2-1-2016

Automated writing evaluation for formative assessment of second language writing: investigating the accuracy and usefulness of feedback as part of argument-based validation

Jim Ranalli
Iowa State University, jranalli@iastate.edu

Stephanie Link
Iowa State University, stephanielink06@gmail.com

Evgeny Chukharev-Hudilainen
Iowa State University, evgeny@iastate.edu

Automated writing evaluation for formative assessment of second language writing: investigating the accuracy and usefulness of feedback as part of argument-based validation

Abstract
An increasing number of studies on the use of tools for automated writing evaluation (AWE) in writing classrooms suggest growing interest in their potential for formative assessment. As with all assessments, these applications should be validated in terms of their intended interpretations and uses. A recent argument-based validation framework outlined inferences that require backing to support integration of one AWE tool, Criterion, into a college-level English as a Second Language (ESL) writing course. The present research appraised evidence for the assumptions underlying two inferences in this argument. In the first of two studies, we assessed evidence for the evaluation inference, which includes the assumption that Criterion provides students with accurate feedback. The second study focused on the utilisation inference involving the assumption that Criterion feedback is useful for students to make decisions about revisions. Results showed accuracy varied considerably across error types, as did students' abilities to use Criterion feedback to correct written errors. The findings can inform discussion of whether and how to integrate the use of AWE into writing classrooms while raising important questions regarding standards for validation of AWE as formative assessment, Criterion developers' approach to accuracy, and instructors' assumptions about the underlying purposes of AWE-based writing activities.

Keywords
Academic writing, argument-based validation, automated writing evaluation, ESL, formative assessment

Disciplines
Bilingual, Multilingual, and Multicultural Education | Curriculum and Instruction | Educational Assessment, Evaluation, and Research | Educational Methods

Comments
This is an Accepted Manuscript of an article published by Taylor & Francis in Educational Psychology: An International Journal of Experimental Educational Psychology on February 1, 2016, available online: http://www.tandf.com/10.1080/01443410.2015.1136407.

Automated Writing Evaluation for Formative Assessment of Second Language Writing: Investigating the Accuracy and Usefulness of Feedback as Part of Argument-Based Validation

Jim Ranalli, Stephanie Link, and Evgeny Chukharev-Hudilainen

Iowa State University

Author Note

Jim Ranalli, Department of English, Iowa State University; Stephanie Link, Department of English, Iowa State University; Evgeny Chukharev-Hudilainen, Department of English, Iowa State University.

Stephanie Link is now part of the English Department at Oklahoma State University.

Correspondence concerning this article should be addressed to Jim Ranalli, 319 Ross Hall, Department of English, Iowa State University, 50011. Email: jranalli@iastate.edu.

This is an Accepted Manuscript of an article published by Taylor & Francis in Educational Psychology: An International Journal of Experimental Educational Psychology on February 1, 2016, available online: http://www.tandf.com/10.1080/01443410.2015.1136407.

Abstract

An increasing number of studies on the use of tools for automated writing evaluation (AWE) in writing classrooms suggests growing interest in their potential for formative assessment. As with all assessments, these applications should be validated in terms of their intended interpretations and uses (Kane, 2012). A recent argument-based validation framework outlined inferences that require backing to support integration of one AWE tool, Criterion, into a college-level ESL writing course. The present research appraised evidence for the assumptions underlying two inferences in this argument. In the first of two studies, we assessed evidence for the evaluation inference, which includes the assumption that Criterion provides students with accurate feedback. The second study focused on the utilization inference involving the assumption that Criterion feedback is useful for students to make decisions about revisions. Results showed accuracy varied considerably across error types, as did students' abilities to use Criterion feedback to correct written errors. The findings can inform discussion of whether and how to integrate the use of AWE into writing classrooms while raising important questions regarding standards for validation of AWE as formative assessment, Criterion developers' approach to accuracy, and instructors' assumptions about the underlying purposes of AWE-based writing activities.

Keywords: Academic writing, argument-based validation, automated writing evaluation, ESL, formative assessment

Automated writing evaluation for formative assessment of second language writing: Investigating the accuracy and usefulness of feedback as part of argument-based validation

Introduction

An increasing amount of research on the use of automated writing evaluation (AWE) in writing classes suggests growing interest among program administrators and instructors in AWE's potential to provide formative assessment; that is, assessment that can support learning and teaching. AWE tools, which employ natural-language processing, machine-learning, or other computational methods in the analysis of text, can provide both scores on writing quality as well as qualitative feedback on aspects of grammar, mechanics, style, discourse, and organization. While the use of AWE for scoring purposes remains controversial because of its connection to large-scale standardized tests, formative uses as support for writing instruction are viewed more positively (Ware, 2011).

Originally developed for use by native speakers of English, the most commonly used commercial AWE systems, such as Educational Testing Service's Criterion Online Writing Evaluation Service, are increasingly marketed as useful for second language (L2) learners. Given that L2 learners have a greater need for feedback on sentence-level correctness, which AWE systems are more computationally adept at providing compared to feedback on higher-level concerns (Weigle, 2013a), a case can be made for the use of AWE as a complement to instructor feedback in L2 writing classrooms. In this role, AWE promises greater autonomy for students while potentially freeing up instructors to devote their feedback efforts to aspects of writing that

require human evaluation (Chen & Cheng, 2008; Li, Link, & Hegelheimer, 2015; Warschauer & Grimes, 2008) such as audience awareness and communicative effectiveness.

Arguments in favor of formative applications rest, however, on the assumption that AWE feedback is accurate and useful. While some recent research has investigated AWE accuracy and usefulness separately and in different ways (Dikli & Bleyle, 2014; Lavolette, Polio, & Kahng, 2015), these studies define accuracy with reference to developer- rather than user-centric standards and leave aside the question of how accuracy problems may affect usefulness. The present research seeks to investigate these issues from the perspective of argument-based validation (Kane, 1992, 2012, 2013, 2015), which provides a framework for combining potentially disparate forms of validity evidence in appraising particular interpretations or uses of an assessment. In two studies described here, we appraise evidence regarding two inferences in an interpretation/use argument for the use of AWE as a formative assessment tool in college-level ESL writing courses. The first inference, called *evaluation*, focuses on the accuracy of implementation and adherence to the conditions of standardization in automated feeback generation (Clauser, Kane, & Swanson, 2002). The second inference, *utilization*, addresses the usefulness of AWE feedback to students in making decisions about revisions. Because "choices made in developing the scoring algorithm may have an important impact on the strength of this aspect of the argument" (Clauser et al., 2002, p. 424), it is important to investigate accuracy and usefulness in conjunction with each other.

Argument-Based Validation for AWE as Formative Assessment

According to Kane, it is the interpretations and uses of assessments, rather than the assessments themselves, that require validation (Kane, 1992, 2012, 2013, 2015). This requirement is seen to extend to formative assessments whose purpose is to support learning and

teaching (Bennet, 2011). If students are the intended audience for the feedback generated by a formative assessment but do not use that feedback, this nonuse poses a threat to the validity of an interpretation stating that the assessment, through its feedback, is beneficial to students (Doe, 2015).

Kane's argument-based approach to validation contains two steps: construction of an interpretation/use argument (IUA) and critical evaluation of the IUA in a validity argument (Kane, 2013, 2015). In the first step, the IUA specifies a network of inferences about the proposed score interpretations and uses of an assessment. Each of these inferences are authorized by a warrant, a generally held principle or statement that links the data to a claim about the interpretations and uses of the assessment. Underlying the warrant are assumptions that connect observations about the scores to conclusions and decisions that can be made based on the scores. Each assumption requires evidence, or backing, that can take a variety of forms as determined through a validity argument. On occasion, circumstances may undermine the inference, thereby rebutting the strength of the IUA, in which case, a rebuttal, or refutation of counterclaims, can be formed and additional evidence would likely be necessary.

In the second step, the validity argument is used to appraise, or critically evaluate the "coherence, completeness, and plausibility of the claims being made" (Kane, 2015, p. 5) in the IUA through theoretical rationales, empirical data, or both. For empirical validity arguments, the assumptions from the IUA turn into research questions and then methods can be formulated to systematically investigate these questions. The amount of evidence necessary to support a claim is dependent on the strength of the claim being made about the proposed interpretations and uses. That is, an IUA with an extensive network of inferences and assumptions would require more support than one with a relatively small number of inferences and assumptions (Kane, 2015).

Although Kane's argument-based approach does not define the methods necessary for evaluating inferences, the validity argument is extremely valuable as a systematic and evidence-based approach (Aryadoust, 2013). Thus, the present study provides evidence for two inferences and a set of assumptions that fit within a more extensive framework proposed by Chapelle, Cotos, and Lee (2015) in order to investigate the use of automated writing evaluation for formative assessment of L2 writing.

Chapelle et al., (2015) presented their IUA based on seven inferences that would require backing to support the integration of Criterion into a college-level ESL writing course at Iowa State University. Five of their inferences (*evaluation, generalization, explanation, extrapolation,* and *utilization*) are outlined in Clauser et al., (2002), who conceptualized how to frame a validity argument on the basis of results from an automated scoring assessment. In addition to those five inferences, Chapelle et al., (2015) included two additional inferences: *ramification*, which is critical for making the claim that learning results from assessment use, and *domain definition*, which is important for defining the real-world domain of interest (see also Chapelle, 2011; Chapelle, Enright, & Jamieson, 2008). The authors of the present study are associated with Iowa State University and its ESL writing program; the first author is currently a course coordinator who decides what materials and tools will be incorporated into the curriculum. To help us determine whether use of Criterion is warranted in our working context and, further, to encourage broader discussion of standards for AWE as formative assessment, we appraise evidence for two inferences from Chapelle et al.'s framework that are of immediate concern in investigating the value of AWE to support L2 writing instruction. We focus on the evaluation inference, which is based on assumptions regarding the accuracy of Criterion feedback, and the utilization inference, which is based on assumptions regarding the usefulness of Criterion

feedback. Research suggesting that inaccuracies in AWE feedback affects students' use of AWE tools, which prompted our selection of these two inferences, is reviewed in the following sections.

Accuracy

Understanding the scope of the accuracy problem -- that is, the extent to which students are provided with inaccurate feedback by AWE tools and what forms these inaccuracies take -- is challenging because of limited information on the actual performance of AWE tools and the different ways accuracy is conceptualized. While Chapelle et al. stipulate that Criterion feedback must provide students with accurate information to target relevant areas for revision, improvement, and learning, they do not specify what constitutes accurate information. Research often cited in discussions of AWE accuracy typically focus on methodological concerns of AWE developers, addressing only one or two linguistic "micro-features," such as articles (Han, Chodorow, & Leacock, 2006) or preposition errors (Chodorow, Tetreault, & Han, 2007). The purpose of such studies is to evaluate the quality of error-detection methods, with little or no consideration given to the ways in which information about errors and suggested remedies are communicated to AWE users. Such studies thus represent a system-centric (i.e., focused on system performance) rather than user-centric (i.e., focused on the user's interaction with the system) viewpoint (Chodorow, Gamon, & Tetreault, 2010).

In such studies, measurement focuses on *precision* and *recall*, two concepts from the field of information science. Precision is a measure of how often the system is correct when it reports finding an error, whereas recall measures the system's coverage; that is, the proportion of actual usage errors that have been detected (Leacock, Chodorow, Gamon, & Tetreault, 2010). For example, data from two system-centric studies (Burstein, Chodorow, & Leacock, 2004;

Chodorow, Gamon, & Tetreault, 2010) show that, of all the features Criterion identifies as preposition errors (Table 1), human annotators will agree in 78% of cases (i.e., precision = .78); however, Criterion will only flag 18% of the preposition errors identified by human annotators (i.e., recall = .18). Comprehensive data on precision and recall for Criterion's array of micro-features is not publicly available.

[INSERT TABLE 1 ABOUT HERE]

Criterion's developers prioritize precision over recall on the assumption that it is better to miss some errors than to incorrectly flag well-formed text as ill-formed (Burstein et al., 2003; Leacock et al., 2010). Perhaps as a result of this policy, recall data are not always reported, and when they are, scant attention is paid to their implications. This is a concern since low recall indicates a large proportion of errors are missed by the system. Given the limitations of system-centric research for use in validating AWE as formative assessment, it is important to investigate how inaccuracies affect student writers.

Recently, two user-centric studies evaluated the accuracy of Criterion feedback. Dikli and Bleyle (2014) compared Criterion feedback to feedback provided by an ESL instructor in an English for Academic Purposes (EAP) class in a university in the southeastern U.S. and found large discrepancies between the two feedback sources. Many of the error types accurately identified by the instructor were missed or mislabelled by Criterion, while the instructor was found in general to provide more, and higher quality, feedback. Criterion was also found to underidentify error types known to frequently occur in university ESL students' writing, such as pronoun and verb-form errors (Ferris, 2011). In another study involving undergraduates enrolled in university-level ESL writing classes, Lavolette et al. (2015) found that Criterion error identifications were correct 75% of the time (high precision), but the system missed at least 46%

of all errors (low recall). Wide variation in accuracy was observed among different error types; for example, human annotators concurred in 95% of cases where Criterion flagged an Ill-Formed Verb but only 49% of cases where it indicated a Run-On Sentence. These findings, taken together, connect accuracy problems to concerns about the usefulness of Criterion feedback.

Usefulness

In their interpretation-use argument, Chapelle et al. (2015) do not define usefulness other than to say that useful Criterion feedback allows students to make decisions about revisions. By contrast, classroom-based studies of AWE have defined the usefulness of AWE feedback in terms of clarity, specificity, and relevance, as well as the types of features that AWE feedback can address. In such studies, students' and teachers' perceptions of usefulness feedback are typically mixed. For example, Dikli and Bleyle (2014) found generally positive views of Criterion feedback among ESL students in the EAP class they studied, although the students preferred feedback from the instructor. In a multi-case study of AWE used in EFL writing classrooms (Chen & Cheng, 2008), one teacher believed use of a tool called MY Access! facilitated more drafting and revising behavior while another felt the program only helped with some basic points of form and organization. Students surveyed in the study were also divided, with 45% believing the tool was unhelpful.

Research involving revision behavior based on Criterion feedback in particular shows pervasive underuse. In a study of Criterion use in grades 6-12 involving thousands of students (both English L2 learners and English native speakers) across the U.S., more than 70% of 33,171 essay submissions were submitted only once for feedback, demonstrating that "... most students did not exploit the revision capabilities of the *Criterion* system" (Attali, 2004, p. 4). A study of Criterion-based revisions made by university-level ESL writing students (Lee, Li, &

Hegelheimer, 2012; see also Chapelle et al., 2015) documented six types of response to Criterion feedback: *no change, remove, add, delete, change,* and *transpose*. In more than 50% of cases, students made no changes, which the authors attributed to frequently inaccurate feedback.

These findings invite comparisons to the problem of widespread nonuse of help options within technology-based learning environments, such as glossaries, annotations, and feedback messages. Research shows learners frequently attribute nonuse of help options to their distracting influence (Aleven, Stahl, Schworm, Fischer, & Wallace, 2003; Cárdenas-Claros & Gruba, 2009), which suggests problems of cognitive load and mental effort. Cognitive load is the load imposed on a learner's cognitive system by performing a particular task, whereas mental effort is the cognitive capacity allocated by the learner to addressing task demands (Paas, Tuovinen, Tabbers, & Van Gerven, 2003). Even if help options are perceived as useful, learners may refrain from using them if addressing the task while simultaneously processing the help imposes too much cognitive load (Aleven et al., 2003). Considering the central role that management of cognitive demands also plays in influential models of the writing process (Flower & Hayes, 1981; Bereiter & Scardamalia, 1987), we therefore included mental effort (described in more detail below in Study 2) in the design of our evaluation of the usefulness of AWE feedback.

The Present Research

To appraise validity evidence for the evaluation and utilization inferences in Chapelle et al.'s (2015) interpretive argument about Criterion as an AWE-based assessment, we conducted two studies based on data from two ESL writing courses at Iowa State University. In Study 1, experts rated the accuracy of Criterion feedback provided for 10 error types commonly identified in the students' writing. In Study 2, which involved student volunteers in the lower (n = 36) and higher-level (n = 46) courses, we evaluated usefulness in terms of students' performance on an

error-correction task involving Criterion feedback, as well as the amount of mental effort students perceived in distinguishing between accurate and inaccurate feedback. The lower-level course in Study 2 focuses on linguistic form and addresses academic writing at the sentence and paragraph levels, whereas the higher-level course focuses on genre, discourse, and rhetorical strategies in essay-length assignments. Students at both levels are given access to Criterion outside of class while the amount of in-class use varies depending on the instructor. Neither level makes use of the writing prompts that accompany Criterion, so the system's holistic scoring feature, which requires use of these prompts, is deactivated.[1]

In the following sections, we report the two studies and then use the results to examine the extent to which they support the assumptions underlying the evaluation and utilization inferences in Chapelle et al.'s (2015) interpretation/use argument.

Study 1

Study 1 addressed the evaluation inference. In their interpretation/use argument, Chapelle et al. formulated the warrant for the evaluation inference this way: "Criterion feedback provides students with accurate information to target relevant areas for revision/improvement/learning" (2015, p. 3). The first assumption underlying this warrant is that "Criterion feedback is accurate" (2015, p. 3). In their study, however, Chapelle et al. do not present evidence supporting this assumption or the evaluation inference more generally, while the data they provide in support of a different inference -- a sample of 294 error identifications by Criterion, 115 of which were deemed inaccurate -- suggests an accuracy rate of only 61%. No information is provided about types of error or how accuracy was evaluated. To address these gaps and determine whether empirical evidence would provide backing for the assumption regarding Criterion's accuracy, we

therefore formulated the following research question: *How accurate is Criterion feedback in terms of the errors it commonly identifies in our students' writing?*

We adopted 70% as a provisional, lower-bound threshold for evaluating the accuracy of Criterion feedback based on the developers' standard of 80%, or precision of .8,[2] that must be achieved before a new micro-feature can be incorporated into *e-rater*, the scoring engine that underlies the Criterion system (Quinlan, Higgins, & Wolff, 2009). We allowed a 10% margin because our definition of accuracy takes into account additional aspects of Criterion feedback not considered by developers; that is, the textual commentary and highlighting that accompany error categorization (Figure 1). Our rationale was that evaluation of AWE feedback for formative purposes must include these presentational elements insofar as they help users locate, understand, and respond to errors. The 70% threshold is provisional because no research has yet investigated levels of AWE feedback accuracy needed to support formative assessment.

[INSERT FIGURE 1 ABOUT HERE]

On the basis of this broader definition, Criterion feedback can be divided into two types: generic and facilitative (e.g., *You may be using the wrong preposition*) or specific and directive (e.g., *You have used* quiet *in this sentence. You may need to use* quite *instead).* The former, henceforth referred to as *generic* feedback, takes the same form regardless of the particular error, while the latter, henceforth referred to as *specific* feedback, incorporates some aspect of the student text in its formulation, either in the recommendation of a specific word to fit the context or by situating a suggested operation with reference to a highlighted textual feature (e.g., *You may need to remove this comma*). Whether generic or specific feedback is generated by the system depends on the type of error.

We expected to find potential differences in accuracy compared to previous research based on two possibilities. First, there might be cases where error categorization was appropriate but the accompanying textual commentary, highlighting, or both (especially in the case of specific feedback) were not; for example, when Criterion identifies a missing article in a clause such as ... *salary is important factor* ... but proposes the use of *a* instead of *an*. Second, feature-detection algorithms have been known to perform differently when deployed in combination compared to their attested performance in isolation during development (Quinlan et al., 2012).

Methods

Study 1 was based on a corpus of all drafts of writing submitted to Criterion by 370 students in the two writing courses described above during the Fall 2013 semester, including 102 in five sections of the lower-level course and 268 in 14 sections of the higher-level course. Online reports were generated in Criterion, saved and converted from html to text format, and then analyzed using AntConc (Anthony, 2014), a simple concordancing tool, to quantify the instances of each error type flagged by Criterion (Table 2). We decided to focus our analysis on the 10 most common error types since many types were flagged infrequently; for example, in the case of Possessive Errors, an average of 1.6 times per student.

Features in the Style category were excluded as too subjective, as were spelling errors, which students can easily address using native spell checkers in word processors. One error type, Compound Words, was omitted because Criterion was observed to frequently flag a non-erroneous string of text.[3] This left a set of 10 common error types representing Criterion's categories of Grammar (4), Mechanics (2), and Usage (4). Of these, Missing or Extra Article errors were the most common, representing 11% of all errors in the corpus, while the least

frequent of the 10, Confused Words, represented 3%. All remaining error types, with the exception of the nonspecific Proofread This!, constituted 2% or less of the total.

[INSERT TABLE 2 ABOUT HERE]

Natural-language processing techniques were used to isolate every sentence in the corpus containing a flagged error. Specifically, sentence boundaries were identified with a regular expression (Friedl, 2006) that looked for sentence-final punctuation marks. For Run-on Sentences and Fragments, the complete sentence was extracted from the corpus and used as material for annotation. For all other errors, text spans of 100 characters to the left and to the right of the flagged error were retrieved to provide contextualization. These units were collected into a subcorpus of Criterion errors, which was used first for training and calibration and then for the rating procedure reported below. The first and second authors conducted the ratings in a specially developed web-based tool using a polytomous, interval scale: 1 for *not accurate*; 2 for *partially accurate*; and 3 for *completely accurate* (Figure 2), with the partially accurate category included to cover cases where Criterion's error categorization was appropriate but some other aspect of the feedback, such as a suggested correction, was not.

[INSERT FIGURE 2 ABOUT HERE]

During training and calibration, we annotated errors randomly selected from the subcorpus while the system recorded our ratings and displayed a constantly updated agreement rate. In between training sessions, the system allowed us to review items where there were disagreements, which we discussed to achieve a shared understanding of the different error types and to develop a list of decision rules. At the end of training and calibration, we had co-rated a set of 360 errors and were consistently achieving an agreement rate of greater than .7 using Krippendorff's alpha[4] (Hayes & Krippendorff, 2007). At this point, we began rating a set of 700

errors (70 of each of the 10 most common types) randomly sampled from the subcorpus. Final inter-rater reliability was high, Krippendorff's $\alpha = .74$, Cronbach's $\alpha = .85$. Discrepancies were resolved by randomly selecting one set of annotations to use in our subsequent analyses.

Results

The results of the accuracy ratings are presented in Table 3. The descriptive statistics include raw frequencies for the errors in each category where Criterion's feedback was rated completely accurate, partially accurate, or not accurate; and percentages of errors in each category in which Criterion's feedback was rated either completely or partially accurate, or completely accurate. The latter represent, respectively, a more lenient and a stricter standard. We present these two standards because it is not currently known whether or to what extent inaccuracies in highlighting or textual commentary affect learners' ability to make use of feedback Criterion feedback, and so it is difficult to determine which is the more appropriate metric for evaluation.

[INSERT TABLE 3 ABOUT HERE]

If the 10 common error types are considered in the aggregate, Criterion's feedback meets or exceeds the 70% threshold for accuracy, regardless of the standard adopted. Among individual error types, some performed well in either case. Ill-formed Verbs topped the list, with a 9% difference between the stricter and more lenient standards. Subject-Verb Agreement errors, Determiner-Noun Agreement errors, and Fragments follow, with little or no differences between the standards.

Other error types show poor performance regardless of the standard adopted. Extra Comma errors showed the lowest accuracy, followed by Run-On Sentences and Confused

Words. Preposition Error, Missing Comma, and Missing or Extra Article underperform according to the stricter standard but reach the 70% threshold using the more lenient standard.

Large discrepancies across the standards are seen in the performance of two error types: Missing or Extra Article (71.4% versus 58.6%) and Preposition Error (87.1% versus 65.7%). With the former error type, the high proportion of partially accurate ratings arose from problems with Criterion's specific suggested alternatives; in the latter, it arose from cases where changing the preposition made sense at phrase-level but was of uncertain value at sentence-level (e.g., ... *Guangzho attracts many people because of city location and good cooking methods* ***with*** *good taste*).

Discussion

Study 1 assessed the accuracy of Criterion feedback on 10 error types commonly identified by the program in our students' writing. The findings showed that in the aggregate, Criterion performs adequately regardless of the standard adopted, but among individual error types there is considerable variation, particularly considering differences between the stricter and more lenient standards.

Some findings of the current study align with those of Lavolette et al. (2015), who also found high accuracy for Ill-Formed Verbs and Subject-verb Agreement and low accuracy for Run-On Sentence errors. However, differences were found between the two studies with regard to Confused Words (.89 in Lavolette et al. versus .60 in the present study) and Fragments (.60 versus .87, respectively). In the case of Confused Words, the difference may be attributable to the fact that Lavolette et al. did not take Criterion's textual commentary into account. Our annotated data included numerous cases where the categorization of Confused Words was appropriate but the suggested alternative word did not fit the context. Regarding Fragments,

many students in our sample neglected to omit titles or headings from their compositions before submission to Criterion, despite being instructed to do so.

Other differences between the two sets of findings must be attributed to different user population characteristics, writing tasks, and other unique features of each context. They point to potentially wide variation in Criterion's accuracy across contexts of use, which suggests a need for research into the scope of this potential variation. This and other implications of Study 1 are addressed in the General Discussion below, following the description of Study 2, to which we now turn.

Study 2

Study 2 addressed the utilization inference, for which Chapelle et al. provide the following warrant: "Diagnostic results on the quality of academic writing obtained from Criterion are useful for students to make decisions on revisions" (2015, p. 3). The following assumptions underlie this warrant:

1. The meaning of the Criterion feedback is clearly interpretable by students.
2. Students are willing to use Criterion in their writing process.
3. Students use diagnostic results to make decisions on how to revise their drafts and correct errors.
4. Criterion provides necessary assistance beyond feedback to help revision [*sic*] process.

(Chapelle et al., 2015, p. 3)

We find these assumptions to be limited, however, in that the nature of students' revisions and the potentially detrimental effects of frequent exposure to inaccurate feedback are not taken into account. Thus, rather than examining these assumptions per se, we augment the Chapelle et al.'s argument with two of our own that more specifically address features of usefulness we deem important based on previous research and the goals of our writing courses: (5) *Criterion feedback specifically supports correction of the errors identified by the system, as opposed to other ways*

that students might use the feedback; and (6) *The need to differentiate between accurate and inaccurate feedback does not overburden users' cognitive-processing capacities, which could affect their willingness to use the feedback.*

Our rationale for Assumption 5 is that, in addition to correcting or ignoring a flagged error, Criterion users may decide to rewrite the relevant section such that the error is avoided or delete the section altogether. While these latter strategies may result in error-free text, they arguably do little to promote L2 development, which is a primary goal of the ESL courses in question. We therefore seek evidence that student writers can use the automated feedback specifically in the task of error correction. Unlike Study 1, a provisional standard for error-correction performance is not specified a priori because there is no research or other precedent to draw on. Study 2 is therefore exploratory in this sense.

Regarding Assumption 6 and the issue of cognitive demands, our rationale is that large amounts of unused feedback have been attributed to issues with the accuracy of the feedback while nonuse of help options has been connected to issues of cognitive demands and perceived mental effort. We therefore seek evidence to show that the cognitive demands of differentiating between accurate and inaccurate feedback are not so high as to compel learners to ignore feedback in Criterion.

Based on these two additional assumptions, we formulated the following research questions:

1. *How well can students use Criterion feedback to correct errors that have been accurately identified by the system?*
2. *Does the need to distinguish between accurate and inaccurate Criterion feedback impose cognitive demands on students that could affect their willingness to use the feedback?*

We also included cross-level comparisons in our analyses to determine if validity evidence supported the utilization inference differentially across the two courses.

Methods

Participants. Eighty-two students in the Spring 2014 semester participated: 36 from two sections of the lower-level course and 46 from five sections of the upper-level course. The sample had an average age of 20.3 years and included 30 females and 52 males. Most spoke Chinese (53) as their first language, followed by Korean (20), Malay (3), Vietnamese (3), Persian (2), and Thai (1).

Task. Our research objectives required a way to constrain participants' choices in responding to Criterion feedback while also allowing measurement of the effects of differentiating between accurate and inaccurate feedback on perceived mental effort. For this reason, we developed an error-correction task involving mock-ups of Criterion feedback extracted from the corpus described in Study 1, rather than real-time feedback delivered on students' own writing within Criterion itself. The feedback was simulated using screenshots from Criterion and design elements in Microsoft PowerPoint (Figure 3). The task was delivered using Qualtrics (Qualtrics Inc. 2015), an online survey tool, and consisted of three parts.

[INSERT FIGURE 3 ABOUT HERE]

Part 1 included one example of each of the 10 error types appearing as an image above a text box in which participants made their corrections. The text containing the error was pre-populated in the box to reduce the chance of new errors being created while typing. Below, a slider bar was used to indicate perceived mental effort on a 7-point scale from 1 = *very little*

mental effort to 7 = *a lot of mental effort*, which is commonly used in studies of cognitive load (Paas et al., 2003).

In written instructions at the beginning of the task and in verbal instructions and a demonstration given before participants began, it was emphasized that the feedback in Part 1 was accurate. As determined by a panel of four experienced, native-speaking ESL writing teachers, only those items that were unanimously agreed to contain completely accurate feedback were included. Thus, participants did not have to evaluate the accuracy of the feedback but instead could devote their cognition to understanding and using it for error-correction purposes.

Part 2 of the task consisted of 20 error-correction items, two of each of the 10 error types, distributed randomly. For each type, one item had been deemed completely accurate and the other inaccurate by unanimous consent of the same panel of ESL instructors. For each item, the simulated Criterion feedback appeared at the top, followed by a yes/no question based on the statement: "This feedback by Criterion is accurate." If the student responded yes, the text box would appear, allowing the participant to correct the error. If the student responded no, the textbox would not appear and no correction was elicited. In either case, participants would indicate perceived mental effort using the 7-point slider bar.

Part 3 was a single survey page containing eight biodata-related questions.

Scoring. Participants' error corrections were scored by the first and second authors separately using a polytomous, interval scale: 0 for *not correct*; 1 for *partially correct*, and 2 for *fully correct*. Inter-rater reliability was excellent (Krippendorff's $\alpha = .86$; Cronbach's $\alpha = .93$). Discrepancies were again resolved by randomly selecting one set of annotations to use in the subsequent analyses.

Procedures. Because the task involved a pedagogical activity seen to enhance the curriculum, it was given to all students in those sections taught by instructors who had agreed to participate in the study. A member of the research team arranged to visit these classes during a regularly scheduled computer-lab session. The task was explained and then students were given 40 minutes to complete it. At the end of the session, an informed consent document was circulated; only those students who signed it were included in our analyses.

To incentivize students to make a sincere effort on the task, it was announced that gift cards would be awarded to the two students (one at each level) who achieved the highest scores on the error-correction component. In addition, all students were provided with feedback on their individual performances via email, while instructors received aggregated results including descriptive statistics and qualitative data on the performance of each participating class, which we encouraged instructors to use in helping students make more effective use of Criterion.

Results

Error correction ability. On Part 1 of the task, where no accuracy determination was required, the average score for the whole sample (N = 82) was 11.67 out of 20 possible points (SD = 2.14), or 58.3%. The average score for the lower-level participants (n = 36) was 12.1 (SD = 2.45), or 60.4%, while the higher-level group (n = 46) scored 11.33 (SD = 1.81), or 56.6%. Average scores on Part 2, where accuracy discrimination was required, were actually higher than on Part 1. The average for the whole sample (N = 82) was 12.82 (SD = 3.71) or 64.1%, with the lower-level group (n = 36) scoring 13.22 (SD = 3.26) or 66.1%, and the higher-level group (n = 46) scoring 12.5 (SD = 4.06), or 62.5%.

The higher standard deviations on Part 2 are attributable to the fact that, despite the higher overall average compared to Part 1, there were more individual scores of 0 because of

participants making corrections in cases where feedback was not accurate or not making corrections when they erroneously deemed feedback to be inaccurate. It is also notable that the lower-level group scored higher on both parts of the task than their higher-level counterparts. This finding will be taken up in the discussion.

Some error types proved more challenging to correct than others (Table 4). Participants scored highly with corrections involving Confused Words, Extra Comma, Missing Comma, Missing or Extra Article, and Subject-Verb Agreement, but scores were lower on items involving Determiner-Noun Agreement, Fragments, and Run-On Sentences. Differences were observed across parts of the task in the participants' ability to correct Ill-Formed Verbs and, to a lesser extent, Preposition Errors, which we attribute to the relative ease or difficulty of the individual items. The relative difficulty of specific items may in fact account for the higher scores on Part 2, despite the addition of the accuracy-determination factor.

[INSERT TABLE 4 ABOUT HERE]

Mental effort. Descriptive statistics showed participants reported lower perceived mental effort on Part 2 of the task, which required accuracy determination, compared to Part 1, which did not. On the 7-point scale, the average rating for the lower-level group (n = 36) on Part 2 was 2.04 (SD = .93) versus 2.09 (SD = .65) on Part 1, whereas the higher-level group (n = 46) reported an average of 2.63 (SD = 1.34) on Part 2 versus 2.32 (SD = 1.37) on Part 1.

To test for significance, we used Wilcoxon signed-rank tests because of a non-normal distribution in the mental effort data for both levels. No statistical differences were found at the lower-level, (n = 36), Z = -.75, p = .46, or the higher level, (n = 46), Z = -.31, p = .75. This suggests that the need to differentiate between accurate and inaccurate feedback did not lead to measurable increases in perceived mental effort for participants at either level.

While overall the mental effort ratings were low, a breakdown according to error type showed some potentially non-random variation between types categorized as specific versus generic (Figure 4).[5] An independent samples t-test was performed to compare mental effort across these categories. Results showed a significant difference in pooled mental-effort ratings for generic feedback (M = 2.55, SD = 1.43, n = 984) and specific feedback (M = 2.09, SD = 1.40, n = 656), $t(1638) = 6.45$, $p < .001$, with an effect size of .31 (Cohen's d), which suggests participants found it more mentally taxing to work with generic feedback than specific feedback.

[INSERT FIGURE 4 ABOUT HERE]

Discussion

Study 2 investigated the degree to which ESL students are able to use Criterion feedback to correct frequently identified error types as well as the amount of mental effort they perceived in differentiating between accurate and inaccurate feedback. Results showed participants were able to make appropriate corrections about 60% of the time and that accuracy determination did not appear to require increased mental effort.

Regarding the lower-level group's higher scores and lower mental effort ratings, it is important to remember that the lower-level course focuses on sentence- and paragraph-level writing, with classroom attention regularly given to presentation and practice of points of grammar, mechanics, and usage. To the extent these issues are addressed in the higher-level course, it is on an ad hoc basis, typically by means of feedback from instructors. It was also reported anecdotally that instructors in the two sections of the lower-level course gave their students more frequent opportunities to use Criterion during computer lab time, which may have given those participants greater familiarity and facility with Criterion feedback.

The fact that participants did not report higher mental effort as a result of accuracy determination suggests this process may simply be integrated into a larger process of making sense of, and deciding whether and how to use, AWE feedback. Importantly, it does not mean the accuracy of Criterion feedback may not in some other way affect students' willingness to use it. In addition, the slight but systematic variation in mental effort ratings observed between specific and generic types of Criterion feedback suggests this aspect of AWE feedback is worth further investigation. This and other implications will be taken up now in a discussion of the findings of the larger project.

General Discussion

In this section, we appraise the results of the two studies in terms of validity evidence for the evaluation and utilization inferences in the interpretation/use argument formulated by Chapelle et al. (2015). This appraisal can help researchers and decision makers with regard to classroom uses of Criterion understand the extent to which Criterion feedback is accurate and useful enough to support L2 writing instruction in contexts like ours, as well as how accuracy and usefulness might be conceptualized, defined, and evaluated in future research addressing applications of AWE for formative assessment.

The Evaluation Inference and the Accuracy of Criterion Feedback

The evaluation inference was based on the warrant that Criterion feedback provides students with accurate information for targeting relevant areas for revision, improvement, and learning. The first assumption underlying this warrant stated that Criterion feedback is accurate. We investigated the accuracy of Criterion's feedback through 10 error types it frequently identified in our students' writing. The results showed the feedback to be accurate between 71-77% of the time when considering the 10 error types in the aggregate, which conformed to our

provisional standard. However, considerable variation in accuracy was found among individual error types, with several performing far below the 70% standard, including errors that are of importance to ESL/EFL writers. We interpret these results as providing only limited support for the assumption regarding the accuracy of Criterion feedback. We also see the potential, noted by Chapelle et al. (2015), for a rebuttal to be added to the interpretation/use argument stating that Criterion's inaccuracies undermine students' confidence in the system, making them reluctant to use it. Assuming support for this rebuttal were found, the validity argument for use of Criterion as a formative assessment tool in our context would obviously be undermined.

Unlike recent studies of classroom uses of AWE, and in a departure from previous system-centric research that only addresses feature detection and categorization, the present investigation took into account the text and highlighting that accompany AWE feedback, which necessitated an intermediate value for rating, *partially accurate*. This value was responsible for a considerable amount of variation in our findings and it is therefore important that validation research be conducted into the effects of inaccurate textual commentary and highlighting on students' corrective abilities and their perceptions of the understandability and usefulness of the feedback. This will be important for clarifying standards for use in future validation research into classroom uses of AWE.

Although the study did not investigate the errors Criterion missed, it raises some concerns regarding recall, which is also encompassed by the first assumption underlying the evaluation inference regarding Criterion's accuracy. It must be remembered that while the 10 error types we investigated were found frequently by Criterion in our students' writing, this does not mean these errors are all serious in terms of their effects on meaning or actually more common in our students writing than other errors Criterion is less adept at finding. For example, in our corpus of

error types, Wrong Part of Speech was found infrequently by Criterion, but a study of college-level ESL learners' written errors showed this to be a frequent problem (Chan, 2010). Another significant potential rebuttal that could undermine the validity argument for Criterion used as formative assessment would be one stating that errors found to be common among ESL writers and considered serious by instructors are detected less frequently by Criterion than other errors in our students' writing. Research addressing this rebuttal could also investigate the effects of low recall on writing quality as well as students' and teachers' perceptions of AWE feedback and tools.

The Utilization Inference and the Usefulness of AWE Feedback

The warrant for the utilization inference stated that Criterion's diagnostic feedback on academic writing is useful for students to make decisions about revisions. Based on relevant research and the goals of our writing program, we specified two new assumptions underlying this warrant. The first of these (Assumption 5) stated that Criterion feedback specifically supports correction of the errors identified by the system. Our results showed students were able to correct errors based on Criterion feedback 55-65% of the time. In the absence of an established standard, it is impossible to say whether this middling finding is an acceptable return on investment of time in using Criterion.

In appraising this finding, it became clear to us that establishing such a standard might require specifying whether the goal of a particular AWE-based writing task was primarily focused on learning to write or on L2 development. Weigle (2013b) has connected validation of use of AWE with English language learners to a distinction between learning to write (LW) and writing to learn (WL), as elaborated in Manchón (2011). If one's focus is writing skills development or support for writing practice (both of which could be characterized as LW), 60%

may be deemed insufficient, since the benefits to the quality of the final written product might not be outweighed by the costs in terms of time spent addressing inaccurate AWE feedback, which could be better used for some other aspect of writing. Conversely, if one considers the purpose of an L2 writing activity as primarily supporting opportunities for L2 development (i.e., WL), then 60% success may be acceptable, since even inaccurate AWE feedback can cause students to notice linguistic forms, which potentially facilitates acquisition (Schmidt, 1994). Therefore, our findings constitute only partial support for the first assumption regarding support for error correction, contingent on specification of the goals of the writing task in question.

Our second assumption (Assumption 6) was that the need to differentiate between accurate and inaccurate Criterion feedback does not overburden users' cognitive-processing capacities. The present findings provided clear support for this assumption, with the caveat that accuracy determination in more authentic writing tasks might elicit different perceptions of mental effort. Taken together, our results provide limited support for the utilization inference in Chapelle et al. (2015) and point to several types of research that will be needed for more unequivocal validation of classroom applications of Criterion.

Implications

The findings of partial support for the evaluation and utilization inferences have implications for the integration of Criterion into L2 classroom writing instruction and assessment. Viewed in isolation, the results from Study 1 could suggest that Criterion is not accurate enough to provide useful formative feedback. However, the argument-based validity framework allows us to consider this finding alongside the results of Study 2, which provide additional insights. Despite problems with the accuracy of the feedback, participants from the lower-level course were able to take better advantage of the feedback in correcting errors,

probably because their course's more explicit attention to linguistic form seemed to predispose them to making better use of corrective feedback. This suggests that the use of Criterion may be more justified on courses where there is a congruent focus on form, supporting previous research showing that the manner in which AWE is integrated into instruction influences its acceptance by students (Chen & Cheng, 2008; Li, Link, & Hegelheimer, 2015). We see this distinction applying to the lower- and higher-level courses in our program, and for this reason, we determined that use of Criterion was more justifiable in the lower-level course.

Whether AWE feedback is generic or specific also makes a difference. The findings showed students can make better use of feedback that provides clear information about the location, nature, and remediation of errors, and that interacting with such feedback is less mentally taxing. AWE tools are constrained in their ability to provide specific feedback in all cases. Moreover, error types such as Run-on Sentences and Fragments may require additional instruction and practice to both diagnose and correct. In addition, the nature of errors is such that individual errors of a particular type may be more or less difficult to understand and address, even when correctly identified. All of this will make specification of standards for usefulness challenging, but such standards will be central to future validation efforts.

In addition to the implications for writing instruction and assessment, there are also implications for AWE development. First, greater transparency is needed from the developers and distributors of AWE systems. These systems should be designed to make institutional evaluation much easier by means of more powerful reporting and querying that facilitate analysis of system performance on a student-, class-, and institution-wide basis. Annotation tools such as those used here could be included to make it easier to perform in situ evaluations. Coupled with

functionality that allows under-performing error types to be toggled on and off, this could permit use of AWE to be fine-tuned to the needs of a particular program.

ETS and other AWE purveyors should also fund classroom-based studies of the effects of variable accuracy on uses and perceptions of their systems. They must also be more forthcoming about recall, reporting testing data for all micro-features as a check on product quality and for use as baseline or comparison data in outsider evaluations. In addition, they should gather and publicize data about how much variation in accuracy might be expected across contexts of use. Evaluation studies that assume ideal conditions are of little practical value for informed use of AWE as formative assessment, as are underperforming types of microfeature.

Finally, there is a need for design-oriented studies which take into account how textual and non-textual, qualitative features of feedback influence interaction with and use by students. A companion study to the present research is investigating these issues, but the field needs larger investigations conducted as part of AWE development projects so they can be explored while students engage in authentic writing tasks. Different forms of phrasing, highlighting, and metalinguistic labeling should be experimentally manipulated. These studies should take into account factors such as the L1 of users, since there is ample reason for believing different user populations will benefit more from feedback on different features, possibly rendered in different forms (Ferris, 2011). Such research should be informed by Multimedia Learning Theory (Mayer, 2002), which offers design principles for managing cognitive demands and has contributed to the design and study of online help options.

Conclusion

This study has looked at two key inferences in an interpretation/use argument for the use of AWE as a formative assessment tool in a specific context. It found some support for these

inferences but raised a number of questions that need to be addressed before definitive statements of support can be made. These results highlight the challenge of validating formative applications of AWE, but also, we hope, make evident the value of the argument-based approach in allowing disparate forms of evidence to be considered alongside each other in a systematic way.

Despite the lack of definitive answers, this study has made clearer the need for greater accountability from those who promote and market AWE tools for use as formative assessment. Assessment experts characterize this as a "low-stakes" application of AWE (Chapelle & Chung, 2010; Williamson, Xi, & Breyer, 2012; Weigle, 2013a), but for students and instructors, time, effort, and funding are limited resources, and decisions about where to invest them are not inconsequential. More work is needed not only to address concerns about the pedagogical value of AWE feedback but to ensure quality experiences for end users, particularly those working in a second or foreign language.

Notes

1. Criterion provides holistic scores for essays based on prompts in the system's built-in prompt library. Holistic scores can also be obtained using prompts developed by the instructor, but these prompts must share certain features with Criterion's built-in prompts and must be created using a special feature within the system.
2. In an early paper about Criterion, Burstein et al. (2003) refer to a 90% precision rate used to evaluate algorithms addressing bigram errors or confusable words. We adopt the standard mentioned in Quinlan et al. because, given recent findings regarding Criterion

feedback accuracy in classroom-based studies (e.g., Lavolette et al., 2015), we believe an 80% threshold to be more realistic.

3. Compound words emerged as a common error category in the frequency analysis, but a review of the raw data showed the vast majority of occurrences consisted of flagging of the word *cannot*, which Criterion analyzed as having been spelled as two words when this was not the case.
4. Krippendorff's alpha is a statistical measure of the agreement achieved among annotators when coding a set of units of analysis. The annotation tool used in this study employs Krippendorf's α because this metric was specifically designed for content analysis applications and can support any number of annotators and categories as well as various metrics of distance between categories (nominal, ordinal, interval, etc.) and incomplete coding data (i.e., when all coders have not coded the entire dataset). Krippendorff's α differs from Cronbach's α in that the latter is a correlation-based consistency index that standardizes annotators' values and measures only covariation. Both indices are reported here on the assumption that Cronbach's α will be more familiar to readers.
5. Unlike the other error types eliciting generic feedback, *Preposition Error* had lower mental effort ratings more in line with the error types providing specific feedback. This may be attributable to the error-correction item involving preposition error in Part 1, which as the error-correction results show, consisted of an item that proved to be easy for most participants.

References

Aleven, V., Stahl, E., Schworm, S., Fischer, F., & Wallace, R. (2003). Help seeking and help design in interactive learning environments. *Review of Educational Research, 73*(3), 277-320. doi: 10.3102/00346543073003277

Anthony, L. (2014). AntConc (Version 3.4.3) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/

Aryadoust, V. (2013). *Building a validity argument for a listening test of academic proficiency.* Newcastle, UK: Cambridge Scholars Publishing.

Attali, Y. (2004). Exploring the feedback and revision features of Criterion. Paper presented at the National Council on Measurement in Education Annual Meeting, San Diego, CA.

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice, 18*(1), 5-25. doi: 10.1080/0969594x.2010.513678

Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Burstein, J., Chodorow, M., & Leacock, C. (2004). Automated essay evaluation: The Criterion online writing service. *AI Magazine, 25*(3), 27-36.

Cárdenas-Claros, M. S., & Gruba, P. (2009). Help options in CALL: A systematic review. *CALICO Journal, 27*(1), 69-90.

Clauser, B. E., Kane, M. T., & Swanson, D. B. (2002). Validity issues for performance-based tests scored with computer-automated scoring systems. *Applied Measurement in Education*, *15*(4), 413–432.

Chan, A. Y. W. (2010). Toward a Taxonomy of Written Errors: Investigation into the Written Errors of Hong Kong Cantonese ESL Learners. *TESOL Quarterly*, *44*(2), 295-319.

Chapelle, C. (2011). Validity argument for language assessment: The framework is simple… *Language Testing, 19*(1), 19-27.

Chapelle, C. A., Cotos, E., & Lee, J. Y. (2015). Validity arguments for diagnostic assessment using automated writing evaluation. *Language Testing.* Advance online publication. doi: 10.1177/0265532214565386

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.) (2008). *Building a validity argument for the Test of English as a Foreign Language.* New York: Routledge.

Chen, C., & Cheng, W. (2008). Beyond the design of automated writing evaluation: Pedagogical practices and perceived learning effectiveness in EFL writing classes, *Language Learning & Technology, 12*(2), 94-112.

Chodorow, M., Tetreault, J. R., & Han, N.-R. (2007). *Detection of grammatical errors involving prepositions*, Proceedinngs of the 4th ACL-SIGSEM Workshop on Prepositions, 25-30.

Chodorow, M., Gamon, M., & Tetreault, J. (2010). The utility of article and preposition error correction systems for English language learners: Feedback and assessment. *Language Testing*, *27*(3), 419–436.

Dikli, S. & Bleyle, S. (2014). Automated essay scoring feedback for second language writers: How does it compare to instructor feedback? *Assessing Writing, 22,* 1-7.

Doe, C. (2015). Student interpretations of diagnostic feedback. *Language Assessment Quarterly, 12*(1), 110-135. doi: 10.1080/15434303.2014.1002925

Ebyary, K. E., & Windeatt, S. (2010). The impact of computer-based feedback on students' written work. *International Journal of English Studies, 10*(2), 121-142.

Ferris, D. R. (2011). *Treatment of error in second language student writing* (2nd ed.). Ann Arbor, MI: The University of Michigan Press.

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication, 32*(4), 365-387. doi: 10.2307/356600

Friedl, J. E. (2006). *Mastering regular expressions: Powerful techniques for Perl and other tools.* Sebastopol, CA: O'Reilly Media.

Han, N.-R., Chodorow, M., & Leacock, C. (2006). Detecting errors in English article usage by non-native speakers. *Natural Language Engineering, 12*(02), 115-129. doi: doi:10.1017/S1351324906004190

Hayes, A.F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, *1*(1), 77-89.

Kane, M. (1992). An argument-based approach to validation. *Psychological Bulletin, 112,* 527–535.

Kane, M. T. (2012). Validating score interpretations and uses: Messick Lecture, Language Testing Research Colloquium, Cambridge, April 2010. *Language Testing, 29*(1), 3-17.

Kane, M. T. (2013). Validating the interpretation and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73.

Kane, M. T. (2015). Explicating validity. *Assessment in Education: Principles, Policy & Practice*, 1-14. doi: 10.1080/0969594x.2015.1060192

Lavolette, E., Polio, C., & Kahng, J. (2015). The accuracy of computer-assisted feedback and students' responses to it. *Language Learning & Technology, 19*(2), 50–68. Retrieved from http://llt.msu.edu/issues/june2015/lavolettepoliokahng.pdf

Leacock, C., Chodorow, M., Gamon, M., & Tetreault, J. (2010). Automated grammatical error detection for language learners. In E. Hirst, *Synthesis Lectures on Human Language Technologies*, San Rafael, CA: Morgan & Claypool Publishers.

Lee, H., Li, J., & Hegelheimer, V. (2012, September). The impact of Criterion on error reduction: A longitudinal study. Paper presented at TSLL conference, Ames, Iowa.

Li, J., Link, S., & Hegelheimer, V. (2015). Rethinking the role of automated writing evaluation (AWE) feedback in ESL writing instruction. *Journal of Second Language Writing, 27*, 1-18. doi: http://dx.doi.org/10.1016/j.jslw.2014.10.004

Manchón, R. M. (Ed.). (2011). *Learning-to-write and writing-to-learn in an additional language*. Amsterdam, the Netherlands: John Benjamins.

Mayer, R. E. (2002). Multimedia learning. In H. R. Brian (Ed.), *Psychology of Learning and Motivation* (Vol. 41, pp. 85-139): Academic Press.

Myers, M. (2003). What can computers and AES contribute to a K-12 writing program? In M. D. Shermis & J. Burstein (Eds.). *Automated essay scoring: A cross-disciplinary perspective* (pp. 3–20). Mahwah, NJ: Lawrence Erlbaum Associates.

Paas, F., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. M. (2003). Cognitive load measurement as a means to advance Cognitive Load Theory. *Educational Psychologist*, *38*(1), 63-71. doi: 10.1207/s15326985ep3801_8

Qualtrics Inc. (2015). Qualtrics survey software. www.qualtrics.com

Quinlan, T., Higgins, D., & Wolff, S. (2009). Evaluating the construct-coverage of the e-rater scoring engine. *ETS Research Report Series*, i-35. doi: 10.1002/j.2333-8504.2009.tb02158.x.

Schmidt, R. W. (1994). Deconstructing consciousness in search of useful definitions for applied linguistics. In J. Hulstijn & R. W. Schmidt (Eds.), *Consciousness in second language learning: AILA Review 11* (pp. 11-26).

Ware, P. (2011). Computer-generated feedback on student writing. *TESOL Quarterly, 45*(4), 769-774. doi: 10.5054/tq.2011.272525

Warschauer, M., & Grimes, D. (2008). Automated writing assessment in the classroom. *Pedagogies: An International Journal, 3*(1), 22-36. doi: 10.1080/15544800701771580

Weigle, S. C. (2013a). English as a second language writing and automated essay evaluation. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluations: Current applications and new directions* (pp. 36-54). New York, NY: Routledge.

Weigle, S. C. (2013b). English language learners and automated scoring of essays: Critical considerations. *Assessing Writing, 18,* 85-99.

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice, 31*(1), 2-13. doi: 10.1111/j.1745-3992.2011.00223.x

Table 1. Precision and Recall Data for Criterion Micro-Features Found in Published Studies

Error Type	Precision (%)	Recall (%)
Subject-verb Agreement	92	-
Articles	91	37
Preposition	78	18
Possessive Marker	95	-
Confusable Word	71	70

Sources: Burstein, Chodorow, and Leacock (2004) and Chodorow, Gamon, and Tetreault (2010).

Table 2. Frequency of Error Types Identified by Criterion in the Fall 2013 Corpus of Student Writing

Error Type	Frequency	Category
1. Repetition of Words	3997	S
2. Missing or Extra Article	**3456**	**U**
3. Spelling	2556	M
4. Missing Comma	**2031**	**M**
5. Preposition Error	**1704**	**U**
6. Fragments	**1567**	**G**
7. Subject-Verb Agreement	**1478**	**G**
8. Extra Comma	**1259**	**M**
9. Ill-formed Verbs	**1235**	**G**
10. Determiner Noun Agreement	**1218**	**U**
11. Run-on Sentences	**1151**	**G**
12. Passive Voice	1128	S
13. Compound Words	1042	M
14. Confused Words	**974**	**U**
15. Proofread This!	759	G
16. Missing Initial Capital Letter in a Sentence	692	M
17. Possessive Errors	599	G
18. Short Sentences	447	S
19. Missing Final Punctuation	392	M
20. Long Sentences	365	S
21. Garbled Sentences	365	G
22. Wrong Article	335	U
23. Capitalize Proper Nouns	279	M
24. Wrong Form of Word	188	U
25. Missing Question Mark	168	M
26. Hyphen Error	147	M
27. Sentences Beginning with Coordinating Conjunctions	145	S
28. Duplicates	94	M
29. Missing Apostrophe	88	M
30. Wrong Part of Speech	79	U
31. Wrong or Missing Word	60	G
32. Pronoun Errors	53	G
33. Faulty Comparisons	20	U
34. Nonstandard Word Form	7	U
35. Fused Words	7	M

36. Negation Error	1	U
37. Inappropriate Word or Phrases	0	S

Note: Bolded items were included in the analyses; *G* = grammar; *M* = mechanics; *S* = Style; *U* = usage

Table 3. Accuracy Ratings for 10 Error Types Most Commonly Identified by Criterion in the *Study*

Error category	*n*	Completely accurate	Partially accurate	Not accurate	% Completely + partially accurate	% Completely accurate
Extra Comma	70	36	4	30	57.1	51.4
Run-On Sentences	70	44	1	25	64.3	62.9
Confused Words	70	42	4	24	65.7	60.0
Missing Comma	70	45	4	21	70.0	64.3
Missing or Extra Article	70	41	9	20	71.4	58.6
Preposition Error	70	46	15	9	87.1	65.7
Fragment	70	61		9	87.1	87.1
Determiner-Noun Agreement	70	60	2	8	88.6	85.7
Subject-Verb Agreement	70	62	1	7	90.0	88.6
Ill-formed Verbs	70	61	6	3	95.7	87.1
Mean Percentage					77.7	71.1

Table 4. Average Scores on Error-Correction Task According To Error Type

Error Type	Part 1		Part 2	
	M	*SD*	*M*	*SD*
Confused Words	1.61	.78	1.95	.31
Determiner-noun Agreement	.22	.59	.35	.73
Extra Comma	1.96	.25	1.63	.78
Fragment	.54	.71	.71	.94
Ill-formed Verb	.10	.43	1.45	.89
Missing Comma	1.93	.38	1.46	.89
Missing or Extra Article	1.94	.33	1.78	.63
Preposition Error	1.71	.71	1.17	.99
Run-on Sentence	.30	.64	.79	.91
Subject-verb Agreement	1.35	.93	1.51	.86

Note: Maximum possible score = 2 based on the scale 0 for *not correct*; 1 for *partially correct*, and 2 for *fully correct.*

Figure 1. Components of Criterion feedback.

Figure 2. Screenshot of the annotation tool interface displaying an error from the subcorpus, with position of Criterion's highlighting denoted by "<<<."

Figure 3. Simulated Criterion feedback and pre-populated textbox in the error-correction task.

Figure 4. Mental effort ratings and standard errors by error type, with gray indicating specific feedback and white indicating generic feedback; scale is 1 = very little mental effort; 7 = a lot of mental effort.