progress over a 15-week semester. Ishikawa (1995) showed change in low-proficiency students using words in error-free clauses and number of error-free clauses. Because the measures showed a difference before and after instruction, they are in a sense validated by a group differences approach.

References

American Psychological Association. (1994). *Publication manual of the American Psychological Association* (4th ed.). Washington, DC: American Psychological Association.

Bardovi-Harlig, K., & Bofman, T. (1989). Attainment of syntactic and morphological accuracy by advanced language learners. *Studies in Second Language Acquisition, 11,* 17–34.

Casanave, C. (1994). Language development in students' journals. *Journal of Second Language Writing, 3,* 179–201.

Carlisle, R. (1989). The writing of anglo and Hispanic elementary school students in bilingual, submersion, and regular programs. *Studies in Second Language Acquisition, 11,* 257–280.

Chastain, K. (1990). Characteristics of graded and ungraded compositions. *Modern Language Journal, 74,* 10–14.

Crookes, G. (1989). Planning and interlanguage variation. *Studies in Second Language Acquisition, 11,* 367–383.

Cumming, A., & Mellow, D. (1996). An investigation into the validity of written indicators of second language proficiency. In A. Cumming and R. Berwick (Eds.), *Validation in Language Testing* (pp. 72–93). Clevedon: Multilingual Matters.

Fischer, R. (1984). Testing written communicative competence in French. *Modern Language Journal, 68,* 13–20.

Frantzen, D. (1995). The effects of grammar supplementation on written accuracy in an intermediate Spanish content course. *Modern Language Journal, 79,* 329–344.

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. (1992). Collaborative oral/aural revision in foreign language writing instruction. *Journal of Second Language Writing, 3,* 255–276.

Hunt, K. (1965). *Grammatical structures written at three grade levels*. Champaign, IL: National Council of Teachers of English.

Ishikawa, S. (1995). Objective measurement of low-proficiency EFL narrative writing. *Journal of Second Language Writing, 4,* 51–70.

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

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

Kobayashi, H. & Rinnert, C. (1992). Effects of first language on second language writing: Translation versus direct composition. *Language Learning, 42,* 183–215.

Kroll, B. (1990). What does time buy? ESL student performance on home versus class compositions. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 140-154). Cambridge: Cambridge University Press.

Polio, C., Fleck, C., & Leder, N. (1996, May). *Second language essay revision and linguistic accuracy*. Paper presented at the meeting for the Canadian Association for Applied Linguistics, London, Ontario.

Polio, C., & Gass, S. (1996). Replication and reporting: A commentary. Submitted to *Studies in Second Language Acquisition*.

Robb, S., Ross, T., & Shortreed, I. (1986). Salience of feedback on error and its effect on EFL writing quality. *TESOL Quarterly, 20,* 83–96.

Schils, E.D.J., van der Poel, M.G.M., & Weltens, B. (1991). The reliability ritual. *Language Testing, 41,* 125-138.

Tarone, E., Downing, B., Cohen, A., Gillette, S., Murie, R., & Dailey, B. (1993). The writing of Southeast Asian-American students in secondary school and university. *Journal of Second Language Writing, 2,* 149–172.

Wesche, M. (1987). Second language performance testing: The Ontario test of ESL. *Language Testing, 37,* 28–47.

Zhang, S. (1987). Cognitive complexity and written production in English as a second language. *Language Learning, 37,* 469-481.

Appendix A

Holistic Measures of Linguistic Accuracy

Hamp-Lyons & Henning (1991)

Linguistic accuracy.

9: The reader sees no errors of vocabulary, spelling, punctuation, or grammar.

8: The reader sees no significant errors of vocabulary, punctuation, or grammar.

7: The reader is aware of but not troubled by occasional errors of vocabulary, spelling, punctuation, or grammar.

6: The reader is aware of errors of vocabulary, spelling, or grammar - but only occasionally.

5: The reader is aware of errors of vocabulary, spelling, punctuation, or grammar that intrude frequently.

4: The reader find the control of vocabulary, spelling, punctuation, and grammar inadequate.

3: The reader is aware of primarily gross inadequacies of vocabulary, spelling, punctuation, and grammar.

2: The reader sees no evidence of control of vocabulary, spelling, punctuation, or grammar.

1: A true nonwriter who has not produced any assessable strings of English writing. An answer that is wholly or almost wholly copied from the input text or task is in this category.

0: This rating should be used only when a candidate did not attend or attempt this part of the test in any way.

Note: From "Communicative writing profiles: An investigation of the transferability of a multiple-trait scoring instrument across ESL writing assessment contexts," by L. Hamp-Lyons and G. Henning, 1991, *Language Learning, 41*, p. 370–373.

Hedgcock & Lefkowitz (1992)

Grammar.

22-25: Excellent to very good: accurate use of relatively complex structures; few errors in agreement, number, tense, word order, articles, pronouns, prepositions.

18-21: Good to average: simple constructions used effectively; some problems in use of complex constructions; errors in agreement, number, tense, word order, articles, pronouns, prepositions.

11-17: Fair to poor: significant defects in use of complex constructions; frequent errors in agreement, number, tense, negation, word order, articles, pronouns, prepositions; fragments and deletions; lack of accuracy interferes with meaning.

5-10: Very poor: no mastery of simple sentence construction; text dominated by errors; does not communicate, or not enough to rate.

Vocabulary.

18-20: Excellent to very good: complex range; accurate word/idiom choice; mastery of word forms; appropriate register.

14-17: Good to average: adequate range; errors of word/idiom choice; effective transmission of meaning.

10-13: Fair to poor: limited range; frequent word/idiom errors; inappropriate choice, usage; meaning not effectively communicated.

7-9: Very poor: translation-based errors; little knowledge of target language vocabulary, or not enough to rate.

Mechanics.

5: Excellent to very good: masters conventions of spelling, punctuation, capitalization, paragraph indentation, etc.

4: Good to average: occasional errors in spelling, punctuation, capitalization, paragraph indentation, etc., which do not interfere with meaning.

3: Fair to poor: frequent spelling, punctuation, capitalization, paragraphing errors; meaning disrupted by formal problems.

2: Very poor: no mastery of conventions due to frequency of mechanical errors, or not enough to rate.

Note: From "Collaborative oral/aural revision in Foreign language writing instruction," by J. Hedgcock and N. Lefkowitz, 1992, *Journal of Second Language Writing, 3*, p. 275–276. Adapted from *Composición, Proceso y Síntesis* by G. Valdes and T. Dvorak, 1989, New York: McGraw-Hill. Copyright 1989 by McGraw-Hill, Inc. Reprinted with permission. (The original version appears in Jacobs, Wormuth, Hartfiel, and Hughey, 1981.)

Tarone et al. (1993)

Accuracy syntax: morphology, word form, sample.

6: Essentially no errors in a pretty complete range.

5: Wide range correctly used for the most part.

4: Some variety but still limited. Generally correct.

3: Some word form problems. Some breakdowns in verbs. Probably limited.

2: Real gaps in syntax. Mixed up structures.

1: Hit or miss. Creates serious difficulties in comprehension. Very limited and wrong.

Note: From "The writing of Southeast Asian-American students in secondary school and university," by E. Tarone, B. Downing, A. Cohen, S. Gillette, R. Murie, and B. Dailey, 1993, *Journal of Second Language Writing, 2*, p. 170, Copyright 1993 by *Journal of Second Language Writing*. Reprinted with permission.

Wesche (1987)

7 Excellent: Appropriate and concise English. Mastery of mechanics. Sophisticated range of vocabulary.

6 Very good: Effective complex constructions. Few errors of grammar, punctuation. Effective word/idiom choice and use.

5 Good: Minor problems in complex sentences. Generally accurate and appropriate language, mechanics and style.

4 Average: Quite a few errors of use, mechanics and vocabulary but meaning is not confused or obscured.

3 Fair: More frequent errors. Meaning occasionally confused or obscured, OR not an essay type format.

2 Poor: Frequent errors. Meaning often confused or obscured.

1 Very poor: Dominated by errors OR not enough to evaluate OR copied.

Note: From "Second language performance testing: The Ontario test of ESL," by M. Wesche, 1987, *Language Testing, 37*, p. 43. Copyright 1987 by *Language Testing*. Reprinted with permission.

Appendix B

Holistic Scale

10–12

mastery of word forms
virtually no errors in lexical choice
virtually no global errors
may be a few minor grammatical errors per page
demonstrates mastery of punctuation conventions

7–11

occasional errors of word form
occasional errors in lexical choice but meaning not obscured
mastery of simple constructions
rare problems in complex constructions
several local errors per page but meaning is seldom obscured
may be a few global errors per page
a few punctuation errors

4–6

frequent errors of word form and choice
meaning may be confused or obscured
some problems complex constructions
frequent global and local errors
meaning is obscured but not unintelligible
some errors in punctuation

1–3

little knowledge of English vocabulary and word forms
virtually no mastery of sentence construction rules
dominated by errors

does not communicate
frequent errors in punctuation

comments: Do not count spelling or capitalization errors. Be conservative about comma errors.

Appendix C

Guidelines for T-units, Clauses, Word Counts, and Errors

T-Units

a. A T-unit is defined an independent clause and all its dependent clauses.
b. Count run-on sentences and comma splices as two T-units with an error in the first T-unit.
ex: My school was in Saudi Arabia, it was the best school there.

T	/	T
1 error		error-free

If several comma-splices occur in a row, count only the last as error free.
c. For sentence fragments, if the verb or copula is missing, count the sentence as 1 T-unit with an error. If an NP is standing alone, attach it to the preceding or following T-unit as appropriate and count as an error. If a subordinate clause is standing alone, attach it to the preceding or following S and count it as 1 T-unit with an error.
d. When there is a grammatical subject deletion in a coordinate clause, count the entire sentence as 1 T-unit.
ex: First we went to our school and then went out with our friends.
e. Count both "so" and "but" as coordinating conjunctions. Count "so that" as a subordinating conjunction unless "so" is obviously meant.
f. Do not count tag-questions as separate T-units.
g. Count S-nodes with a deleted complementizer as a subordinate clause as in: I believe that A and (that) B = 1 T-unit.

h. But, direct quotes should be counted as:
John said, "A and B."
1 T-unit 1 T-unit
i. Assess the following type of structures on a case-by-case basis:
If A, then B and C.
As a result, A or B.
j. Count T-units in parentheses as individual T-units.

Clauses

a. A clause equals an overt subject and a finite verb. The following are only one clause each:
He left the house and drove away.
He wanted John to leave the house.
b. Only an imperative does not require a subject to be considered a clause.
c. In a sentence that has a subject with only an auxiliary verb, do not count that subject and verb as a separate clause (or as a separate T-unit. (e.g. John likes to ski and Mary does too; John likes to ski, doesn't he?; John is happy and Mary is too)

Error Guidelines

a. Do not count spelling errors (including word changes like "there/their").
b. Be conservative about counting comma errors; don't count missing commas between clauses or after prepositional phrases. Comma errors related to restrictive/non-restrictive relative clauses *should* be counted. Extraneous commas should also be considered errors.
c. Base tense/reference errors on preceding discourse; do not look at the sentence in isolation.
d. Don't count British usages as errors, (e.g. "in hospital," "at university," collective nouns as plural).
e. Be lenient about article errors from translations of proper nouns.

f. Don't count errors in capitalization.
g. Count errors that could be made by native speakers (e.g. between you and I).
h. Do not count register errors related to lexical choices (e.g. lots, kids).
i. Disregard an unfinished sentence at the end of the essay.

Word Count

a. Count contractions as one word whether correct or not.
b. Count numbers as one word.
c. Count proper nouns in English and in other languages as they are written.
d. Do not count hyphenated words as single words. (e.g. well-written = 2 words)
e. Don't include essay titles in word count.
f. Count words as they are written, even if they are incorrect. (e.g. alot = 1 word)

Appendix D

Error Classification System

New	Kroll (1990)	
1	1	whole sentence or clause aberrant
2	2	subject formation (including missing subject and existential, but not wrong case)
3	3	verb missing (not including auxiliary)
4	4	verb complement/object complement
	5	prepositional phrase/infinitive mixup
5	6	dangling/misplaced modifier
6	7	sentence fragment
7	8	run-on sentence (including comma splice)
8	9	parallel structure
9	10	relative clause formation (not including wrong or missing relative pronoun or resumptive pronoun)

10	11	word order
11	12	gapping error
12	13	extraneous words (not included elsewhere in descriptors)
13		missing word (not including preposition, article, verb, subject, relative pronoun)
	14	awkward phrasing
14		wrong modal
15	15	tense/aspect (incorrect tense, not incorrect formation)
16		voice (incorrect voice, not incorrect formation)
17	17	verb formation (including no auxiliary verb, lack of "to" with infinitive, participle misformation, gerund/infinitive problem)
18	18	subject-verb agreement
19	19	two-word verb (separation problem, incorrect particle)
20	20	noun-pronoun agreement (including wrong relative pronoun)
21	21	quantifier-noun agreement (much/many, this/these)
22	22	epenthentic pronoun (resumptive pronoun in relative clause, pronominal copy)
23	23	ambiguous/unlocatable reference
	24	voice shift
24		wrong case
25	25	lexical/phrase choice (including so/so that)
26	26	idiom
27	27	word form
28		wrong noun phrase morphology (but not word form)
29		wrong comparative formation
30	28	singular for plural
31	29	plural for singular
32	30	quantity words (few/a few, many kinds of, all/the whole)

33	31	preposition (incorrect, missing, extra)
34		genitive (missing/misused 's, N of N misuse)
35	32	article (missing, extra, incorrect)
36		deixis problem (this/that; the/this; it/that)
37	33	punctuation (missing, extra, wrong including restrict/non-restrictive problem—do not include capitalization)
38		negation (never/ever, any/some, either/neither, misplaced negator)

a. If sentence at the end of an essay is not finished, don't code it.
b. Code errors so that sentence is changed minimally. If there are two possible errors requiring equal change, code the first error.
c. If tense is incorrect and misformed, count it only as 15.
d. If error can be classified as a relative clause error, or a verb formation error (I know a man call John.), count it only as verb formation.
e. Don't double penalize for SV agreement:
ex. Visitor are pleased with the sight. (only a 30)
The man generally likes to go to work. (only a 30)

Appendix E

Categories of Errors Causing Rater Disagreement

1. *Lexicon, phrase, collocation* (Lex)

ex. It is *necessary that we really need* to talk face to face if we have time.

2. *Tense/aspect, voice* (T/A)

ex. Historically Japanese *had been influenced* by Confusionism and Buddism.

3. *Number of noun* (Num)

ex. The hobbies of my *families* are very typical in my country.

4. *Parallelism* (Par)

ex. They have a complete love, trust each other and respect each other.

5. *Punctuation* (Pun)

ex. So, this is the biggest and the most difficult problem in my country*(;)* Saudi Arabia.

6. *Preposition* (Prep)

ex. All parents have responsibility *for* their children.

7. *Reference* (Ref)

ex. Sometimes it is impossible to talk with *their* whole family face to face.

8. *Subject-verb agreement (SV)*

ex. In addition the teaching of all religions *encourage* a typical family.

9. *Article* (Art)

ex. My sister studies French language.

10. *Other determiner* (Det)

ex. So part of *my* reason why I'm now here is because of my father.

11. *Word form* (WF)

ex. Actually, Tokyo had a wonderful city *planning*.

12. *Modal* (Mod)

ex. But she wouldn't like to be.

13. *Verb form* (VF)

ex. They*('re)* always take care of me.

14. *Case* (Case)

ex. I'd rather *her* find something to do than worry.

15. *Missing word* (MW)

ex. Though some of the systems are co-ordinated, the others not.

16. *Reflexive* (Refl)

ex. All you have to do is relax *yourselves* and go home.

17. *Coordinator* (Coor)

ex. Since I am the last one in my family, all of my brothers and sisters, *also* my parents, love me very much.

18. *Word order* (WO)

ex. Why this day is so important for us?

Note: Parentheses indicate illegible writing. Many of the above sentences are obviously incorrect but were miscoded by mistake.