2015

Using the Developmental Path of Cause to Bridge the Gap between AWE Scores and Writing Teachers' Evaluations

Hong Ma
Iowa State University

Tammy Slater
Iowa State University, tslater@iastate.edu

Using the Developmental Path of Cause to Bridge the Gap between AWE Scores and Writing Teachers' Evaluations

Abstract
Supported by artificial intelligence (AI), the most advanced Automatic Writing Evaluation (AWE) systems have gained increasing attention for their ability to provide immediate scoring and formative feedback, yet teachers have been hesitant to implement them into their classes because correlations between the grades they assign and the AWE scores have generally been low. This begs the question of where improvements in evaluation may need to be made, and what approaches are available to carry out this improvement.
This mixed-method study involved 59 cause and effect essays collected from English language learners enrolled in six different sections of a college level academic writing course and utilized theory proposed by Slater and Mohan (2010) regarding the developmental path of cause. The study compared the results of raters who used this developmental path with the accuracy of AWE scores produced by Criterion, an AWE tool developed by Educational Testing Service (ETS), and the grades reported by teachers.
Findings suggested that if Criterion is to be used successfully in the classroom, writing teachers need to take a meaning-based approach to their assessment, which would allow them and their students to understand more fully how language constructs cause and effect. Using the developmental path of cause as an analytical framework for assessment may then help teachers assign grades that are more in sync with AWE scores, which in turn can help students gain more trust in the scores they receive from both their teachers and Criterion.

Disciplines
Educational Assessment, Evaluation, and Research | Language and Literacy Education

Comments
This is a manuscript of an article published as Ma, Hong, and Tammy Slater. "Using the Developmental Path of Cause to Bridge the Gap between AWE Scores and Writing Teachers' Evaluations." Writing & Pedagogy 7 (2015). 10.1558/wap.v7i2-3.26376 Posted with permission.

Using the developmental path of cause to bridge the gap between AWE scores and writing teachers' evaluations

Hong Ma, Tammy Slater

Abstract

Supported by artificial intelligence (AI), the most advanced Automatic Writing Evaluation (AWE) systems have gained increasing attention for their ability to provide immediate scoring and formative feedback, yet teachers have been hesitant to implement them in their classes because correlations between the grades teachers assign and AWE scores have generally been low. This begs the question of where improvements in evaluation may need to be made, and what approaches are available to carry out these improvements.

This mixed-method study involved 58 cause and effect essays collected from English language learners enrolled in seven different sections of a college level academic writing course and utilized a theory proposed by Slater and Mohan (2010) regarding the developmental path of cause. The study compared the results of raters who used this developmental path with the AWE scores produced by *Criterion*, an AWE tool developed by Educational Testing Service (ETS), and the grades reported by teachers.

Findings suggest that if *Criterion* is to be used successfully in the classroom, writing teachers need to take a meaning-based approach to their assessment, which would allow them and their students to understand more fully how language constructs cause and effect. Using the developmental path of cause as an analytical framework for assessment may then help teachers assign grades that are more in sync with AWE scores, which in turn can help students gain more trust in the scores they receive from both their teachers and *Criterion*.

KEY WORDS: [PLEASE INSERT KEYWORDS HERE]

Introduction

English writing has been widely recognized as vital for academic success, effective for academic language development and valuable for mastering subject matter (Wang, Shang, & Briody, 2013; Warschauer, 2010). Although through the years there has been a preference for assessing students' writing performance directly through essay writing (Attali, Bridgeman, & Trapani, 2010), this emphasis on performance-based testing requires a significant time-consuming effort (Attali et al., 2010; Burstein, Chodorow & Leacock, 2003). Automated Writing Evaluation (AWE) systems have been created with hopes of reducing the enormous workload of essay evaluation.

As early as the 1960s, the first pioneering automated essay scoring system, Project Essay Grade (PEG), was created (Page, 2003; Page & Peterson, 1995; for a brief overview of automated essay scoring, see Wang & Brown, 2007). Score generation of PEG used surface quantifiable features such as essay length, sentence length, and word length to generate scores, and relied on feature weights obtained through multiple regression. Supported by artificial intelligence (AI), the mot advanced Automatic Writing Evaluation (AWE) systems, such as *Criterion* by the Educational Testing Service (ETS)

and *My Access!* by Vantage Learning, have gained increasing attention for their ability to provide immediate scoring and formative feedback (Chen & Cheng, 2008).

Although AWE systems differ in terms of specific approaches for generating scores, they all rely on statistical and linguistic methods to identify relevant language in examinees' essay responses and then predict scores by establishing the connection between these relevant linguistic features and human raters' grading (Chapelle & Chung, 2010). Scores provided by these AWE systems are claimed to be immune to factors that cause inconsistency in human grading, such as "fatigue, halo, hand writing, length effects and the effects of specific content" (Ben-Simon & Bennett, 2007, p. 4). AWE scores have been frequently used in testing contexts, mainly because interrater human-machine agreement is comparable to interrater performance in these contexts (Keith, 2003), where the interrater human performance has been shown to be used as "the gold standard" against which human-system agreement is compared (Burstein et al., 2003). Table 1, adapted from Ben-Simon and Bennett (2007), provides a brief overview of interrater human performance and human-machine agreement of different AWE tools in testing contexts.

AWE program	Scoring system	Empirical studies	Writing tests	Sample size	Interrater human performance	Human-machine agreement
PEG	PEG	Petersen, 1997	GRE	497	.75	.74-.75
MY Access!	IntelliMetric	Elliot, 2001	K-12 norm-referenced test	102	.84	.78-.85
Criterion	E-rater	Burstein et al., 1998	GMAT	500-1,000 per prompt	.82-.89	.79-.87
Criterion	E-rater	Burstein & Chodorow,1999	TWE	270	.69	.75
Criterion	E-rater	Attali et al., 2010	GRE & TWE	GRE: 3,000 essays TWE: 205,566	.70-.79	.72-.80

Table 1: *Interrater Human Performance and Human-machine Agreement in Testing Context*

This psychometric approach to validation has also been introduced into the classroom context, where the correlation between AWE scores and instructor grades is

employed as an important index underlying the appropriate pedagogical use of AWE. Since human-machine agreement in the classroom context has generally been very low, instructors have questioned the accuracy of AWE scores and have hesitated to implement AWE sores into classroom writing evaluation (Ebyary & Windeatt, 2010; James, 2006; Li, Link, Ma, Yang, & Hegelheimer, 2014; Wang & Brown, 2007). Therefore, a pressing question is how AWE systems, including AWE scores, can be used to achieve more desirable learning outcomes, and in particular how AWE as a pedagogical tool can be brought into the L2 writing classroom in ways that aim to "strike a balance between form and meaning" (Chen & Cheng, 2008, p. 108). Different from the previous literature that has investigated how AWE scores were implemented in classes and how students and instructors perceived the use of AWE scores, this study proposes an approach to validating/ justifying AWE score use in classrooms by drawing upon theoretical form-meaning connections proposed by Slater and Mohan (2010), with what they refer to as the developmental path of cause. Before we describe the developmental path of cause, we will briefly review the literature on instructor-machine agreement and AWE score use in the classroom context.

Human-machine Agreement and AWE Score Use in Classrooms

Compared to the high correlation values consistently reported in testing contexts (shown in Table 1), human-machine correlation in classroom-based studies has generally been low and has tended to vary considerably (see, for example, Chen & Cheng, 2008; Ebyary & Windeatt, 2010; Grimes & Warschauer, 2010; James, 2006; Li et al., 2014; Wang & Brown, 2007). Given the scarcity of classroom-based studies on automated scores, instructor-machine agreement has more frequently been reported as a by-product of research on other topics. For instance, Wang and Brown (2007) tested the null hypothesis that there was no significant difference between group mean scores generated by IntelliMetric™ and those grades assigned by human raters. The participants were 107 native English-speaking students from a Hispanic-serving institution in South Texas, taking the highest level of a Developmental English Writing course. All participants took the WritePlacer *Plus*, a standardized test that measures entry-level college students' writing skills, and produced writing samples in response to a prompt eliciting a persuasive essay. Scores generated by IntelliMetric™ range from 2 to 8. Students' essays produced for the Texas version of WritePlacer *Plus* were scored by IntelliMetric™ and and two faculty members. As summarized above, a correlational analysis using Spearman Rank Correlation Coefficient suggested that the correlation between IntelliMetricTM overall holistic scores and faculty human raters' overall holistic scores was very low (r_s = .11, $p < .017$). Such discrepancy between the AWE scores and instructor grades can cause problems for students and their teachers in classroom contexts, especially when students' perceptions of score reliability favors one form of assessment over another (Link et al., 2014).

Grimes and Warschauer (2010), following a naturalistic classroom-based approach, conducted a longitudinal and large-scale study on how the AWE program *MY Access!* was used in eight middle schools in Southern California over a three-year period. Observations, interviews, and a survey were employed to collect data on the classroom use of the program, and teachers and students' attitudes towards the program. Like Chen and Cheng (2008), Grimes and Warschauer also reported using AWE scores for both formative and partial summative purposes. Although automated scoring was perceived as unreliable, teachers still encouraged students to score higher. The reliance on automated scores to determine students' final grades varied among teachers from 0% to 90% with an average of only 18% of students' grades being determined by AWE scores. Students

reported that although they considered AWE scores ungrounded, the immediacy of score provision motivated them to focus more on writing, but they still took teachers' grades more seriously.

Such findings hint towards using AWE as only one form of feedback to students. For example, although the primary purpose of Ebyary and Windeatt's (2010) research was to investigate effects of *Criterion* feedback on language learners' L2 writing, instructor-machine agreement was reported, and the researchers suggested a baseline for appropriate pedagogical use of *Criterion* scores. Among the 31 instructors and 549 Egyptian potential EFL teachers who filled out the pretreatment questionnaire, two instructors and 24 volunteer students participated in the treatment with *Criterion*. The students were required to write about four topics over eight weeks and submit two drafts (an initial draft and a revised draft) for each topic. The inter-rater reliability between the two instructors and *Criterion* was moderate (r =.624 with the first rater, and r =.499 with the second rater), when both drafts of the first assignment were included. When the first submission for their first essay and the second submission of the fourth assignment were considered, the first rater and *Criterion* scores correlated significantly (r =.839). However, only moderate inter-rater reliability between the second rater and *Criterion* was found (r =.539). Given the generally moderate level of agreement between *Criterion* holistic scores and those provided by trained professional readers, the researchers suggested that *Criterion* scores "should be used as just one piece of evidence about the quality of students' writing" (p. 137), a sentiment echoed by other authors, such as Wang and Brown (2007), Lai (2010), Link, Durson, Karakaya, and Hegelheimer (2014), and Li et al (2014). These findings again call for a pedagogical approach that aims to bring the levels of agreement between AWE tools and instructors closer together, which may in turn raise the level of reliability on AWE scores and improve students' and teachers' perceptions of AWE use.

In sum, the low correlation between AWE scores and instructors' grades has raised issues regarding the implementation of AWE scores in pedagogical practice. It is certainly problematic to justify the inclusion of AWE scores for a high percentage of students' grades if the correlation between AWE scores and instructor grades is low. Despite providing both a psychometric approach and a naturalistic classroom-based approach to the question of AWE use in classroom settings, Li et al. (2014) still struggled to identify how low instructor-*Criterion* agreement can justify different uses of these scores by instructors, especially if they are using *Criterion* scores for summative purposes. With all these findings in mind, the current study aims to shed light on specific ways AWE score use in classrooms by introducing the developmental path of cause as a theoretical framework and seeing if and how its use reduce the disconnect between AWE scores and teachers' perceptions. The next section will provide a brief explanation of the developmental path of cause and the functional theory of language on which it is based.

The Developmental Path of Cause and a Functional Theory of Language

The developmental path of cause, initially proposed at a conference by Mohan, Slater, Luo, and Jaipal (2002) to illustrate the findings of a corpus-based causal discourse analysis of two encyclopedias targeted for different age and education levels, and later described in more detail in Slater and Mohan (2010), arranges linguistic features typical in causal discourse into hierarchical order and "supports the validity of judgments that rate one performance of causal discourse over another" from a systemic functional linguistics (SFL) perspective (Slater & Mohan, 2010, p. 261). The SFL framework, on which the developmental path of cause is based, has the potential to resolve the dilemma

in contemporary language assessment between assessing content and assessing language simultaneously (Mohan, Leung, & Slater, 2010). In contrast to a view of language that emphasizes accuracy in terms of form and structure, a functional view sees language (the wording of a discourse) as central to the construction of content (the meaning of a discourse) (Mohan & Slater, 2005). As language is considered to be the primary evidence for assessing an individual's knowledge (Mohan et al., 2010), SFL has the potential to provide an integrated assessment of language and content.

The SFL framework offers two complementary and interconnected approaches to the assessment of language and content: a genre approach and a register approach. The former relies on the analysis of prominent genres in education and how they are constructed and ordered. Veel (1997), for example, described the genres of science, arguing that there is a progression of genres that is generally followed in school science to help students learn science knowledge. To illustrate, he suggested that in moving from the genre of recounting procedures to explaining, students learn how to move from a here-and-now context to one that is more theoretical and abstract, a move that is a critical part of knowledge construction in science. Coffin (1997) argued the existence of a similar pathway for school history. These different genres use various language features that can be used to assess students' ability to construct them.

A register approach targets the meaning-wording relation directly through the analysis of what SFL refers to as ideational meaning (Mohan et al., 2010). From an SFL perspective, three variables of a discourse determine the use of language: field, tenor, and mode, which make up the register of the situation. These variables are associated with three main areas of meaning in language respectively: "ideational meaning, the resources for representing our experience of the world; interpersonal meaning, the resources for enabling interaction; and textual meaning, the resources for constructing coherent and connected texts" (Mohan & Slater, 2005, p. 156). Although all three meanings coexist in discourse, ideational meaning is a useful target for analyzing wording-meaning relationships in academic discourse, since it is closest to the everyday sense of content (Halliday, 1994). Moreover, within content/ideational meaning, the expression of causality has been considered fundamental to logical and scientific thought (Painter, 1999). Halliday and Martin (1993) in fact argued that "the language of science has become the language of literacy" (p. 11). With this in mind, the wording of causal relations within a genre offers a fruitful area of investigation for assessing academic writing.

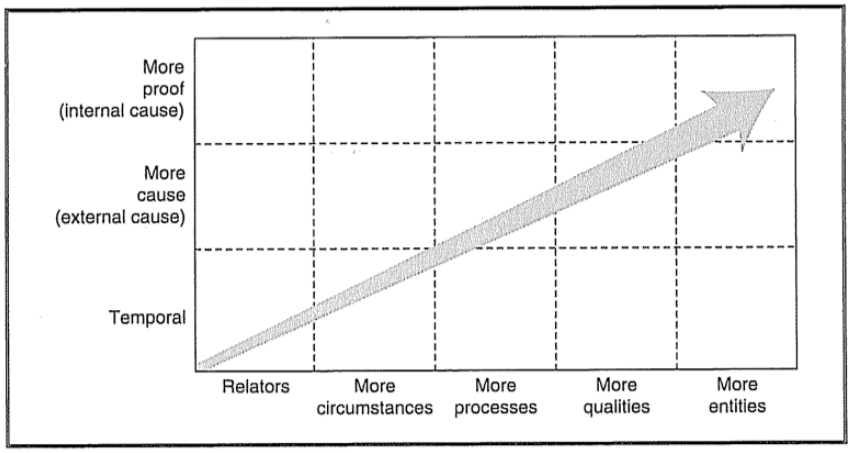

Figure 1: The developmental path of cause (Slater & Mohan, 2010)

From this SFL approach, the developmental path of cause (shown in Figure 1) captures the phenomenon that "causal language develops along two dimensions: a

lexicogrammatical dimension (the horizontal axis) and a semantic dimension (the vertical axis)" (Slater & Mohan, 2010, p. 261). Reflecting the "general drift" of grammatical metaphor' put forth by Halliday (1998), the horizontal axis of the model suggests that the sophistication of causal language increases by moving away from relators (conjunctions) towards more grammatically metaphoric constructions, including circumstances, processes, qualities, and entities (Slater & Mohan, 2010). The vertical axis of the model illustrates the semantic dimension of causal language that moves from time through cause to proof, reflecting the evolution that scientific thought has made throughout history to represent and explain physical phenomena (see Halliday & Martin, 1993, p. 66). Further support for the theoretical model of the developmental path of cause was offered by Slater (2004), who investigated the oral causal explanations in science of ESL and non-ESL students in primary and high school and found that students with higher English proficiency levels tended to produce causal language that generally skewed more towards proof and entities than their counterparts with lower English proficiency levels, as shown in Table 2.

	primary vs. high school (native)		ESL vs. non-ESL	
Linguistic feature	primary	high school	ESL	non-ESL
External temporal conjunctions *(indicating real time or sequence)*	25.35	51.11	29.68	51.11
External causal conjunctions *(indicating real causal relations)*	29.11	12.81	30.53	12.81
Internal conjunctions *(indicating textual organization)*	0	0.28	0	0.28
Temporal circumstances	15.96	22.56	30.31	22.56
Causal circumstances	3.76	0.56	1.47	0.56
Temporal processes	0	1.39	0	1.39
Causal processes	1.88	6.41	4.21	6.41
Proof processes	0.94	0.7	0	0.7
Temporal entities	0	2.51	0	2.51
Causal entities	0.94	4.46	0	4.46
General metaphoric entities	0	16.99	11.37	16.99

Table 2: *Use of Causal Language: primary vs. high school, ESL vs. non-ESL* (Give citation for this table)

A functional approach to evaluating cause-effect discourse with a focus on field and ideational meaning can facilitate the domain definition of an argument-based approach to validity. According to the TOEFL validity argument framework proposed by Chapelle, Jamieson, and Enright (2008), a domain definition is essential for obtaining an observation of student performance, which in turn serves as the basis for generating an observed score. Domain definition (the starting point of validation) from a SFL perspective brings changes to the traditional evaluations of causal discourse. For both content teachers and language teachers who "assess the meaning of text on the basis of the wording," responsible evaluators "should be able to explain or justify their judgment of meaning of discourse by pointing to wording in the discourse that expresses that meaning" (Mohan & Slater, 2010, p. 227). We argue then that the developmental path of

cause can provide teachers with adequate criteria for evaluating students' causal discourse, thus supporting the validity of teachers' judgments. Therefore, with a goal to investigate how accurately instructors' grades and AWE scores reflect the quality of students' causal discourse by using the developmental path of cause, the following research questions were asked:

1. To what extent do *Criterion* scores and teacher scores on students' cause and effect essays correlate with scores generated according to 'the developmental path'?
2. How well do scores generated according to 'the developmental path' support experienced teachers' intuitive judgment on the quality of students' writing when both teacher and *Criterion* scores are not able to decide?

Methodology

This study adopts an explanatory sequential design, a two-phase mixed methods design beginning with quantitative data collection and analysis and followed by qualitative data, which is collected and analyzed to "explain or build upon initial quantitative results" (Creswell & Plano Clark, 2007, p. 72). In this study, *Criterion* scores, instructor grades, and grades generated from an SFL perspective (referred to in this study as third party grades) were collected and analyzed quantitatively by calculating the correlation values between each pair. The results further informed the collection of qualitative data in the next phase. Since the third party grades agreed with *Criterion* scores more strongly than with instructor grades, essay pairs ranked identically by *Criterion* and the third party but differently by the class instructors were selected intentionally to see whether decisions made by *Criterion* were supported by the intuitive judgment of experienced ESL writing teachers. The rest of this section will cover information about the design of this study, including participants, context, materials, and procedures.

Participants

The participants in this study included 58 college-level international students, their five ESL (English as a second language) writing instructors, three additional experienced ESL writing teachers, and three trained SFL coders. Except for the international students, all other participants were graduate students in the English Department. Fifty-eight essay responses were collected from the international students enrolled in seven sections of an introductory academic writing course Fall 2014, a course designed to develop the abilities of undergraduate ESL students whose writing samples demonstrate grammatical errors that do not impede comprehension, but who require further work on organization before they can register for the standard first-year composition courses.

The five course instructors (two males and three females) included two native speakers of English, one Turkish, one Korean, and one Chinese and their experience teaching this class varied from one semester to three years. These instructors graded their own students' writing using the rubric for the current study (see Table 3). Three additional experienced ESL writing teachers attended a focus group interview; these were two female American native speakers of English and one female Chinese. All had three to four years of teaching experience. The three trained SFL coders were all female—one Turkish, one Vietnamese, and one Chinese—and coded students' essays using the developmental path of cause.

Study Context

The study took place in an ESL composition class provided by the English department of a university located in the American mid-west. Students majoring in various fields are in the course to improve their general writing proficiency without reference to the specific requirements of their own fields. With a focus on preparing students for mainstream course writing assignments, this academic composition course covers a range of different genres, including expository paragraphs, classifications, critiques, evaluations, and consequential essays. The course textbook, *Engaging Writing 2, 2nd Edition* (Fitzpatrick, 2011), presents each genre, focusing on specific topics. The unit on the cause-effect essay, for example, contains reading materials concentrating on causes and/or effects of specific economic phenomena. Following the combination of genre and content knowledge, the major assignment required students to search for resources and discuss one of the following or related issues: (1) the effects of globalization on a country, region, or city; (2) the reason why a country has a strong, weak, or variable economy; and (3) the effects of a specific event on economy.

Scaffolding activities introduced both composition strategies and linguistic features typical of cause and effect. Composition strategies included different ways of writing an introduction to a formal essay and the organization of body paragraphs based on temporal, sequential, or causal order. As to linguistic features, nouns (e.g., cause, reason, factor, result, effect), verbs (e.g., cause, result in, lead to, affect) and conjunctions (e.g., because, so, therefore) indicating causality were introduced. In addition, this course adopted a process-based approach, where each writing assignment required multiple drafts, and students received formative feedback from both their instructors and the AWE system *Criterion* between drafts. During this process, students were able to receive scores generated by *Criterion* immediately after each submission.

Materials

Materials in this study included *Criterion*, the writing prompt stimulating students' essay responses, and the rubric used for evaluation. The following sections will provide detailed information for each material used.

The AWE System Criterion

The AWE system used in this study, *Criterion,* is widely used and has been purchased by many institutions, including elementary, middle and high schools, universities, and military institutions in the US; it has also moved to EFL contexts such as China, Taiwan, and Japan (Burstein et al., 2003). This web-based service developed by Educational Testing Service (ETS) is able to provide submitted essays with immediate diagnostic feedback and a holistic score. These two functions are realized through two complementary applications relying on natural language processing (NLP): the feedback provision application, Critique, and the scoring application, E-rater. The latest version of Critique used in this study includes several programs that evaluate and provide feedback along five dimensions: (1) grammar, usage, mechanics, and style; (2) organization and development; (3) topical analysis (i.e., prompt-specific vocabulary); (4) word complexity; and (5) essay length (Attali & Burstein, 2005). Although the five dimensions are not directly related to scoring, they indicate features that designers of *Criterion* intend to evaluate (Ben-Simon & Bennett, 2007). As to automated scoring, E-rater generates a holistic score by extracting from an essay linguistic features that reflect characteristics covered in the scoring guide and determining the weight of these features in the overall writing quality using a statistical model. The *Criterion* holistic score is presented on a six-point scale, with a higher score indicating a higher quality of essay. In addition, a score

description (scoring guide) is provided for each score level to function as general formative information.

The Writing Prompt

To obtain *Criterion* scores, only prompts included in the system can be selected. The TOEFL level topic requiring students to explain reasons for people attending college was selected as the prompt for this study due to its appropriate difficulty level for the target participants and the causal discourse it intends to elicit. To draw students' attention to causality, questions were added to the prompt for students (see Appendix 1). Students were required to produce 250 to 300 words in 30 minutes, the same time set for the TOEFL iBT (https://www.ets.org/Media/Tests/TOEFL/pdf/TOEFL_at_a_Glance.pdf).

The Grading Rubric

The rubrics used in this course for grading major writing assignments were topic-specific. Although one of the assignments was a cause-effect essay, the writing prompt used in this study was not part of the curriculum. Therefore, a specific rubric was created by adapting the existing rubric. As shown in Table 3, the grading rubric used for this study differed from the course rubric mainly in its requirement for a thesis statement (context) and citation (style). The course rubric required an extended introduction and an accurate citation of the article assigned. The study's rubric, due to a more limited number of words and the intention to elicit causal expressions, required only a brief introduction. In addition, this timed writing task was based on a prompt rather than stimulated by reading, eliminating the need for citations. Both rubrics required instructors to grade holistically with a maximum score of fifty.

	Rubric for the current study	Rubric shared by sections of the writing course
context	A brief introduction sets the context. (pp. 90-91)	Full introduction sets the context (time period, people, place) and introduces the major factors involved. (pp. 90-91)
	A thesis states the reasons for attending college.	A thesis states the causes and effects of the phenomenon discussed in the essay.
substance	Include an extended discussion of (1) what types of experiences/career preparation/knowledge/other ideas that attending a college or university can provide; (2) how attending a college or university can provide people with these experiences/ chances for career preparation/useful knowledge or other benefits (3) what are the possible benefits of obtaining these experiences/ career preparation/ useful knowledge/other ideas?	The original article is explained and developed fully with sufficient examples. • Includes an extended discussion of the points made in the original article, either in agreement or in disagreement.
	Unity of topic is maintained by eliminating unrelated material and keeping only connected ideas. (pp.	Unity of topic is maintained by eliminating unrelated material and keeping only connected ideas. (pp. 96-

	96-99)	99)
	A logical order is followed and cohesion is created – either time, sequence, or order of importance of the reasons. (pp.94) Extended commentary is integrated into the paragraph as a unified part of the whole discussion and conclusion.	A logical order is followed and cohesion is created – either time, sequence, or order of importance of the factors. (pp.94) Extended commentary is integrated into the paragraph as a unified part of the whole discussion and conclusion.
	Verb tense is correct and consistent. Cause and effect vocabulary structures are used. (pp.103-111) Problems with grammar and mechanics are minimal and do not distract the reader. Use required document formatting.	Verb tense is correct and consistent. Cause and effect vocabulary structures are used. (pp.103-111) Problems with grammar and mechanics are minimal and do not distract the reader. Use required document formatting. Provides an accurate APA or MLA citation of the article.

Table 3: *Comparison between Rubrics*

Procedure

Data collection and analysis of the study can be divided into three steps: obtaining *Criterion* scores and teacher grades, obtaining the third party grades, and carrying out a focus group interview.

Obtaining Criterion scores & teacher grades

Students' essays were collected in three ways according to the instructors' teaching schedule, and all students had 30 minutes to write, as in the TOEFL iBT. Three instructors had students complete the task during lab classes, where computers were available. Students from the other two sections wrote their essays as an assignment after class. Students from these five sections composed their essays with access to the *Criterion* interface and submitted their essays directly to *Criterion*. All instructors posted the writing prompt on their course webpages and had students read the prompt before composing their essays using *Criterion*. After the students submitted their writing to *Criterion*, the instructors downloaded their essays for grading. One instructor had her students write with pen and paper during class, and thus for these students, a hard copy of the writing prompt was provided, the essays were later retyped using a word processor, and the electronic version of these essays were submitted to *Criterion* and sent to the instructor for scores. The two different ways of composing essays, using computers versus using paper and pencil, may have influenced students' performance (Taylor, Kirsch, Eignor, & Jamieson, 1999), since it has been reported that students writing with computers tend to produce work of "greater length and higher quality" (Goldberg, Russell, & Cook, 2003: 2). These differences were not considered influential for our study, though, as we were primarily interested in examining the texts themselves and comparing these across *Criterion*, instructor scoring, and SFL analysis.

Obtaining the third party grades

After receiving *Criterion* scores and instructor grades, essays were coded according to the developmental path of cause to generate the third party grades. To ensure accuracy of the coding process, three coders (one major coder and two additional coders) were employed. The major coder trained the other two regarding the linguistic features on which to focus, using a coding rubric (shown in Appendix 2). The coders were informed that they did not need to memorize the terminology or categorize the causal language. The purpose of the rubric was to remind coders what linguistic features they needed to identify from essays. Each coder analyzed three essays independently and compared their coding with the major coder to clarify their task and establish reliability.

After the training session, the three coders coded twenty essays randomly selected from the 59 essays, then discussed any discrepancies in their coding until an agreement was reached. When they could not decide, an expert in SFL was consulted. The major coder then continued coding the rest of the essays. For each causal expression appearing in students' essays, a score was granted following Table 2: The lowest level expressions, external temporal conjunctions, scored 0.5 each, and causal expressions categorized at each level higher corresponded with a 0.5-point increase. For causal expressions that were not used correctly, no point was granted since the expressions did not contribute to meaning making. Finally, the third-party grade of each essay based on the developmental path of cause was calculated by adding up these points.

Focus group interview

Three experienced English writing instructors who were not trained to use the developmental path of cause—two native English speakers and one Chinese—attended a focus group interview to judge intuitively the quality of three sets of essays (two papers in each set). These six essays (2*3) were selected because the essays in each set were ranked identically by *Criterion* and the third party but differently by the class instructors. All three teachers ranked these essays and provided justifications for their ranking by set. The researcher organized the focus group interview, posed questions to clarify teachers' explanations, and audiotaped the interview, which lasted about ninety minutes.

Data Analysis

To answer the first research question, we calculated the correlations between *Criterion* scores and instructor grades, between *Criterion* scores and the third party grades, and between instructor grades and the third party grades using the Pearson *r*. The teachers graded the essays holistically according to the rubric for this study as shown in Table 3 with a total score of fifty. The third party grades for each essay were calculated by adding up the points for all appropriately used occurrences of linguistic features. The following two short examples help to illustrate how these third party grades were calculated.

> Example 1: Education is the key tool that ***shapes*** and ***ameliorates*** our future. It responsible for ***making*** humans civilized and live in harmony. (Three causal processes = 10.5 points)
>
> Example 2: ***If*** people only study at home, they will not touch machines more than they study in university. Colleges have lots of professional machines for different majors. (One external causal conjunction = 1 point)

Explanations for the second research question relied on coding and analysis of teachers' intuitive judgments obtained from the focus group interview. Data analysis of this research question started with *in vivo* coding (Saldana, 2009), which enabled the researcher to note down all the related ideas emerging from the interviewees' speech. The coding eventually revealed six categories: grammatical accuracy, word choice and pronouns, accuracy of causal language, maintaining of correct genre, achieving development, convincing reasons, and different views toward structure by native and nonnative teachers. Finally, three themes—language accuracy, content, and structure—emerged as important aspects that the teachers attended to when grading students' cause-effect essays.

Results

The following sections describe our correlational analyses between the instructors' scores, *Criterion scores,* and third party scores as well as discussions of the teachers' intuitive judgments. The results provide information concerning how *Criterion* scores can connect to potential classroom use.

Correlational Analysis: Research question 1

Similar to the majority of previous empirical efforts at validating *Criterion* scores in classroom use, the correlation between *Criterion* scores and instructor grades was very low (r = .39). Replicating the reasoning that dominates research in this field, this result might suggest that *Criterion* scores need to be used with caution in classroom contexts. However, an examination of Table 4 reveals that the correlation between the third party grades and *Criterion* scores (r = .61) was remarkably higher than that between the third party grades and class instructor grades (r = .35). In other words, the grades based on SFL theory, calculated from the appropriate use of causal linguistic features in context, tended to support *Criterion* scores more strongly than it did for instructor grades. This suggests that teachers who become familiar with the SFL approach may provide their students with information about causal discourse that can improve their *Criterion* scores.

	The 3rd party	*Criterion*	Instructors
The 3rd party	1		
Criterion	0.61	1	
Instructors	0.35	0.39	1

Table 4: *Correlation Metrics, the 3rd Party, Criterion, & Instructors*

Given this result and its implications, the next section focuses on the three experienced ESL writing teachers' intuitive judgments and their justifications.

Teachers' Intuitive Judgments: Research Question 2

To explore how well the scores generated according to 'the developmental path' (i.e., the third party grades) support experienced teachers' intuitive judgments on the quality of students' writing when both teacher and *Criterion* scores are not able to decide, three sets of essays (two essays each set) were selected for examination and discussion. It is important to reiterate here that these teachers were not trained in the use of the developmental path of cause; nor did they have special training in SFL theory. The three

teachers ranked the problematic essays set by set and talked about the features they attended to when evaluating cause-effect essays. All three teachers ranked these essays in the same order as *Criterion* and the third party grades, but this ranking differed from that of the course instructors. The last set of essays caused a debate, and an agreement was not easily reached. The two native English-speaking teachers ranked the essays the same as *Criterion* and the third party raters, as they had with the first two sets of essays. The nonnative teacher, however, only agreed with the two native teachers when considering which essay provided adequate explanation of the reasons for attending college. When considering the essay's structure, though, she argued for a reverse order. The following sections present the teachers' justifications for their ranking from three interrelated aspects that affect essay quality: language, content, and structure.

Language accuracy/appropriateness

An important factor influencing the teachers' intuitive judgments was the effort they needed to make to comprehend the meaning students intended to express. Grammatical accuracy was not the teachers' primary concern when they evaluated these essays, unless too many errors impeded understanding.

> *It's not really good writing. So many grammatical errors that really affect the comprehension, including subject-verb agreement, run-on sentences, articles, conjunctions, different types of errors, missing comma.*
>
> *And then using the second essay as a comparison,… the few issues that were there in grammar and usage is very minor, you can still understand the student's meaning.*

Compared to grammatical errors, word choice and use of pronouns affected the teachers' evaluation more strongly, since inappropriate word choice prevented the teachers from understanding the intended meaning, and unclear pronouns had the teachers wondering what they referred to. Both issues obscured meaning and were associated with weak writing ability.

> *This is one sentence that has cause-effect, but it has two pronouns that have very unclear antecedent, so it makes it confusing what is the cause and what is the effect. (use of pronoun)*
>
> *Well, it is a pretty good cause effect sentence, but the fact again, the pronouns, not sure who they are referencing, really weaken the sentence and doesn't really give him credit. (use of pronoun)*
>
> *The second paragraph, it says "we can make a lot of friends in different areas," but what does it mean by "areas," geographic regions, majors. (word choice)*
>
> *Then "I made many local persons," I assume they meant "I made many local friends," but that's not clear either. "I believe social net," I assume they mean network. And they "the people everyone knows compose new around," so the student's English skill is far weaker than the other paper too. (word choice)*

Another important issue for the teachers was the accurate use of causal linguistic features. Although students made mechanical errors when using causal conjunctions, teachers were not very concerned with punctuation, since those errors did not impede understanding. Rather, if causal conjunctions did not construct appropriate logical relations, the teachers found those conjunctions misleading:

"Now this society need people being at multiple skills, so attending college is a satisfactory way to approach this need," that's not an appropriate way to use "so," that doesn't justify the claim that the society needs "multiple skills." So yes they are using cause effect language, but they are not using it appropriately.

Causal language & content

Connected to the language was content that the teachers focused on when ranking these essays. Teachers discussed content or meaning that students intended to express in terms of adherence to genre conventions, development of argument, and persuasiveness of reasons. All agreed that the appropriate use of causal language helped students maintain the correct genre, a cause-effect essay. Inferior essays used fewer causal markers, leaving the readers with the impression that these essays were more descriptive or argumentative rather than cause-effect essays. These raters admitted that cause-effect essays needed to include causes or effects and well-developed arguments; however, the absence of causal features where they were supposed to appear suggested that the writer lacked the ability to control the cause-effect genre:

The student's essay is very much descriptive. …This essay did not have the cause-effect. Actually, this is only one sentence that has strong cause and effect, where here he/she says, "If we have experience about those in college, we would do it better in the future."

This was really not a cause effect essay. It was borderline argumentative, but it wasn't cause effect. … "They are trying to assist their student to get success, so they will provide a lot of chance for students to get exercise to improve themselves." This is the only sentence in the whole essay trying to be cause effect.

Frequent occurrences of causal language or a chain of causal language facilitates development of a cause-effect essay. The teachers noticed that superior essays generally tended to use causal language to provide more insight into the relationships between the cause (reasons) and the effect (attending university) that students pointed out in their topic sentence(s).

The student was able to back up the effects of going to attending college very clearly. For example, there were three cause effect connections, where the student used words like "so that," "because," and the very last sentence, which gave a very detailed insight.

I also think Essay A is much better than Essay B. First of all, the cause effect signals, there are some relative clauses at the end of their paragraph, "students get working experiences through co-ops, which gives them…" This is pretty good.

Whether the reasons students provided in their essays were convincing was also likely to affect teachers' evaluation of a specific essay. This phenomenon reflected the claim from the SFL perspective that language is the evidence for assessing content. One of the teachers found it important for students to identify unique reasons for attending university.

I appreciated that the student made the claim that "the primary reason for everyone to attend college is to obtain education," and then in order to justify that claim, he/she provides universally believed reason. … It's not a justification people like to argue with, so it was powerful. (convincing reasons)

They didn't justify their claim "there is almost no job opportunity for those who don't attend college," not everyone believes that, because not everyone attends college. (convincing reasons)

"We continue to have habits for experiences for friends." That's suddenly true for people from different areas, or regions, it can be any friends, not just college ones. (uniqueness of reasons)

"First thing get more opportunity, get ideal job, have a better career development in the future," and then, in the rest of the paragraph, the student doesn't explain how the evidence that they give "is a unique feature of college life," so why is that the reason people go to college. It's not convincing that they couldn't get the same benefit in other ways. (uniqueness of reasons)

Causal language vs. structure

During the focus group discussion, the debate over which feature was more important in determining the quality of students' essays—the appropriate use of causal language or structure—contributed to the disagreement in the teachers' rankings of the two essays in the last set. Essay A, ranked higher by the two native English-speaking teachers, did not attempt to follow the conventional five-paragraph organization, while Essay B organized the essay in a conventional manner. When evaluating the general quality of the two essays, the nonnative speaker appeared to judge structure as being more important than causal language and had a strict view towards appropriate organization. For this essay, her comments showed she was expecting a clear thesis statement followed by reasons for attending university, with each paragraph covering more detailed explanation for each reason, and a straightforward topic sentence stating the reason being covered in each paragraph. Therefore, she ranked Essay B higher:

So when I grade students' papers, the structure is really important…I think this one is a little better than the next one. It has clear structure. This [the essay rated higher by native teachers] *is lower, because for the second paragraph, there is no topic sentence here at all.*

Although the two native English-speaking teachers also appreciated a clear thesis and topic sentences, they appeared to demonstrate a more flexible view towards the appropriate organization of a cause-effect essay. They appreciated the idea that Essay A focused on one reason for career preparation and devoted the whole essay to elaborating on this reason.

Actually I am not very sensitive to structure. I like the way the student approached this essay given the time limit. It is good to focus on one reason and provide more insight to it. It did not follow the conventional way of TOEFL writing.

These teachers admitted that the absence of clear topic sentences influenced the quality of this essay and that this essay was not a particularly good cause-effect essay. However, they still rated this essay higher because they felt it demonstrated the appropriate use of several causal language features, which led the reader smoothly

through the piece. In addition, one of these teachers pointed out that the strategies of comparing the effects of attending and not attending university that were used in Essay A was the best overall.

> *You don't know that this paragraph is going to be a cause effect paragraph, because the very beginning has no sense of the fact that the student is going to talk about all these benefits. So it would be helpful to have a strong topic sentence that clear…but did it in subtle ways for making cause and effect, which makes it stronger than the other essay, but not a strong cause effect essay over all in comparison of others.*
>
> *The student said, "if people only study at home, they will not touch machines more than they study in the university," This is excellent, this is giving the contrast why you need to go to university. Where the other paper had nothing even close to that.*

Overall, the qualitative data suggested that despite ***not*** being trained to look for the evidence of causal features, teachers' intuitive judgments generally focused on whether these were included, appropriate, and logical—feedback that is very useful for their students—and their views appeared to correspond with the *Criterion* scores that were assigned. These views contradicted the classroom instructors' scores, which were based on the course rubrics. This finding supports Mohan and Slater (2004), who observed that raters' intuitive evaluations of texts could at times be suppressed by the scoring rubrics being used. Further, the *Criterion* scores, in this case, successfully ranked one essay that did not follow a conventional organization.

Conclusion

This study was motivated by the general sense of usefulness of automated essay evaluation, by the low correlations identified in previous research between teachers' scores and scores generated by *Criterion,* and by the SFL view that wording and meaning are inherently connected. In essence, if students are learning to write cause-effect essays, it makes sense to hope that they will eventually be able to construct discourse that makes use of appropriate causal discourse features, and that their teachers as well as the AWE system that may be incorporated into classroom practice will agree that students have met the challenge well. It therefore seemed evident to explore the correlations that can result from an approach that examines both AWE use and an SFL approach. Thus, from this motivation, we set out to see whether the use of Slater and Mohan's developmental path of cause could lessen the gap between scores given in the classroom context and those assigned by *Criterion,* with regards to causal discourse.

While we agree that the instructors' input, the third-party raters' discussions, and *Criterion* feedback and scores all have much to offer the writing classroom, both qualitative and quantitative data in this study suggested that compared to the in-class instructors' grades assigned in this study, *Criterion* scores appear to be more strongly supported by SFL theory, suggesting that SFL theory has a place in teaching students how to write causal discourse. Not only were the third party grades—those determined from the developmental path of cause—more closely correlated to the *Criterion* scores, comments made by the teachers in the focus group revolved around the appropriate use of causal discourse by the students. And interestingly, their recommendations resembled the *Criterion* and third party raters more closely than the rubric-informed course instructors' scores. Thus insight from our findings suggests that the students' choices of causal discourse offer linguistic evidence of a developing ability to construct causal meanings

appropriately. *Criterion* seems to be picking up on this development, as are the third-party raters who were trained to use the developmental path of cause. The course instructors appear to be attending to something different, perhaps as a consequence of their interpretations of the scoring rubric. These findings are not unexpected since *Criterion* scores and the third party grades are highly language based, and the language complexity that the developmental path of cause targeted is also rated by *Criterion* using computational methods. The instructors, on the other hand, attended to a much broader scope of writing constructs and rated through a reconciliation of their intuitive impressions and the application of their rubrics (Lumley, 2002; Weigle, 2010). We argue from our findings that by raising instructors' awareness of the developmental path of cause, these teachers can help their students develop more sophisticated linguistic resources for constructing accurate and appropriate causal texts, and because texts assessed using this approach appear to be more closely matched by *Criterion's* formative and summative assessment, teachers' adoption of *Criterion* in the classroom may be better accepted by students.

While not intending to diminish the value of teacher ratings (especially with feedback) or advocate the blind adoption of *Criterion* scores, the results of this study suggest that *Criterion* could be used successfully in classrooms where writing teachers adopt a heightened awareness of the connections between wording and meaning in their assessment. In such a context, using the developmental path of cause as a theoretical framework for the assessment of causal essays may not only help instructors develop students' causal linguistic repertoires but may also result in grades that are more in sync with AWE scores, which in turn can help students gain more trust in the scores they receive from both their teachers and from *Criterion*. In other words, teaching students about the developmental path of cause can provide positive washback from both teachers and AWE systems. It has been reported that students were motivated to make revisions to increase their AWE scores largely through addressing linguistic accuracy (Chen & Cheng, 2008; Grimes & Warschauer, 2010). The connection between *Criterion* scores and grades based on the SFL theory established in this study goes beyond accuracy by potentially encouraging more revisions as a result of internalizing how appropriately constructed linguistic form expresses causal meanings. The underlying concept is that teachers using the developmental path of cause can help students understand and manipulate for higher scores from both teachers and the AWE system.

Certainly, the findings of this study should be interpreted with caution. In addition to the small number of essays analyzed, correlation values involving teachers' grades were highly dependent on a specific group of teachers. More research needs to be undertaken to confirm whether the correlation between third party grades and *Criterion* scores would still hold true for other teachers working in different contexts. But the use of the developmental path of cause is promising for this type of assessment.

Scores assigned from AWE programs offer reasonable consistency, and thus their use has the potential to not only promote learner autonomy, but free up time that can be spent on further writing development. But both students and teachers need to feel confident that the scores being assigned are similar and are highlighting the same aspects of writing development. For the writing of short cause-and-effect essays, focusing on a theoretical model which combines wording and meaning, such as the developmental path of cause, will not only bring the two authorities together, but will also expand students' linguistic resources for constructing causal discourse—in other words, the use of the path in formative assessment by AWE tools and course instructors carries good potential for helping students develop their academic language proficiency.

References

Attali, Y., & Burstein, J. (2005). *Automated essay scoring with e-rater version 2.0* (ETS RR-04-45). Princeton, NJ: Educational Testing Service.

Attali, Y., Bridgeman, B., & Trapani, C. (2010). Performance of a generic approach in automated essay scoring. *The Journal of Technology, Learning and Assessment*, *10*(3), 1-17. Retrieved from http://ejournals.bc.edu/ojs/index.php/jtla/article/view/1603/1455

Ben-Simon, A., & Bennett, E.R. (2007). Toward more substantively meaningful automated essay scoring. *The Journal of Technology, Learning and Assessment, 6*(1), Retrieved from Retrieved from http://ejournals.bc.edu/ojs/index.php/jtla/article/view/1631/1475

Burstein, J., & Chodorow, M. (1999, June). Automated essay scoring for nonnative English speakers. *In Proceedings of the ACL99 Workshop on Computer-Mediated Language Assessment and Evaluation of Natural Language Processing*. Retrieved from http://www. ets.org/Media/Research/pdf/erater_acl99rev.pdf

Burstein, J., Chodorow, M., & Leacock, C. (2003). Criterion online essay evaluation: An application for automated evaluation of student essays. Retrieved from http://www.ets.org/Media/Research/pdf/erater_iaai03_burstein.pdf

Burstein, J., Kukich, K., Wolff, S., Lu, C., & Chodorow, M. (1998). Computer analysis of essays. Retrieved from https://www.ets.org/Media/Research/pdf/erater_ncmefinal.pdf.

Chapelle, C.A. & Chung, Y. (2010). The promise of NLP and speech processing technologies in language assessment. *Language Testing, 27*(3), 301-315.

Chapelle, C.A., Jamieson, J. & Enright, M.K. (Eds.). (2008*). Building a validity argument for the test of English as a foreign language*. London: Routledge.

Chen, C.-F.E., & Cheng, W.-Y.E. (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. Retrieved from http://llt.msu.edu/vol12num2/chencheng.pdf

Coffin, C. (1997). Constructing and giving value to the past: An investigation into secondary school history. In F. Christie & J.R. Martin (Eds.), *Genre and Institutions: Social Processes in the Workplace and School* (pp. 196-230). London: Continuum.

Creswell, J.W., & Plano Clark, V.L. (2007). *Designing and conducting mixed methods research*. Thousand Oaks, CA: SAGE Publications.

Ebyary, K., & Windeatt, S. (2010). The impact of computer-based feedback on students' written work, *International Journal of English Studies, 10* (2), 121-142. Retrieved from http://revistas.um.es/ijes/article/view/119231/112351

Elliot, S. (2001). IntelliMetric: From here to validity. Paper presented at *the annual meeting of the American Educational Research Association*. Seattle, Washington.

Fitzpatrick, M. (2011). *Engaging writing 2: Essential skills for academic writing* (2nd ed.). New York: Pearson Longman.

Goldberg, A., Russell, M., & Cook, A. (2003). The effect of computers on student writing: A meta-analysis of studies from 1992 to 2002. Journal of Technology, Learning, and Assessment, 2(1). Retrieved from http://ejournals.bc.edu/ojs/index.php/jtla/article/view/1661

Grimes, D., & Warschauer, M. (2010). Utility in a fallible tool: A multi-site case study of automated writing evaluation. *Journal of Technology, Learning, and Assessment, 8*(6), 4–44. Retrieved from http://ejournals.bc.edu/ojs/index.php/jtla/article/view/1625/1469

James, C. (2006). Validating a computerized scoring system for assessing writing and

placing students in composition courses. *Assessing Writing, 11*(3), 167-178. http://dx.doi.org/10.1016/j.asw.2007.01.002

Keith, T.Z. (2003). Validity of automated essay scoring systems. In M.D. Shermis & J.C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective*. Mahwah, NJ: Erlbaum.

Halliday, M.A.K. (1994). *An introduction to functional grammar* (2nd ed.). New York, NY: Edward Arnold.

Halliday, M.A.K. (1998). Things and relations: Regrammaticising experience as technical knowledge. In J.R. Martin and R. Veel (Eds.), *Reading science: Critical and functional perspectives on discourses of science* (pp. 185-235). New York: Routledge.

Halliday, M.A.K., & Martin, J.R. (1993). *Writing science: Literacy and discursive Power*. Washington DC: The Falmer Press.

Lai, Y.-H. (2010). Which do students prefer to evaluate their essays: Peers or computer program. *British Journal of Educational Technology, 41* (3), 432-454. http://dx.doi.org/10.1111/j.1467-8535.2009.00959.x

Li, Z., Link, S., Ma, H., Yang, H., & Hegelheimer, V. (2014). The role of automated writing evaluation holistic scores in the ESL classroom. *SYSTEM Journal, 44*, 66-78. http://dx.doi.org/10.1016/j.system.2014.02.007

Link, S., Durson, A., Karakaya, K., & Hegelheimer, V. (2014). Towards best ESL practices for implementing automated writing evaluation. *CALICO Journal, 31* (3), 323-344. http://dx.doi.org/10.11139/cj.31.3.323-344

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing, 19*(3), 246-276. http://dx.doi.org/10.1191/0265532202lt230oa

Mohan, B., Leung, C., & Slater, T. (2010). Assessing language and content: A functional perspective. In A. Paran & L. Sercu (Eds.), *Testing the untestable in language education* (pp. 219-242). Bristol, UK: Multilingual Matters.

Mohan, B., & Slater, T. (2004). The evaluation of causal discourse and language as a resource for meaning. In J. A. Foley. (Ed.), *Language, education & discourse: Functional approaches* (pp. 255-269). London: Continuum.

Mohan, B., & Slater, T. (2005). A functional perspective on the critical 'theory/practice' relation in teaching language and science. *Linguistics and Education, 16*, 151-172. http://dx.doi.org/10.1016/j.linged.2006.01.008

Mohan, B., Slater, T., Luo, L., & Jaipal, K. (2002). *Developmental lexicogrammar of causal explanations in science.* Paper presented at the International Systemic Functional Linguistics Congress (ISFC29), Liverpool, UK.

Page, E.B. (2003). Project Essay Grade: PEG. In M.D. Shermis & J.C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 43-54). Mahwah, NJ: Lawrence Erlbaum Associates.

Page, E.B., & Peterson, N.S. (1995). The computer moves into essay grading: Updating the ancient test. *The Phi Delta Kappan, 76* (7), 561-565.

Painter, C. (1999). *Learning through language in early childhood*. London: Continuum.

Petersen, N.S. (1997) *Automated scoring of written essays: Can such scores be valid?* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Saldana, J. (2009). *The coding manual for qualitative researchers.* Washington DC: SAGE.

Slater, T. (2004). The discourse of causal explanations in school science. PhD thesis, University of British Columbia.

Slater, T., & Mohan, B. (2010). Towards systematic and sustained formative assessment of causal explanations in oral interactions. In A. Paran & L. Sercu (Eds.), *Testing the untestable in language education* (pp. 256-269). Bristol, UK: Multilingual Matters.

Taylor, C., Kirsch, I., & Eignor, D. (1999). Examining the relationship between computer familiarity and performance on computer-based language tasks. *Language Learning, 49*(2), 219-274. http://dx.doi.org/10.1111/0023-8333.00088

Veel, R. (1997). Learning how to mean-scientifically speaking: Apprenticeship into scientific discourse in the secondary school. In F. Christie & J.R. Martin (Eds.), *Genre and Institutions: Social Processes in the Workplace and School* (pp. 161-195). London: Continuum.

Wang, J., & Brown, M.S. (2007). Automated essay scoring versus human scoring: a comparative study. *Journal of Technology, Learning, and Assessment, 6*(2). Retrieved from http://www.jtla.org. Retrieved from http://files.eric.ed.gov/fulltext/EJ838612.pdf

Wang, Y.-J., Shang, H.-F., & Briody, P. (2013). Exploring the implact of using automated writing evaluation in English as a foreign language university students' writing. *Computer Assisted Language Learning, 26* (3), 234-257. http://dx.doi.org/10.1080/09588221.2012.655300

Warschauer, M. (2010). Invited commentary: New tools for teaching writing. *Language Learning & Technology, 14*(1), 3-8. Retrieved from http://llt.msu.edu/vol14num1/commentary.pdf

Weigle, S.C. (2010). Validation of automated scoring of TOEFL iBT tasks against non-test indicators of writing. *Language Testing, 27*(3), 335-353. http://dx.doi.org/10.1177/0265532210364406

Appendix A: The Writing Prompt

The Original Prompt:

Reasons for Attending College (Expository)

People attend a college or university for many different reasons (for example, new experiences, career preparation and increased knowledge). Why do you think people attend college or university? Use specific reasons and examples to support your answer.

The Edited Prompt:

Level: TOEFL

Word count: 250-300 words

Time limit: 30 minutes

The Prompt:

Reasons for Attending College (Expository) – Cause and Effect Essay

People attend a college or university for many different reasons (for example, new experiences, career preparation and increased knowledge). Why do you think people attend college or university? Use specific reasons and examples to support your answer.

Requirements:

Your introduction/conclusion should be no more than two sentences each.

You need to provide detailed discussion on (1) what types of experiences/career preparation/knowledge/other ideas that attending a college or university can provide; (2) how attending a college or university can provide people with these experiences/ chances for career preparation/useful knowledge or other benefits (3) what are the possible benefits of obtaining these experiences/ career preparation/ useful knowledge/other ideas.

Appendix B: The SFL Coding Rubric

Features	Initials	meaning	Examples
External temporal conjunctions	ETC	Conjunctions indicating time sequence	When, then…
External causal conjunctions	ECC	Conjunctions indicating causality	If, because, therefore, as, for, so, hence, thus, in that, only if…
Internal conjunctions	IC	Logical conjunctions, rather than following a natural sequence of events	Firstly, secondly, thirdly, additionally, furthermore, in addition, moreover, finally, lastly…
Temporal circumstances	TC	Adverbials indicating time sequence	After, eventually, …
Causal circumstances	CC	Adverbials indicating causality	As a consequence, due to the fact that, in of view, owing to, due to, to, with, by, through, under… **With the help of such kind of equipment**, students will get fully understand the knowledge about computer. **By** registering to a university, you could have a brighter future. **Through** this meeting and work experiences, people can realize what kind of job they really want to do. **Under the guiding of professors**, college students can learn much more efficiently.
Temporal processes	TP	Verbs indicating time	Follow, proceed, initiate…
Causal processes	CP	Verbs indicating causality	Brings about, cases, contributes to, gives rise to, is responsible for, leads to, produces, results in, is due to, occurs as the result of, results from, prevent, improve, cause, affect…

Proof processes	PP	Verbs indicating proof	Prove… (May not be applicable for this study)
Temporal entities	TE	Nouns indicating time	May not be applicable for this
Causal entities	CE	Nouns indicating causality	Cause, reason, effect, consequence, result, factor,
General metaphoric entities	GME	Nominalization (noun as transformation of a verb)	Reactant, product, function, circulation, nutrition, prevention (e.g. react => reactant; produce => product; circulate => circulation…) study-> studying, attend-> attending