Computer Assisted Language Instruction Consortium

CALICO JOURNAL (ONLINE) ISSN 2056–9017

Article

Connecting *Criterion* scores and Classroom Grading Contexts: A Systemic Functional Linguistic Model for Teaching and Assessing Causal Language

Hong Ma and Tammy Slater

Abstract

This study utilized a theory proposed by Mohan, Slater, Luo, and Jaipal (2002) regarding the Developmental Path of Cause to investigate AWE score use in classroom contexts. This 'path' has the potential to support validity arguments because it suggests how causal linguistic features can be organized in hierarchical order. Utilization of this path enabled this study to investigate AWE scores by comparing them to ratings based on teachers' intuitions as well as to scores based on assignment rubrics. Qualitative focus group data suggested that the path can help teachers articulate their intuitions. Quantitative results showed that the grades provided by raters trained to use the path tended to support AWE scores from Criterion, a web-based AWE system, more strongly than did rubric-generated grades. The findings from this study suggest that Criterion scores not only closely correlated with teachers' intuitions and with raters trained to use the path, but that the use of the path for teaching may support the implementation of AWE systems in classroom contexts, and would help students focus on the core of a cause-effect essay: appropriateness and sophistication of causal language.

KEYWORDS: AWE SCORES; CAUSAL DISCOURSE; SYSTEMIC FUNCTIONAL LINGUISTICS; THE DEVELOPMENTAL PATH OF CAUSE

Affiliation

Iowa State University, Ames IA 50011, USA.
email: mahong84@hotmail.com (corresponding author)

CALICO JOURNAL VOL 33.1 2016 1–18

doi : 10.1558/cj.v33i1.26562

Introduction

The development of academic language and the mastering of diverse subject matter are indispensable for academic and occupational success (Grimes & Warschauer, 2010). In academic writing, causal explanations play a critical role since they occur across subject areas, from explaining causes and effects in science, to documenting historical conflicts, to arguing motives in literary characters. Even though various automated evaluation systems have been developed since the 1960s to compensate for the shortage of teacher time available for the individual responses needed to hone students' writing skills (Burstein, Chodorow, & Leacock, 2003a; Dikli, 2006), few studies have investigated the extent to which AWE scores can accurately reflect the quality of causal explanations and whether automated evaluation systems can scaffold students' development of causal explanations.

Different from earlier counterparts such as Project Essay Grade, created in the 1960s exclusively for automated scoring, the most up-to-date Automatic Writing Evaluation (AWE) systems, *Criterion* by Educational Testing Service and *My Access!,* can provide both immediate scoring and formative feedback (Chen & Cheng, 2008), thus holding potential for classroom implementation. However, some writing instructors are dissuaded from using AWE systems because AWE scores have not consistently correlated well with instructor grades (Ebyary & Windeatt, 2010; Li, Link, Ma, Yang, & Hegelheimer, 2014; Wang & Brown, 2007). Re-examining this psychometric method of validating AWE scores, we argue that using rubric-based instructor grades as the benchmark may be inappropriate, since these rubrics can at times subvert instructors' intuitive judgments by forcing them to focus on elements peripheral to the writing task (Slater, 1998; Mohan & Slater, 2004). Slater (1998; see also Mohan & Slater, 2004), examining connections between a functional perspective on the assessment of causal discourse and raters' intuitions of the discourse, suggested that the scores assigned can be valid and reliable if what is being assessed matches raters' intuitions.

Another reason to hesitate implementing AWE systems in writing instruction is that using AWE tools without appropriate pedagogical intervention could cause negative washback (CCCC, 2006; Cheville, 2004; Ericsson, 2006; Ziegler, 2007). Research has reported that an AWE system 'favors lengthiness … overemphasizes the use of transition words … ignores coherence and content development … [and] discourages unconventional ways of essay writing (Chen & Cheng, 2008: 104). Students thus could manipulate their writing to achieve higher scores instead of improving the overall quality of their writing.

We therefore propose a theoretically based model that appears to take instructors' intuition into consideration. We hypothesize that this model, which targets cause-effect discourse, can achieve two main goals: (1) given that

there appears to be a close connection between the model and teachers' intuitive judgments, using the model in classroom writing evaluation may result in a higher correlation between AWE scores and instructors' grades; and (2) if implemented, this model may yield positive washback as it encourages students to use more sophisticated causal language to obtain higher grades from both AWE systems and instructors. In this paper, we focus on the first point in hopes that we can provide a foundation for future testing of the second.

The following sections review previous literature on AWE score use in classroom contexts and then introduce the model we are proposing, the Developmental Path of Cause.

Previous research on AWE score use

Two approaches, the psychometric approach and the naturalistic classroom-based approach, have been used to examine the appropriateness of AWE score use in classroom contexts. The psychometric studies have relied on the correlational value between teacher grades and *Criterion* scores as evidence. Compared with the high correlation values consistently reported in testing contexts (Attali, Bridgeman & Trapani, 2010; Burstein *et al.*, 2003a), human-machine correlation in classroom-based studies has generally been much lower and has tended to vary considerably (Ebyary & Windeatt, 2010; Li *et al.*, 2014; Wang & Brown, 2007). When AWE scores do not correlate well with instructor grades, students receiving conflicting responses to their writing can become confused as to what features to focus on for revision. However, the method of validating AWE score use exemplified in these studies, requiring instructors to assess students' essays using rubrics, may be less valid because rubrics can force raters to pay attention to other elements, suppressing instructors' intuition (Mohan & Slater, 2004; Slater, 1998).

Because of issues associated with the psychometric approach, an increasing number of studies have adopted a naturalistic classroom-based approach, relying on classroom observations, questionnaires, and interviews to collect evidence about how AWE systems are used, and what teachers' and students' perceptions are of AWE systems (Chen & Cheng, 2008; Grimes & Warschauer, 2010; Li *et al.*, 2014). Findings from these studies emphasize the importance of teachers' intervention to counteract the potentially negative washback from using AWE scores. For example, Chen and Cheng (2008) argued for careful intervention by teachers because, as noted above, students perceive automated scores as favoring lengthiness, overemphasizing the use of transition words and failing to judge coherence and content development, and discouraging unconventional ways of organizing information. Grimes and Warschauer (2010) reported that the immediacy of automated scoring succeeded in motivating students to revise more frequently; however, low disagreement between

teachers' grades and automated scores caused confusion and put teachers in the awkward position of having to defend scores.

Given the issues existing in research on AWE score use in classroom contexts, we propose a theoretically based model that appears to match instructors' intuitions and may enable us to achieve two main goals: Using the model: (1) will result in a higher level of agreement between AWE scores and instructor grades; and (2) may promote positive washback by helping students improve their ability to construct sophisticated causal discourse. The following section will introduce the theoretical model we are suggesting, the Developmental Path of Cause.

The Developmental Path of Cause

Adopting a sociolinguistic perspective, researchers in Systemic Functional Linguistics (SFL) view language as formed in and for social communication, and they have developed their analytical categories by examining authentic discourse in use (Halliday & Martin, 1993). The Developmental Path of Cause was the result of a corpus analysis of written cause-effect discourse in two encyclopedias (Mohan *et al.*, 2002). The analysis revealed patterns that provided support that 'causal language develops along two dimensions: a lexicogrammatical dimension (the horizontal axis) and a semantic dimension (the vertical axis)', as shown in Figure 1 (Slater & Mohan, 2010: 261). The horizontal axis of the model describes the metaphoric progression of causal linguistic features from relators (conjunctions) towards circumstances, processes, qualities, and entities (as described by Halliday, 1998), while the vertical axis describes the semantic progression of causal language from time, to cause, and proof (Slater & Mohan, 2010).

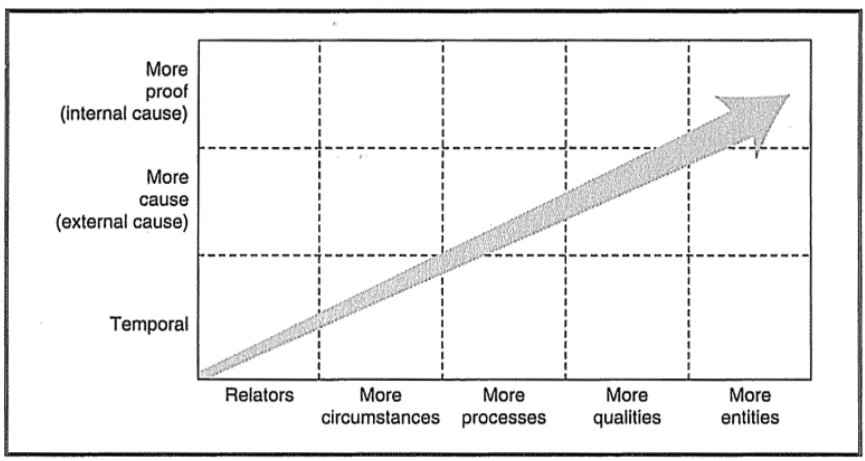

Figure 1: The developmental path of cause. (Mohan *et al.*, 2002)

Slater (2004) combined both dimensions of causal linguistic features and illustrated linear progression in terms of linguistic difficulty from external temporal conjunctions to general metaphoric entities (as shown in Table 1).

Table 1: Linear Progression of Developmental Path of Cause and Examples

Features	Meaning	Examples
External temporal conjunctions	Conjunctions indicating time sequence	When, then
External causal conjunctions	Conjunctions indicating causality	If, because, therefore
Internal conjunctions	Logical conjunctions organizing text	Firstly, additionally, furthermore
Temporal circumstances	Adverbials indicating time sequence	After
Causal circumstances	Adverbials indicating causality	As a consequence, due to, through
Temporal processes	Verbs indicating time	Follow, proceed,
Causal processes	Verbs indicating causality	Causes, contributes to,
Proof processes	Verbs indicating proof	Prove…
Temporal entities	Nouns indicating time	No The beginning, the following
Causal entities	Nouns indicating causality	Cause, effect, consequence,
General metaphoric entities	Nominalization (noun transformed from a verb)	Reactant, product, circulation…

(adapted from Slater, 2004)

At the heart of our study is the argument that the use of the Developmental Path of Cause not only has the potential to help students produce better causal discourse (positive washback), but within assessment, its use can help provide more valid evaluations of cause-effect essays, since the model appears to capture teachers' intuitive judgments. Our study examines whether teachers who are assessing cause and effect essays are attending to the features that this path captures (i.e., whether they can intuitively use this model), and how scores given by raters who are trained to use this path compare to those given by classroom instructors' ratings and those assigned by *Criterion*. The specific research questions are:

1. What are the causal discourse features experienced teachers attend to, based on their intuitions, when evaluating students' cause-effect essays, and how do these compare to the Developmental Path of Cause?
2. How do scores generated by *Criterion* compare with scores assigned by raters trained to use the Developmental Path of Cause (whom we refer to as SFL raters)? How do these compare with scores based on the holistic scoring rubric that classroom instructors were encouraged to use?

The next section will describe the methodology used in the study.

Methodology

Study context

The study was conducted in seven sections of an undergraduate ESL composition course in a mid-western American university. The course was offered to help ESL students meet the requirements needed to enroll in first-year composition classes. Data were collected with the approval of the institution and the informed consent of the participants.

This ESL composition course was designed to develop students' ability to write academically. Because the students were required to write consequential (i.e., cause and effect) texts, the course offered a highly useful context for exploring the connections between the use of the Developmental Path of Cause and the AWE system being used.

Participants

This study involves four different groups of participants: 58 tertiary level ESL students enrolled in seven sections of the ESL composition course, five writing instructors (two of the writing instructors were teaching two sections), three experienced teachers, and three raters trained to use the Developmental Path of Cause (hereafter SFL raters). Of these SFL raters, one was considered the major rater and two were considered 'auxiliary raters', used to establish coding reliability of the essays. Essays that the student participants produced during the university-created English Placement Test identified them as Advanced Low according to the ACTFL proficiency guidelines (ACTFL, 2002). All teachers and raters were graduate students in the English Department. Table 2 summarizes demographic information and the responsibilities of the graduate participants.

Table 2: Graduate Participants: Demographic Information & Responsibilities

	Native language	Gender	Education	Responsibilities
5 writing instructors	American (2), Turkish (1), Korean (1), Chinese (1)	Males (2) Females (3)	Master students (2) PhD students (3)	Graded own students' writing
3 experienced teachers	American (2), Chinese (1)	Males (0) Females (3)	Master students (0) PhD students (3)	Attended focus group interview and ranked 6 essays intuitively
3 SFL raters	Turkish (1), Vietnamese (1), Chinese (1)	Males (0) Females (3)	Master students (0) PhD students (3)	Coded students' essays using the Developmental Path of Cause

The three groups of graduate participants were independent of each other with the exception of one Chinese doctoral student, who was both a writing instructor and a teacher in the focus group.

Materials

Materials in this study included *Criterion*, a cause-effect writing prompt, and the grading rubric as the following sections detail.

The AWE System Criterion

Criterion is a web-based AWE system developed by Educational Testing Service (ETS) to provide holistic scores and diagnostic feedback. Instantaneous holistic scores and immediate feedback are generated with support of the E-rater scoring application and the Critique writing analysis function respectively (Attali & Burstein, 2005).

E-rater was trained on a large sample of human scored essays ranging from one to six points (Burstein, Marcu, & Knight, 2003b). It was designed to extract more than 50 features specified in scoring guides from students' essays and predict scores using regression analysis with the identified features as predictor variables (Burstein *et al.*, 2003a). Along with holistic scores, the writing analysis function (Critique) detects undesirable features in terms of grammar, usage, mechanics, vocabulary choice, undesirable style, and essay-based discourse elements and provides feedback on these (Attali & Burstein, 2005; Burstein *et al.*, 2003a).

The cause-effect writing prompt

The topic 'Reasons for Attending College' in the *Criterion* TOEFL-level topic library was chosen to elicit students' writing response because of its appropriate difficulty level and its focus on causality. Although *Criterion* allows teachers to create their own topics, holistic scoring can only be reported for topics in the *Criterion* library. This writing task was a 30-minute timed task with a word limit of 250 to 300 words. Several questions were added to the original wording of the prompt (shown in Appendix A) to encourage students to provide detailed cause-effect reasoning.

The Grading Rubric

A rubric for grading students' essays was designed to closely resemble the rubric for other cause-effect writing assignments so that the instructors would maintain their grading style. Adjustments were made, however, because the essays assigned for this study did not require a fully developed introduction (context) or citations (style). In addition, questions intended to evaluate whether students provided extended discussions of the reasons for attending

college were added to the study rubric (under substance) and were intended to draw attention to the specific writing task for this study. Otherwise, both the course rubric and the rubric used in this study had the same primary content and supported holistic grading based on a 0- to 50-point scale. This rubric can be found in Appendix B.

Data collection

All students spent 30 minutes writing responses using computers during lab sessions (three sections), at home as a journal assignment (two sections), or during class using pen and paper (two sections). The pen-and-paper essays were retyped and submitted to *Criterion* for scoring. The instructors graded the versions of the essays written by their own students according to the rubric presented in Appendix B.

Once the essays were written, submitted, and graded, three experienced English writing teachers participated in a focus group interview that asked them to justify an evaluation of three sets of essays (two essays in each set) using their intuitions and experience as guides (as per Low, 2010). These six essays were selected because the instructor grades and the *Criterion* scores ranked the two essays in each set differently.

Three raters (the major rater and the two auxiliary raters) evaluated 20 essays randomly selected from the pool of essays (see Table 1, which was used as the SFL coding rubric), calibrating their evaluation through discussion and consultation with an expert in SFL. After all three reached an agreement on these 20 essays, the major rater assessed all remaining essays. Final SFL-based grades were calculated by adding up the points generated by the use of appropriate cause-effect resources. As we could not locate any studies that present an appropriate grading process, we followed the hierarchy of causal discourse expressions in Table 1 and awarded half a point to each causal feature that was a step further along the path, with the lowest value being external temporal conjunctions (worth 0.5 each), and each expression higher in the hierarchy (moving down the list) earning half a point more (e.g., the second feature 'external cause conjunctions' was worth one point, compared with one half point for the first feature 'external temporal Conjunctions'; the feature 'general metaphor entities' was worth 5.5 points). This grading system aimed at reflecting the progression in terms of linguistic complexity, as per the Developmental Path of Cause. Moreover, given that the essays required in this context were of limited and similar length, we felt that this method would reveal that higher scores within these similarly long essays would signify a greater use of those linguistic resources that appear at the higher end of the Developmental Path of Cause. The score of each essay was thus obtained by totaling the scores of each use of a causal linguistic feature to reward the frequency of the target feature.

Since this rating is a frequency count, the length of the text in particular may affect the scoring and would need to be addressed in future work. However, in this case, we argue that scores based on the Developmental Path of Cause reflect participants' writing proficiency levels and are directly comparable in that this model is rewarding writers that use more complex syntactic features.

Data analysis

To respond to the first research question regarding intuitive judgments of cause-effect essays, interviews were coded to capture the features teachers attended to when evaluating student work. We employed open coding for the first round (see Saldana, 2009). The data suggested that although grammatical accuracy and organization affected the teachers' impressions of the quality of the essays, causal language received a great deal of attention. The Developmental Path of Cause was implemented in our coding scheme during the second round of analysis, and observations about causal discourse were thus related to the semantic and lexicogrammatical dimensions of this model.

For the second research question, we began by using Spearman's ρ to calculate correlations between (1) *Criterion* scores and teachers' rubric-based scores and between (2) *Criterion* scores and those given by raters using the theoretical model. Spearman's ρ is appropriate when one variable is ordinal data and the other variable is interval data (Bachman, 2004). In this case, *Criterion* scores were ordinal, and teachers' rubric-base scores and those given by raters using the model were interval data. We then calculated the correlation between (3) scores given by the model-based raters and the rubric-based scores using Pearson's r. Pearson's r was adopted since both variables, the scores given by the model-based raters and rubric-based scores, were interval variables (Bachman, 2004). Finally, to test whether the correlation between (1) and the correlation between (2) were significantly different, we calculated the z-score value using the following formula, as per Kleinbaum, Kupper, and Muller, 1988 (p. 92):

$$H_0\colon \rho_{12} = \rho_{13}$$

$$Z = \frac{(r_{12} - r_{13})\sqrt{n}}{\sqrt{(1 - r_{12}^2)^2 + (1 - r_{13}^2)^2 - 2r_{23}^3 - 2(r_{23} - r_{12}r_{13})(1 - r_{12}^2 - r_{13}^2 - r_{23}^2)}}$$

Results & Discussion

The following sections describe our findings and argue that the features identified in the Developmental Path of Cause not only reflect those that teachers

focus on when assessing cause-effect essays, but that the scores based on this path correlate well with *Criterion* scores. Each of the two research questions will be addressed in turn.

1. What are the causal discourse features experienced teachers attend to, based on their intuitions, when evaluating students' cause-effect essays, and how do these compare to the Developmental Path of Cause?

The qualitative data revealed that experienced teachers rating student essays during the focus group interview attended intuitively to the aspects of causal discourse—both semantic and lexicogrammatical—that the path attempts to illuminate. Semantically, the focus group teachers stated they were expecting texts that went beyond temporal or sequential descriptions into the realm of cause and effect. These teachers made comments such as 'the student's essay is very much descriptive… did not have the cause effect' (T3), and 'it uses 'can' several times and I didn't find so many causes and effects' (T1). The teachers claimed that the lower-rated 'descriptive' essays were using modals such as 'can,' 'may,' and 'might' to suggest options rather than constructing cause-effect texts.

The general consensus was that essays needed to have logical causal relations to receive higher ratings. The teachers commented that in some lower-rated essays, there was an 'absence of the actual cause and effect language [and] the absence of a clear topic sentence in the paragraph [which] really affected the overall genre of it being cause and effect' (T3). But it was not simply the existence of causal discourse that was noticed. The raters consistently used terms such as 'evidence,' 'claims,' and 'warrants' to capture what they felt was a necessary part of a well-constructed causal text, supporting the move to what the path refers to as 'proof.' Comments such as the following captured these observations:

'The logic was problematic. I have problems with their claims.' (T2)
'It's not clear that the student is trying to build their evidence for their claims.' (T3)
'The data that they used to support their points made sense. I believe their points; I accept their claims.' (T3)

Thus the teachers' comments highlighted their beliefs that the meanings of the essays needed to involve causality but also needed to provide evidence (proof) for the causal relations, thereby suggesting that these teachers were intuitively following the semantic progression described by the Development Path of Cause.

The data also showed that experienced teachers' intuitive judgment corresponded with the path's lexicogrammatical progression. Causal resources

mentioned by the raters were relators/conjunctions, such as 'so,' 'so that,' 'because,' 'in order to,' and 'if' (in conditional clauses), causal circumstances such as 'through,' and causal processes such as 'make,' 'shape,' and 'form.' Other processes (verbs) that may not always suggest causality were used in the better-rated essays to create subtle causal relations of means/end. For example, one student wrote that reasons to attend college include 'to form strong will' or 'to obtain education.' The same writer could have used simple relators to raise the same argument: 'If I go to college, I can form strong will or obtain education.' The use of the process in a means/end construction, however, moved this student farther right on the path. The experienced teachers picked up on these subtle differences in their oral evaluations of the essays.

Another example of highly rated causal language was the use of 'due to' in the sentence: 'due to getting these experiences and knowledge, students can find job easily in the future.' This feature requires a following noun or gerund—an entity (nominal group), which is considered to be at the high end of the path. T1's observation of the writer's use of 'due to' and her elaboration of that clause as an example of 'pretty good' writing supports the view that the better written causal essays are ones that attempt to move more towards the higher end of the lexicogrammatical axis, an observation made in Slater (2004) and Slater and Mohan (2010). Less highly rated essays on the other hand, seemed to exhibit problems trying to use entities. Instead of using clear nominal forms, less competent writers tended to use pronouns, but these at times caused comprehension breakdowns; as the teachers pointed out, pronouns with a lack of clear reference 'really weaken the sentence' (T3).

The qualitative data suggest that the experienced teachers were intuitively attending to the features captured by the Developmental Path of Cause. These findings suggest that incorporating the path into teaching may help students understand what the task is asking them to do. But does this understanding help narrow the gap in existing correlations between classroom instructors' grades and *Criterion* grades? We attempted to explore this issue by addressing our next research question:

2. How do scores generated by *Criterion* compare with scores assigned by raters trained to use the Developmental Path of Cause (whom we refer to as SFL raters)? How do these compare with scores based on the holistic scoring rubric that classroom teachers were encouraged to use?

The correlation between scores generated by *Criterion* and scores assigned by SFL raters was $\rho(56) = .604$ ($p < .05$). While not a high correlation, it was nonetheless higher than the correlation between the scores assigned by our SFL raters and the classroom instructors' ratings ($r(56) = .346$, $p < .05$), or the

correlation between scores generated by *Criterion* and the classroom instructors' ratings (ρ (56) = .484, $p < .05$). In addition, the result of the z-score test suggested that the correlation between scores generated by *Criterion* and scores assigned by our SFL raters was significantly higher than the correlation between the scores assigned by our SFL raters and the classroom instructors' ratings ($z = 2.2$, $p < .05$, two-tailed). These findings lead us to believe that *Criterion* scores, rather than instructors' ratings from the rubric, were more aligned to the SFL ratings. Such an increase in correlation is promising as it suggests that the implementation of the Developmental Path of Cause as a theoretical model for teaching and testing in classrooms may result in higher agreement between grades assigned by teachers and the *Criterion* scores, which may in turn provide teachers and students with higher levels of confidence in the *Criterion* scores.

It may not be surprising to see that the correlation between *Criterion* and scores assigned by our SFL raters was higher than the correlation between the scores assigned by these raters and the instructors' ratings. As reported in Deane (2013), the scoring engine of *Criterion* operationalizes the construct of text quality as computable features that we believe appear to overlap with the linguistic progression specified in the Developmental Path of Cause. We could reasonably assume that papers employing causal linguistic features towards the higher end of the path are scored higher by *Criterion*, which measures the number of discourse elements, length of elements, average word length, and sophistication of word choice.

Given that the scoring engine of *Criterion* only captures a subset of features specified in most rubrics facilitating holistic grading, *Criterion* by no means assesses the same construct as human scoring. Nevertheless, the fact that *Criterion* can achieve high levels of agreement with human raters is substantially supported, since 'those who have developed high fluency and control over text production processes are precisely those who have the cognitive resources needed to practice the skills needed to master a broader writing construct' (Deane, 2013: 18). Previous research on AWE score use in classroom contexts, on the contrary, has yielded teacher-machine agreement varying from .11 (Wang & Brown, 2007) to .839 (Ebyary & Windeatt, 2010). We therefore question whether instructors rating students' essays applied the rubrics consistently, given that the machine-instructor agreement varies so greatly (see, for example, Ebyary & Windeatt, 2010, Li *et al.*, 2014). With such a notorious inconsistency of instructors' rating and the resulting skepticism toward using *Criterion* in classroom contexts as well as our favorable findings connecting teachers' intuitions, the Developmental Path of Cause, and *Criterion*, we argue for continued testing of the Developmental Path of Cause in classrooms which use *Criterion* scores.

Conclusion

The use of assessment tasks can influence teaching and learning (Green, 2007). We suggest that the essays described in this paper, because they are being scored formatively and summatively, are assessment tasks and that as such, students want to know how to obtain better grades for future performance. As the literature has shown, however, when AWE scores differ from instructors' assigned grades, teaching and learning may become frustrating and difficult. From our findings, we argue that if instructors adopt the Developmental Path of Cause in the classroom as a theoretical model for teaching students how to improve their causal discourse, the assigned grades may be more aligned to the scores suggested by *Criterion*, making the implementation of AWE scoring more valuable in the classroom context. This alignment in turn may help establish construct validity and trustworthiness of these assignments and promote positive washback, good teaching practice 'that is evidentially linked to the introduction and use of the test' (Messick, 1996: 16).

We have attempted to show that experienced teachers are intuitively paying attention to elements that are reflected clearly on the Developmental Path of Cause. However, these elements appear not to be explicit in the assessment rubric, a tool that may be suppressing teachers' impressions of student writing, as Mohan and Slater (2004) observed. Teaching from the Developmental Path of Cause may thus help students expand their resources for constructing written causal discourse by raising their awareness of sophisticated cause-and-effect language. Furthermore, we suggest that students can use *Criterion* to show them how their writing of causal discourse is improving. Moreover, findings from our small-scale study suggest that if teachers use the Developmental Path of Cause to evaluate the drafts that have gone through *Criterion*, their grades may correlate well with the AWE feedback, thus minimizing the issues that previous literature has reported and providing a more successful implementation of the scores in classroom contexts.

Findings from this study should be interpreted with caution. First, these results are based on a small sample of essays and a limited number of instructors assessing them. While machine scoring is consistent, it is possible that different instructors may grade student essays differently despite using the same rubric; thus in future studies, larger numbers of essays should be scored and reliably calibrated to compare the scores with SFL raters' scores. Moreover, as we mentioned, essay length should also be addressed, as longer essays may involve a greater use of causal discourse features, and essays from more fluent writers could therefore result in higher scores even when their features occur lower on the path. Yet despite our small sample size, our theory-based approach to validating AWE scores in the classroom context has shown

potential and merits further empirical research. A second limitation is that the Developmental Path of Cause is restricted to the evaluation and teaching of causal discourse alone. The issue of AWE score use in classrooms for other genres needs to be examined using comparable theories.

This study set out to achieve two goals. The first was to see if the use of the Developmental Path of Cause could more closely connect *Criterion* scores to teachers' grades. Not only did our analysis show that teachers appeared to attend to the features of the Developmental Path of Cause (but may have been limited by the rubric they were using), statistical tests revealed higher correlations between *Criterion* scores and grades given by raters scoring from this theoretical model than between these AWE scores and rubric-based assessments. Despite study limitations, our results are promising in that they propose a way to increase the agreement between teacher grades and *Criterion* scores. The second goal, however, has yet to be examined. Further research needs to be carried out to address the consequential aspect of construct validity (Messick, 1996), to see whether the use of the Developmental Path of Cause in tandem with regular use of *Criterion* feedback can indeed promote improvement in students' writing of cause-effect essays.

About the Authors

Hong Ma is a PhD candidate in Applied Linguistics and Technology at Iowa State University. Her primary research interests lay in computer-assisted language learning and language testing. She is currently leading multiple research projects, which intend to develop and evaluate a vocabulary-learning tool and extract a more pedagogy-informed vocabulary list using programming language.

Tammy Slater is an associate professor in Applied Linguistics and Technology at Iowa State University. Her research draws upon Systemic Functional Linguistics to understand the development of academic language through content-based and project-based teaching and learning, particularly as it informs English language education.

References

American Council on the Teaching of Foreign Languages. (2002). *Program standards for the preparation of foreign language teachers* (Initial level- undergraduate & graduate) (For K-12 and secondary certification programs). Retrieved from http://www.actfl.org/sites/default/files/pdfs/public/ACTFLNCATEStandardsRevised713.pdf.

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.

Bachman, L. (2004). *Statistical Analysis for Language Assessment.* New York: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511667350

Burstein, J., Chodorow, M., & Leacock, C. (2003a). Criterion online essay evaluation: An application for automated evaluation of student essays. *AI Magazine,* 25 (3), 27–35. http://dx.doi.org/10.1609/aimag.v25i3.1774

Burstein, J., Marcu, D., & Knight, K. (2003b). Finding the WRITE stuff: Automatic identification of discourse structure in student essays. *IEEE Intelligent Systems: Special Issue on Natural Language Processing,* 18 (1), 32–39. http://dx.doi.org/10.1109/MIS.2003.1179191

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.

Cheville, J. (2004). Automated Scoring Technologies and the Rising Influence of Error. *English Journal,* 93 (4), 47–52. http://dx.doi.org/10.2307/4128980

Conference on College Composition and Communication (2006). *Writing assessment: A position statement*. Retrieved July 20, 2007, from http://www.ncte.org/cccc/resources/positions/123784.htm.

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing,* 18, 7–24. http://dx.doi.org/10.1016/j.asw.2012.10.002

Dikli, S. (2006). An overview of automated scoring of essays. *Journal of Technology, Learning, and Assessment,* 5 (1). Retrieved from http://www.jtla.org

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

Ericsson, P. F. (2006). The meaning of meaning: Is a paragraph more than an equation? In P. F. Ericsson & R. H. Haswell (Eds.), *Machine scoring of student essays: Truth and consequences*, 28–37. Logan: Utah State University Press.

Green, A. (2007). *IELTS washback in context: Preparation for academic writing in higher education.* Cambridge, UK: Cambridge University Press.

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.

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

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

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

Klenbaum, D. G., Kupper, L. L., & Muller, K. E. (1988). Applied regression analysis and other multivariable methods. Boston: PWS-KENT Publishing Company.

Low, M. (2010). Teachers and texts: Judging what English language learners know from what they say. In A. Paran & L. Sercu (Eds), *Testing the untestable in language education* (pp. 241–255). Bristol, UK: Multilingual Matters.

Messick, S. (1996). *Validity and washback in language testing.* Princeton, NJ: Education Testing Services.

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*, 255–269. London: Continuum.

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.

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

Slater, T. (1998). *Evaluating causal discourse in academic writing.* MA thesis. University of British Columbia.

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*, 256–269. Bristol, UK: Multilingual Matters.

Wang, J., & Brown, M. S. (2007). Automated essay scoring versus human scoring: a comparative study. *Journal of Technology, Learning, and Assessment,* 6 (2).

Ziegler, W.W. (2007). Computerized Writing Assessment: Community College Faculty Find Reasons to say 'Not Yet'. In P. F. Ericsson & R. Haswell (Eds.), *Machine Scoring of Human Essays: Truth and Consequences*, 138–153. Logan, Utah: Utah State University Press.

Appendix A: The Writing Prompt

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.

Edited Prompt:

Level: TOEFL
Word count: 250-300
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: Comparison between rubrics

	The rubric used in this study	**The typical 101C rubric**
context	Brief introduction sets the context (pp. 90–91)	Full introduction sets context (time period, people, place) and introduces major factors involved. (pp. 90–91)
	Thesis states the reasons for attending college.	Thesis states causes and effects of the phenomenon discussed.
substance	Includes extended discussion of (1) what types of experiences/career preparation/knowledge/other ideas that attending college or university can provide; (2) how attending college or university can provide people with these experiences/ chances for career preparation/ useful knowledge or other benefits (3) what are possible benefits of obtaining these experiences/ career preparation/ useful knowledge/other ideas?	Original article is explained and developed fully with sufficient examples. Includes extended discussion of points made in the original article, either in agreement or disagreement.
	Unity of topic is maintained by eliminating unrelated material and keeping only connected ideas	Unity of topic is maintained by eliminating unrelated material and keeping only connected ideas
organization	Logical order is followed and cohesion created – either time, sequence, or order of importance of the reasons.	Logical order is followed and cohesion created – either time, sequence, or order of importance of the factors.
	Extended commentary is integrated into the paragraph as a unified part of the whole discussion and conclusion.	Extended commentary is integrated into the paragraph as a unified part of the whole discussion and conclusion.
style	Verb tense is correct and consistent. Cause and effect vocabulary structures are used.	Verb tense is correct and consistent. Cause and effect vocabulary structures are used.
	Problems with grammar and mechanics are minimal and do not distract the reader. Required document formatting used.	Problems with grammar and mechanics are minimal and do not distract the reader. Required document formatting used.
		Provides an accurate APA or MLA citation of the article.