2015

Validity Arguments for Diagnostic Assessment Using Automated Writing Evaluation

Carol Chapelle
Iowa State University, carolc@iastate.edu

Elena Cotos
Iowa State University, ecotos@iastate.edu

Jooyoung Lee
Iowa State University, jylee@iastate.edu

Validity Arguments for Diagnostic Assessment Using Automated Writing Evaluation

Abstract
Two examples demonstrate an argument-based approach to validation of diagnostic assessment using automated writing evaluation (AWE). *Criterion*®, was developed by Educational Testing Service to analyze students' papers grammatically, providing sentence-level error feedback. An interpretive argument was developed for its use as part of the diagnostic assessment process in undergraduate university English for academic purposes (EAP) classes. The *Intelligent Academic Discourse Evaluator* (*IADE*) was developed for use in graduate EAP university classes, where the goal was to help students improve their discipline-specific writing. The validation for each was designed to support claims about the intended purposes of the assessments. We present the interpretive argument for each and show some of the data that have been gathered as backing for the respective validity arguments, which include the range of inferences that one would make in claiming validity of the interpretations, uses, and consequences of diagnostic AWE-based assessments.

Keywords
automated writing evaluation, classroom assessment, computer-assisted language testing, diagnostic assessment, validity argument

Comments
This is a manuscript of an article from Language Testing 32 (2015): 385, doi: 10.1177/0265532214565386. Posted with permission.

Validity Arguments for Diagnostic Assessment Using Automated Writing Evaluation

Carol A. Chapelle, Elena Cotos, and Jooyoung Lee
Iowa State University

New diagnostic assessments of writing with automated writing evaluation (AWE) promise individualized diagnostic feedback aimed at guiding students' revision and raising their consciousness about specific aspects of their writing to help them learn. However, like any assessment, diagnostic assessments need to be evaluated in view of the validity of their intended interpretations, uses and consequences. Despite the need for validity arguments, the evaluation of AWE has tended to focus on accuracy of the system (Dikli, 2006), even if other approaches to validation have been recognized (Bennett and Bejar, 1998; Yang, Buckendahl, Juszkiewicz, & Bhola, 2002; Cotos & Pendar, 2008; Powers, Burstein, Chodorow, Fowles, & Kukich, 2001; Yang et al. 2002). In particular, Clauser, Kane and Swanson (2002) conceptualized how to frame a validity argument by identifying five types of inferences that one might wish to make on the basis of results from an assessment with automated scoring. This paper demonstrates how such inferences can be used to develop validity arguments for two AWE-based diagnostic assessments. This paper illustrates 1) the claims that users of AWE systems want to be able to make about the value of the classroom assessment and the data used to support such claims, as well as 2) the need for inferences and claims to be developed to reflect the goals of the AWE system and its use.

We provide two examples to demonstrate that the variety of approaches taken to AWE-based diagnostic assessment can be evaluated through the use of an argument-based validity framework. The first system, *Criterion*®, was developed by Educational Testing Service to analyze students' papers grammatically and provide sentence-level error feedback. This system was used as part of the diagnostic assessment in five undergraduate university English for academic purposes (EAP) classes. The second one, *Intelligent Academic Discourse Evaluator* (*IADE*), was developed for use in six graduate EAP university classes, where the goal was to help students improve the effectiveness of their discipline-specific research writing. The validation for each was designed to support claims about the intended purposes of the assessments, and therefore the first step was to make those claims explicit by formulating an interpretive argument for each. We present the interpretive argument for each and give an example of the data that have been gathered as backing for their respective validity arguments, which include a range of claims and inferences that one would make in claiming the validity of the inferences, uses, and consequences of diagnostic AWE-based assessments.

CASE 1: *Criterion®* IN UNDERGRADUATE EAP CLASSES

Criterion® was used as one means of evaluating writing tasks in English 101C, a course offered at Iowa State University which is designed to help international students to improve their academic writing skills in preparation for a first-year writing course, English 150. For each writing assignment, students were encouraged to use *Criterion* in the process of drafting and revising their essays. Three of the course objectives prompted the use of AWE in the writing process:

- Use the process of multiple drafts and feedback to revise and improve written work across several drafts of a composition
- Become independent writers who can identify weaknesses, evaluate effectiveness, and revise compositions
- Proofread, edit, and correct drafts for common errors of mechanics and word choice

This is a manuscript of an article from *Language Testing* 32 (2015): 385, doi: 10.1177/0265532214565386. Posted with permission.

In order to make explicit all that was expected of the assessment in this context, an interpretive argument was developed, as shown in Figure 1. The interpretive argument includes the inferences as outlined by Clauser, Kane and Swanson in addition to those of Ramification and Domain Definition. Ramification is critical for classroom diagnostic assessment, where we need to be able to claim that learning results from assessment use. Domain Definition is important in all assessment as illustrated by Chapelle, Enright and Jamieson (2008). The interpretive argument is developed by identifying the inferences to be made, the warrant associated with each inference and assumptions underlying each warrant. The general meaning of each of the inferences is the same across interpretive arguments, but certain warrants and assumptions are formulated to reflect the specific intended interpretations, uses, and consequences of the assessment.

Inference: Warrant *Assumptions (numbered under each warrant)*

G. Ramification: Use of the system is beneficial for learning.

1. Students perceive this new system positively.
2. The use of the system provides learning opportunities by offering feedback on relevant grammatical features in a way that complements teacher feedback.
3. The use of the system provides learning opportunities by allowing frequent access to feedback.
4. Students understand the degree to which they have mastered what was learned as well as their strengths and weaknesses.
5. Students acquire grammatical features and increase grammar awareness.
6. The use of Criterion facilitates their writing process.
7. The use of Criterion fosters autonomous learning.

F. Utilization: Diagnostic results on the quality of academic writing obtained from *Criterion* are useful for students to make decisions on revisions.

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

E. Extrapolation: Diagnostic results are relevant to academic writing context.

1. Diagnostic results from *Criterion* are related to teachers' feedback on student essays.
2. The characteristics of the assessment tasks offered via *Criterion* closely correspond to the writing tasks required in university courses.
3. The criteria and procedures of *Criterion* for evaluating the performance on the English 101C assessment tasks are pertinent to those identified by instructors as important for assessing writing assignments in mainstream university courses.

D. Explanation: Diagnostic results given by *Criterion* reflect the aspects of a construct of academic writing ability targeted by English 101C courses.

1. Diagnostic results address some of the important aspects of academic writing skills learned in English 101C courses.

C. Generalization: *Criterion* feedback is consistent (provides the same results) across the instances of the same error.

1. Writing tests provide sufficient opportunities for students to obtain frequent error feedback for learning.
2. Error recognition is consistent in identifying types of errors from student writing.

B. Evaluation: *Criterion* feedback provides students with accurate information to target relevant areas for revision/improvement/learning.

1. *Criterion* feedback is accurate.
2. *Criterion* feedback is clear and specific.

A. Domain Definition: Observations of performance via *Criterion* reveal relevant knowledge, skills, processes, and strategies representative of those required for academic writing in English 101C.

	1. Assessment tasks that require important knowledge and skills taught in English 101C can be simulated and offered via *Criterion*. 2. Assessment tasks offered via *Criterion* encourage students to follow the writing process of planning, drafting, and multiple revisions, which is emphasized in English 101C.

Figure 1: Inferences, Warrants, and Assumptions in the Validity Argument for the English 101C Writing Assessment using *Criterion*

To illustrate how support is obtained for assumptions in the validity argument, we describe the study conducted to support the warrant that diagnostic results on the quality of academic writing obtained from *Criterion®* were useful for students to make decisions about revisions. The assumption targeted in this study was F3 (in Figure 1) that students use diagnostic results to make decisions on how to revise their drafts and correct errors.

Context and Method

A classroom-based investigation was conducted as part of a larger project investigating the use of *Criterion®* to provide formative feedback and encourage multiple revisions. These classes used *Criterion®* in more or less similar ways for the same purpose, promoting a multi-stage writing process. Students were taught how to use the system and encouraged to submit drafts of their papers to obtain *Criterion®* feedback before handing it into the teacher, who would ultimately provide the final evaluation. Researchers were invited to observe this process and gather data on Criterion use, which they did by accessing the students' writing with feedback received from Criterion and comparing the original drafts and the drafts revised based on Criterion feedback.

Participants

Participants were taking required ESL writing courses, where they had been placed based on the scores they received on the English Placement Test, whose purpose it is to identify students unlikely to succeed on the writing tasks required in mainstream courses. All students had obtained a high enough score on the TOEFL IBT (71 plus a minimum score of 17 in both Speaking and Writing sections) or the IELTS (6.0 plus minimum scores of 5.5 in all subsections) to be admitted to the university. Twenty students taking four ESL writing courses, which were taught by three different instructors in Spring 2012, participated in the first study. Five students were randomly selected from each class to participate in the study and they agreed. They were international students from diverse countries such as China, Korea, India, Malaysia, Spain, and Saudi Arabia and from different disciplines including Business, Sociology, Design, Agriculture, and Engineering.

Criterion®

Criterion® is an online AWE program powered by a scoring engine called *e-rater* developed by Educational Testing Service. Unlike *e-rater* used in large-scale high-stakes exams and only providing numerical scores, *Criterion®* was designed as a classroom instructional tool, and its target audience is high school and college students. It provides not only holistic scores ranging from one to six but also diagnostic feedback on various aspects of writing including grammar, usage, mechanics, style, and organization and development. The writing tasks that *Criterion®* evaluates includes a wide range of topics as well as several different types/genres of essays from which teachers can select for writing assignments. *Criterion®* also has planning tools which provide eight different planning templates including outline, compare and contrast, cause and effect, and persuasive so that learners can outline and structure their essay before writing. In addition, students have access to a writer's handbook that provides helpful advice that can be consulted to follow up feedback received from *Criterion®* or at any time in the writing process. Table 1 shows the error comments that *Criterion®* provides to students.

Table 1. Types of error comments provided by *Criterion®*

Grammar	Fragment or missing comma, Run-on sentences, Garbled sentences, Subject-verb agreement, Ill-formed verbs, Pronoun errors, Possessive errors, Wrong or missing word, Proofread this
Usage	Wrong article, Missing or extra article, Confused words, Wrong from of word, Faulty comparisons, Preposition error, Nonstandard word form, Negation error
Mechanics	Spelling, Capitalize proper nouns, Missing initial capital letter in a sentence, Missing question mark, Missing final punctuation, Missing apostrophe, Missing comma, Hyphen error, Fused words, Compound words, Duplicate
Style	Repetition of words, Inappropriate words or phrases, Sentences beginning with coordinating conjunctions, Too many short sentences, Too many long sentences, Passive voice
Organization and development	Introduction, Thesis, Main ideas, Supporting ideas, Transitional words and phrases, Conclusion

Data Collection

The data included students' writing samples consisting of two drafts of one paper for three of the classes (5 students x 2 drafts x 3 classes = 30) and three drafts of one paper (5 students x 3 drafts x 1 class = 15) for one class along with feedback given by *Criterion®* and teachers. The number of drafts students wrote to complete an essay depended on the teacher's design of the assignment. Also, some teachers asked students to turn papers in to the *Criterion®* website while others requested them to submit them via email in a Word file. In the former case, the data were collected by accessing the *Criterion®* webpage, and in the latter case, word files were collected from the instructors. For this study, a total of 45 drafts written by 20 students were analyzed; 35 of these were revisions. All of them were personal essays in which students were supposed to write about their past experience.

Analysis

To analyze how students used the feedback from *Criterion®* to make decisions on how to revise their drafts and correct errors, Lee and Hegelheimer (2012) used coding schemes regarding revision types and success of revision, respectively. These coding schemes were developed based on Ferris (1997) but modified and refined to fit the data and context of the study, as shown in Table 2. The revision classification is divided into six categories depending on the presence and the manner (adding, deleting, changing, and transposing) of revision. The second category, *remove*, applies to a case in which a student deletes an erroneous part to change content or organization in a considerable way. An avoidance strategy was coded only when i) only the problematic (grammatically wrong) part was deleted; ii) the deletion did not lead to any improvement; and iii) no change was made in terms of content or organization.

Table 2: Coding scheme definitions and examples for revision types

Coding	Definition	Examples (Types of errors identified by *Criterion®*)	
		Draft 1	Draft 2
No change	Students did not make a change attempting to correct identified error.	So I do not like staying at home, the only thing I can feel is loneliness instead of love. (Run-on sentence)	So I do not like staying at home, the only thing I can feel is loneliness instead of love.

Remove	The part of an essay including a highlighted error was removed in the process of changing content and organization.	After classes began I planned on talking to the football coach to go to practice and hopefully get accepted in the team. (Fragment or missing comma)	Not found in the second draft
Add	Students revised the essay by adding word(s), phrase(s), or sentence(s).	And that was **first** time I feel so hopeless about my life. (Missing article)	And that was **the first** time I feel so hopeless about my life.
Delete	Students revised the essay by deleting word(s), phrase(s), or sentence(s).	Because, I believe **the English** is the most important language in the world. (Extra article)	Because, I believe **English** is the most important language in the world.
Change	Students revised the essay by changing word(s), phrase(s), or sentence(s) with alternatives.	There are full of every kinds of difficult in our lives and every person has him/her own difficult **should to undertake.** (Ill-formed verb)	There are full of every kinds of difficult in our lives and every person has him/her own difficult **should be undertaken**.
Transpose	Students revised the essay by changing the order of words, phrase(s), or sentence(s).	In my part, I feel that study abroad would be a very good and impressive experience. **Just like the coin has two sides, study abroad also has two sides.** In order to get one thing, you must to give up another thing, that what we called opportunity cost. (Self-revision)	In my part, I feel that study abroad would be a very good and impressive experience. In order to get one thing…and then you may know what you truly want! **Just like the coin has two sides, study abroad also has two sides.**

Results

Students revised their drafts in diverse ways based on *Criterion* feedback as shown in Table 3, but nearly 50% of the *Criterion®* feedback was disregarded by students. They did not change anything even when direct feedback or indirect feedback including suggestions was provided. Although the reasons for this tendency should be further investigated and confirmed by conducting student interviews, researchers observed that the frequent inaccuracy of feedback may have been one reason. For example, many proper nouns were flagged by *Criterion®* as having spelling errors, and some correctly written sentences were identified as fragments or run-on sentences. Among the revisions prompted by *Criterion® f*eedback, changing a word/phrase was the predominant type of revision.

Table 3: Revision attempts following *Criterion®* feedback (n= 20; the number of revised drafts examined = 35)

Revision Attempts	Definition	Frequency*	Example (Types of errors identified by *Criterion®*)
No change	Students did not attempt to correct identified errors.	150 (51%)	**Before**: So I do not like staying at home, the only thing I can feel is loneliness instead of love. (run-on sentence) **After**: So I do not like staying at home, the only thing I can feel is loneliness instead of love.
Remove	The part of an essay including a highlighted error was removed in the process of changing content and organization.	11 (4%)	**Before**: After classes began I planned on talking to the football coach to go to practice and hopefully get accepted in the team. (fragment or missing comma) **After**: Not found in the second draft
Add word/phrase	Students revised the essay by adding word(s) or phrase(s).	24 (8%)	**Before**: And that was **first** time I feel so hopeless about my life. (missing article) **After**: And that was **the first** time I feel so hopeless about my life.
Add sentence	Students revised the essay by adding sentence(s).	0 (0%)	None
Delete word/phrase	Students revised the essay by deleting word(s) or phrase(s).	18 (6%)	**Before**: Because, I believe **the English** is the most important language in the world. (extra article) **After**: Because, I believe **English** is the most important language in the world.
Delete sentence	Students revised the essay by deleting sentence(s).	12 (4%)	**Before**: Family for me, becoming a place just for food and sleep. (fragment or missing comma) **After**: Not found in the second draft
Change word/phrase	Students revised the essay by changing word(s) or phrase(s) with alternatives.	50 (17%)	**Before**: There are full of every kinds of difficult in our lives and every person has him/her own difficult **should to** undertake. (ill-formed verb) **After**: There are full of every kinds of difficult in our lives and every person has him/her own difficult **should be undertaken**.
Change sentence	Students revised the essay by changing sentence(s) with different contents/structures.	29 (10%)	**Before**: When I told him I was the same skinny boy that came to his office a few months ago he was astounded. (fragment or missing comma) **After**: He came up to me to offer me to practice for the team next season. Noticing he didn't recognize me, I told him who he spoke to and the coach was astounded.

Transpose word/phrase	Students revised the essay by changing the order of word(s) or phrase(s).	0 (0%)	None
Transpose sentence	Students revised the essay by changing the order of sentence(s).	0 (0%)	None
Total		294 (100%)	

* Frequency refers to the raw frequency of Criterion-identified errors across all 35 revised drafts analyzed .

Table 4 shows the frequencies of attempted revision and their success divided according to the accuracy of *Criterion®* error identification. Overall, the success of the students' revisions are related to the accuracy of the feedback. Given that the proportion of successful revision is over 70%, *Criterion®* feedback can be considered as positively influencing the revision process, even if substantial room for improvement exists.

Table 4: Success of revision efforts under conditions of accurate and inaccurate feedback

Feedback Accuracy	Revision Attempt	Success of Revision	Example (Types of errors identified by *Criterion®*)
Correct 179	Change 122	Correct 106	**Before**: Different from other parents, instead of sending their children go abroad, my mother sent me to the Mudu high school which is famous **<u>of</u>** its strict education. (Preposition) **After**: Different from other parents, instead of sending their children go abroad, my mother sent me to the Mudu high school which is famous **<u>for</u>** its strict education.
		Incorrect 16	**Before**: Without any other choices, I sat on the bench while planning my mischievous plot to kill the hour, like complaining of hurting fingers which would always bring in short but frequent breaks or using the bathroom for thirty minutes after which it would be **<u>bed time</u>**. (Compound words) **After**: Without any other choices, I sat on the bench while planning my mischievous plot to kill the hour, like complaining of hurting fingers which would always bring in short but frequent breaks or using the bathroom for thirty minutes after which it would be **<u>bed-time</u>**.
	No change 57	Correct 0	None

		Incorrect 57	**Before**: In the last few day before I go to **US**, my mother became my best friend and we love sharing dinner at one table talking everything happened that day. (Article) **After**: In the last few day before I go to **US**, my mother became my best friend and we love sharing dinner at one table talking everything happened that day.
Incorrect 115	Change 22	Correct 16	**Before**: Everyone worked hard to attach their goals, of **cause** included me. (Article: consider using "the") **After**: Everyone worked hard to attach their goals, **of course** included me.
		Incorrect 6	**Before**: Although I liked listening to music, playing piano was a different story. (Fragmented sentence) **After**: Listening to music is my favorites, however playing piano was a different story.
	No change 93	Correct 90	**Before**: When I "wander" in house for half year, I found that study seemed to be **the** most interest things. (Article: remove this article) **After**: When I "wander" in house for half year, I found that study seemed to be **the** most interest things.
		Incorrect 3	**Before**: I remember I went to book the ticket back to home on the second day after my mom gave me the call and there was always a idea in my mind that is **study** one more year and enter the next year's examination. (Article) **After**: I remember I went to book the ticket back to home on the second day after my mom gave me the call and there was always a idea in my mind that is **study** one more year and enter the next year's examination.
Total	294		

These data partially support one of the assumptions underlying the warrant that diagnostic results on the quality of academic writing obtained from *Criterion* are useful for students to make decisions on revisions. By observing the decisions about revision that students made, we were able to quantify a specific level of support for this assumption. These data do not provide a definitive answer, but rather serve as some evidence in an overall argument. Examination of the overall argument presented in Figure 1 reveals the precise place of this evidence in addition to additional needs for research.

CASE 2: *INTELLIGENT ACADEMIC DISCOURSE EVALUATOR (IADE)* IN GRADUATE EAP CLASSES

The *Intelligent Academic Discourse Evaluator* (*IADE*) program was developed at Iowa State University to provide feedback to ESL graduate students learning how to write discipline-specific research articles. It was used in a writing course designed to help students write papers required in their disciplines. The genre given the most attention is the research article. In addition to the AWE-based assessment, materials for the course consist of a corpus of research articles in 50 academic disciplines, readings, and a concordancing program. Similar to the use of the concordancer, revising with *IADE* was introduced as a learner-centered activity. Extensive data on multiple aspects of *IADE* use was gathered (described by Cotos, 2014). We demonstrate how some of these data can be used in support of a validity argument for use of *IADE* for assessment and feedback. Figure 2 outlines the interpretive argument that is the starting point for the validity argument. It consists of an eight-step argument each with assumptions

that are identified for each of the warrants. Warrant D claims that feedback provided by *IADE* on students' writing is pertinent to the quality of their research article writing. This warrant rests on the assumption that the feedback provided by *IADE* helps students to focus on how meaning is construed in research articles (D2).

Assumptions (numbered under each warrant

Inference: Warrant

H. Positive Impact: Students have a positive learning experience using *IADE*.

1. Students perceive *IADE* positively.
2. Students judge the use of *IADE* to be beneficial to their writing.
3. Students increase their metalinguistic awareness of how language is used to convey the moves needed to construct meaning in particular genres.
4. The use of *IADE* is beneficial for students' writing processes.
5. The use of *IADE* fosters autonomous learning.

G. Language learning potential: Use of *IADE* is beneficial for learning.

1. The use of *IADE* provides students with opportunities for noticing and focus on discourse form.
2. The use of *IADE* results in the improvement of the rhetorical quality of writing research articles.
3. The use of *IADE* results in learning gains.

F. Utilization: Feedback on *IADE* is useful for students to make decisions about revisions.

1. The meaning of the *IADE* feedback is clearly interpretable by students.
2. Students are willing to use *IADE i*n their writing process.
3. Students use feedback to make decisions on how to revise their drafts in order to improve the effectiveness of their discourse.
4. *IADE* provides necessary assistance beyond feedback to help revision process.

E. Extrapolation: Writing tasks and feedback are relevant to writing research articles in students' disciplines.

1. The writing tasks resemble the task of writing research articles in students' disciplines.
2. *IADE* criteria and procedures for evaluating the performance resemble those students will encounter as they write academic research articles in their disciplines after taking the class.

D. Explanation: Feedback given by *IADE* on students' writing is pertinent to the quality of their research article writing.

1. Feedback about moves in research articles within specific disciplines is relevant for students' improvement of their writing.
2. Feedback provided by *IADE* helps students to focus on how meaning is construed in research articles.

C. Generalization: *IADE* feedback is consistent in providing the same feedback across the instances of the same rhetorical move within a particular register of a research article in a particular discipline.

1. The writing tasks provide sufficient opportunities for students to obtain frequent discourse-level feedback for learning.
2. Feedback is consistent in identifying moves in students' writing of the relevant discipline.
3. Feedback is consistent in orienting students towards research article writing norms in their particular discipline.

B. Evaluation: *IADE* feedback provides students with accurate and appropriate information to target relevant areas for revision/improvement/learning.

1. *IADE* feedback is accurate.
2. *IADE* feedback is clear and specific.
3. *IADE* feedback is appropriate for targeted learners.

A. Authenticity: The writing tasks required students to produce texts conforming to the conventions of research articles within their own disciplines.

1. Research articles from the relevant disciplines had been identified and compiled in the corpus used as a basis for the writing tasks.
2. Conventions of the research article in relevant disciplines had been identified and coded in the corpus used as a basis for assessment of writing tasks.

3. Students recognize their discipline-specific genres in the writing tasks.

Figure 2. Inferences, warrants and assumptions in the validity argument for the use of *IADE* in English for Academic Purposes ESL classes for graduate students

Context and Method

The data serving as support for the assumption were gathered in a study investigating student use of *IADE* in the graduate-level ESL writing class. Pursuing a process-product research approach (Warschauer & Ware, 2006), this study employed a mixed-methods design with a concurrent transformative strategy, where quantitative and qualitative data were collected and integrated during the analysis and interpretation phases. The larger study and its rationale are described by Cotos (2014). Some of these data served well for the support of assumption D2 about the utility of the feedback for focusing students' attention on meaning.

Participants

The study was carried out with 105 participants (59 male and 46 female) ranging from 22 to 44 in age. They were international graduate students at Iowa State University (37 master's and 68 doctoral students) specializing in one of the following disciplines: Accounting, Aerospace Engineering, Agronomy, Analytical Chemistry, Animal Science, Biochemistry, Bioinformatics, Biology, Biomedical Science, Bionanotechnology, Chemical Engineering, Computer Engineering, Computer Science, Curriculum and Instruction, Economics, Electrical Engineering and Power Systems, Environmental Engineering, Ergonomics, Food and Lodging Management, Genetics, Human Health and Public Performance, Industrial Engineering, Journalism, Materials Science and Engineering, Mathematics, Mechanical Engineering, Physics and Astronomy, Plant Breeding, Public Administration, Sociology, Statistics, Textiles and Clothing, Urban and Regional Planning, and Veterinary Medicine. They represented different language backgrounds including Chinese Thai, Italian, Spanish, Korean, Portuguese, Arabic, Turkish, Telugu, Filipino, and Greek. Their level of English language proficiency ranged between 80 and 104 on TOEFL IBT, 243 and 255 on TOEFL CBT, and 520 and 667 on TOEFL PBT.

The participants were students in six sections of the English 101D writing course in the fall of 2008 and in the spring of 2009. Eighty six students were in the first year of their graduate program, of which 84 had not written a research article before enrolling in this course. Although 43 students had written research articles in their native language, only 22 had published their papers. Not all the participants had research writing experience in English; while 19 students had written research articles in English, 14 of them submitted their papers as course assignments without the intent to pursue publishing. Only 5 students had research articles published in English; however, those articles were co-authored with faculty.

Intelligent Academic Discourse Evaluator

IADE is an AWE program developed to complement writing instruction and help students practice with and make incremental improvements on their drafts of research article introductions. Since the focus of the course is on the discourse patterns and linguistic conventions of research articles, *IADE* does not detect and provide feedback on writing errors. To ensure that the program's feedback is meaningful, i.e., a "response that provides a learning opportunity for students" (Heift, 2003, p. 533), it embodies the following characteristics considered beneficial for language learning: individual-specific (Hyland, 1998), explicit (Caroll & Swain, 1993; Lyster, 1998), metalinguistic (Rosa & Leow, 2004), negative cognitive (Mitchell & Myles, 1998), detailed iterative (Hyland & Hyland, 2006), and short (Van der Linden, 1993). *IADE* generates feedback intended to serve as color-coded modified input, which is in fact analyzed student's output. The color codes are indicative of the three moves characteristic of introduction sections as described by Swales (1981, 2004). Move 1 has the purpose of establishing a territory

representing the breadth of disciplinary knowledge on the topic of the study. Move 2 identifies a niche in the existing knowledge territory. Move 3 introduces the reported study, demonstrating how it addresses the niche. Figure 3 shows how the color-coded feedback renders the rhetorical structure of an Introduction submitted by a student in Journalism. The text in blue (Move 1) provides background about international communication, national policies, and political image making, which is needed to contextualize the topic of the study. Interspersed is Move 2 in red, where the student highlights the problem that mass media could magnify certain issues to the extent of distorting the real image of a country. The student also argues that simply broadcasting positive news cannot solve this problem. The green portion (Move 3) of the introduction informs how the present study addresses this problem by focusing on measuring the impact of foreign media coverage on the effectiveness of international public relations campaign.

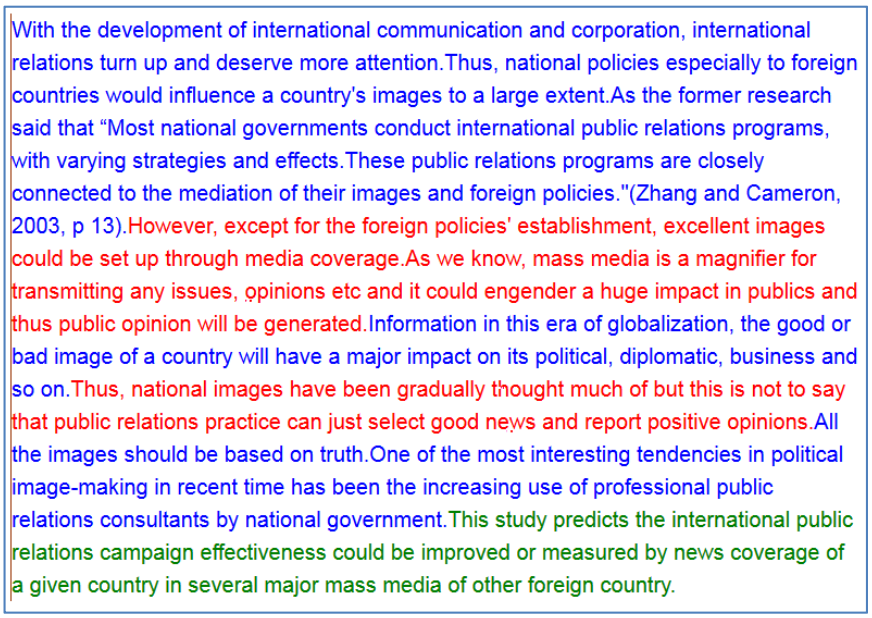

Figure 3. *IADE* color-coded feedback shown to a student who has written a draft introduction section for her research

In addition, *IADE* generates numerical feedback based on statistical analyses on the moves' lengths and distribution in the annotated corpus for each discipline. For example, the distribution of moves in Journalism is as follows:

Establishing the territory: minimum 22.22%, average 51.35%, maximum 77.78%
Identifying a niche: minimum 0%, average 19.61%, maximum 41.67%
Addressing the niche: minimum 4.17%, average 29.03%, maximum 52.63%

Therefore, the numerical feedback shown in Figure 4 indicates the distribution of moves as well as the length of the student's text compared to the respective minimum, average, and maximum percentages In Journalism. Figure 4 indicates that this student's first draft is too short compared to published introductions in Journalism (267 words versus an average of 712). More importantly, it shows that the distributions of moves 1 and 3 are different from the extent to which published authors develop these particular moves to accomplish their communicative goals. Within the length of this draft, move 2 is distributed more similarly to the introductions in the student's field. This type of feedback operationalizes the goal orientation quality propagated in formative assessment (Fisher & Ford, 1998).

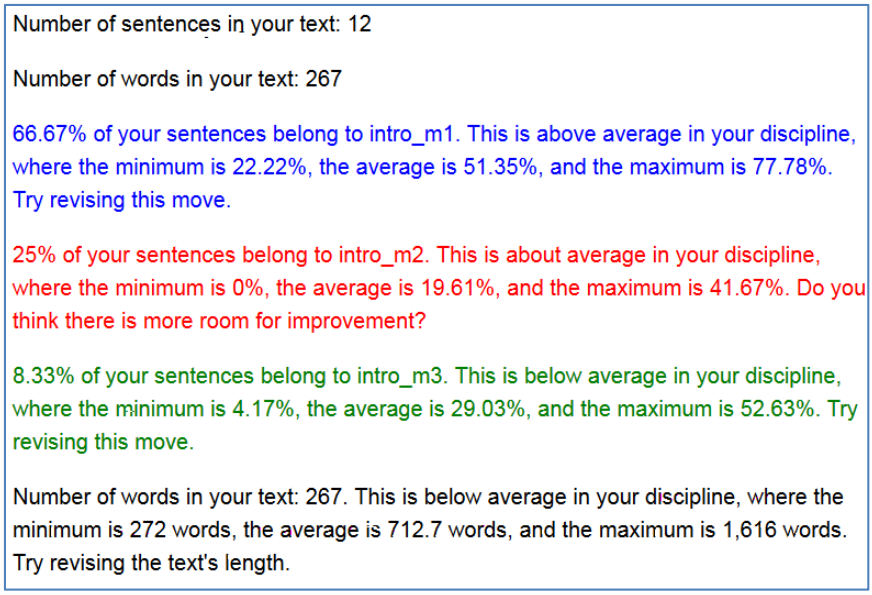

Figure 4. *IADE* numerical feedback to the writer of the introduction shown in Figure 3 for the journalism research

Data Collection

Quantitative data were gathered through 3 Likert-scale and 3 yes/no survey questions asking 1) whether the students thought about the meaning they wanted to express when they were revising with *IADE*, 2) whether they noticed that their intended meaning expressed in their move/s was reflected by a different color in the feedback, and 3) whether the feedback helped them modify their writing to better express the intended meaning. The yes/no questions were followed by open-ended questions such as "Why?", "If so, what did you think and what did you do?", and "How?", respectively. Other sources of qualitative data were think-aloud protocols, screen recordings of participants' interaction with *IADE*, semi-structured interviews, and observations elicited from a random sample of 16 participants. Camtasia Studio 5 software by TechSmith was used to screen capture their interaction with the tool, and Camtasia's audio recording function was used to record the participants thinking aloud. Concurrently conducted observations yielded notes on each participant's behavior during the interaction with *IADE* (e.g., cursor movements, verbal reactions, body language) as well as clarification questions for the semi-structured interviews.

Analysis

Percentages for Likert-scale and yes/no responses were calculated and interpreted as evidence pertaining to the assumption. The four levels of the Likert-scale response choices were interpreted as follows: "a lot" was considered as excellent evidence, "somewhat" as good evidence, "a little" as weak evidence, and "not at all" as poor evidence. Participants' responses to the open-ended questions were analyzed by identifying emerging themes, which were then quantified in terms of percentages of students who mentioned them. All the qualitative data were transcribed, and the analysis was done according to a coding taxonomy developed in view of SLA constructs and based on the results of the pilot conducted prior to this study. For coding, data were segmented into semantic "idea units" defined as "a chunk of information which is viewed by the speaker/writer cohesively as it is given a surface form […] related […] to psychological reality for the encoder" (Kroll, 1977, p. 85). A second coder was not involved since that would have required extensive resources for training; however, to estimate the

reliability of coding, the researcher coded the transcripts from the pilot study data twice with an interval of eight months (Cohen's Kappa .88), which helped confirm and refine the initial coding categories.

Results and Discussion

Evidence for the assumption about discourse meaning is summarized in Table 5. According to participants' responses to the Likert-scale questions, a total of 92% indicated that they focused on the meaning that they intended to express as they were revising with *IADE*. Only one student responded "not at all," while 81% focused on meaning "a lot," 15% – "somewhat," and 3% – "a little." Participants' answers to the survey question that inquired about their focus on the functional meaning of the moves were also more positive than negative. Of 83 respondents, 92% focused on the functional meaning of the moves, and only 8% did not think they did.

Table 5. Summary of evidence of focus on discourse meaning from five data sources

Data source	*# participants*	*Data*	*Evidence*	*No evidence*
Likert-scale q-ns	88	Q-n 8: [think about meaning]	99%	1%
		Q-n 9: [notice mismatch]	100%	0%
		Q-n 10: [modify for better meaning]	100%	0%
Yes/No & open-ended q-ns	83	Q-n 13: [reasons for focus on meaning when revising]	92%	8%
		Q-n 14: [noticing mismatch; follow-up action]	100%	0%
		Q-n 15: [role of feedback for better meaning]	76%	24%
Think-aloud protocols/Camtasia	16	16 Think-aloud/Camtasia transcripts	252 idea units	
		- Noticing a mismatch	30%	
		- Reflection on meaning	17%	
		- Connection between functional and lexical choice	38%	
		- Construction of new meaning	15%	
Semi-structured interviews	16	16 interview transcripts	54 idea units	
		- Noticing a mismatch	44%	
		- Reflection on meaning	22%	
		- Connection between meaning and lexical choice	33%	
Observations	16	16 observations	77 idea units	
		- Noticing a mismatch	32%	
		- Reflection on meaning	40%	
		- Connection between meaning and lexical choice	17%	
		- Construction of new meaning	10%	

Explanations of why the participants thought they focused on the meaning of the moves fell into several thematic categories. In their positive responses, the majority of the participants (68%) reasoned that they thought of the functions of their sentences when the color-coded feedback displayed a move-color differently than expected. For example, according to Student 42, "*This was the hard part. Because, I cannot express well what I want to express, sometime the program shows other moves instead of what I want to express. And then I think what I'm saying."* Other students (12%) explained that they had to think about the meaning of their moves in order to ensure the effectiveness of a particular communicative purpose. Student 14 explained, "*If I want reader to know that I point to gap, I think about how I should say to point to gap."* In addition, 9% realized that only by paying more attention to the functional meaning of the moves and steps they could better understand what they had been taught in class and make better corrections to their drafts. "*Actually, to know the moves you have to understand how they work, not only know them from lessons. That's why I thought of move meaning,"* wrote Student 7. No follow-up explanations were by provided by 10%.

The role of the color-coded negative feedback became more evident when the students explained that they noticed having miscommunicated functional meaning due to a mismatch between intended and conveyed meaning displayed by *IADE*'s colors. According to the open-ended survey data, 92% noticed that *IADE* color-coded feedback displayed their moves in a color different than what they had in mind (see Table 5). The Likert-scale data supported this finding with evidence that 44% noticed such a mismatch "a lot," 41% – "somewhat," and 14% – "a little." Explaining how they reacted to such feedback, the participants commented on their thoughts (40%) and actions (60%) at the moment. Their thoughts included self-questioning as to what might have caused the mismatch (12%), self-verification as to whether *IADE* was wrong in identifying the move and whether they were right in expressing its function (23%), and self-planning speculations as to what should or should not be done when expressing the meaning of a given move (5%). Participants' actions upon receiving color-coded feedback, which was in disagreement with their communicative intent, consisted of immediate attempts to modify their output (19%), consult the help options in *IADE* (13%), and search for move-specific phraseology (28%), which in fact also resulted in output modification. These results suggest that in their attempts to modify their output, the learners were discovering a connection between certain vocabulary and the functional meaning of moves and steps and, therefore, directed their attention to key words and expressions indicative of such meaning. For instance, Student 8 wrote in the survey, "*I realized that one word could change my thoughts. So, as I want to maintain my ideas of movements, I changed some expression to convert the sentences into other move or step*." Student 12 implied a similar idea saying, "*I was trying to change and insert specific words because I realized that my steps depends on the right 'words.'*"

As shown in Table 5, in response to the question about noticing mismatch, all the participants reported that the feedback helped them focus on discourse meaning to a certain extent: 30% thought it helped "a lot," 54% thought it "somewhat" helped, and 16% thought it helped "a little." They also reflected about the role of the feedback on their revisions. "Yes" answers, meaning that the feedback helped them better express the intended functional meaning, were provided by 76%; "No" by 18%; and 6% were not sure. For the latter two groups, the feedback was not entirely helpful in this respect because it only pointed to miscommunicated functional meaning, without providing a specific direction for remediation. The first group, on the other hand, believed that the feedback helped them in a number of ways. For instance, it made them think about what they were trying to express (13%), it made them focus on negative evidence making them aware of a mismatch between intended and expressed functional meaning (30%), and, similar to a perception mentioned above, it helped them identify a connection between functional meaning and specific lexical means of expressing it (57%).

Furthermore, all the introspective data sources contained evidence of phenomena that were identified in participants' perceptions, indicating how focus on functional meaning was manifested. Table 5 presents percentages for such themes as noticing a mismatch between intended and expressed

functional meaning, reflection on functional meaning, connection between functional meaning and lexical choice, and construction of new functional meaning, which emerged from the transcripts. More specifically, the students focused on meaning by noticing that in some cases the intended functional meaning was displayed with the color of a different move (30%, 32%, 44% in think-aloud/Camtasia, observation, and interview transcripts, respectively). Student 27 said while thinking aloud, "*The last sentence is recognized as m1, but I meant it as m3."* Having received the color-coded feedback marking their moves, the participants took time to reflect about how effectively or ineffectively they had expressed the intended meaning (17%, 40%, 22%). For example, as noted in the observation, Student 58 "*Looked back at his sentence and decided that it sounds like m1 Topic Prominence (Centrality)"*. In addition, the students seemed to realize that the effectiveness of expressing functional meaning is directly related to certain lexical choices (38%, 16%, 33%) and, therefore, further modified their output in view of the move-specific phraseology that they could find in *IADE*'s Help Options. Think/aloud Camtasia transcripts contain 15% of such instances, which are similar to the one for Student 65: "*[goes down to the revision box. highlights "may also" and changes it to "may contribute to"] I changed just ... m2. I think it maybe will let the problem change it into m3."*

It is worth noting here that noticing a mismatch between intended and expressed meaning appeared to lead to reflection on functional meaning, and that making connections between functional meaning and lexical choice led to the construction of new discourse meaning. These two patterns emerged from statements like the ones below, where the arrow represents the transition from one process to another. In the first example, it is from noticing a mismatch to reflection; in the second example, from connecting meaning with particular linguistic means to output modification.

> *He still sees a problem. Part of m3 appears as m1. → He thinks that maybe it sounds as Review (which it does). (Student 40, observation)*
>
> *Now I found that I add one sentence for m2, but without the "although" my previous sentence changed to m1. → I'll add "although" because I want to show contrast. (Student 32, think-aloud/Camtasia)*

The semi-structured interviews provided additional insights into learners' focus on meaning. One such insight is that the students did not think of meaning when they started revising their draft; they seemed to focus more consistently on functional meaning later in the revision process, once they realized the importance of functional meaning because of discrepancies displayed by the color-coded feedback. Another observation is that, initially, noticing a mismatch in the colors was mostly accidental, and, as the revision continued, it became intentional. In other words, the learners began to intentionally verify their sentences to see if they are displayed in "the right color" (Student 29).

A less encouraging insight is related to discovering the lexical realizations of move/step functions. Nine out of 16 students mentioned that once they realized that certain words can help them build certain moves, they tended to rely more on making lexical modifications. Along these lines, Student 64 said in his interview, "Yes, I used those words there because they work. Actually, this is good that I know that because after that I changed many words many times." Indeed, developing awareness that certain functional meanings have certain linguistic realizations is a positive thing; however, learners' tendency to replace some vocabulary items with others, making this their main revision strategy, is limiting and undesirable. The overwhelming findings from these data, however, provide support for the assumption that the feedback provided by *IADE* helps students to focus on how meaning is construed in research articles.

CONCLUSION

These examples of AWE-based assessments were used to demonstrate the types of claims and inferences that one might hope to be able to make in support of their use for diagnosis, feedback and

learning. They show how the concepts and methods of validity argument help to transcend issues of accuracy and efficiency that typically preoccupy evaluators of such systems (Chapelle & Douglas, 2006). The two AWE-based assessments provide different types of diagnostic information, feedback, and learning for two types of EAP writing classes. They underscore the fact that AWE is not a monolithic assessment type that can be evaluated based on a single metric or set of metrics. The interpretive arguments we developed for the two systems illustrate points of similarities and differences between these two, but other AWE systems used in different contexts would need to be evaluated on the basis of specific interpretive arguments depending on the objectives of those assessments.

The interpretive argument accomplishes three objectives for AWE-based diagnostic assessments First, it provides a framework for classroom assessment in a manner that is consistent with validation of other types of assessments, thereby situating classroom assessment in a domain where rigorous and useful evaluation methods can be employed. Such rigorous methods are important for the use of AWE-based assessments because of their sophistication and the potential systematicity of their application (in contrast to the more idiosyncratic nature of oral teacher-based assessment in the classroom). Second, it provides a framework for planning research relevant to the claims that are made about the use of an AWE system. In view of the quantity of performance data that can come from the use of AWE in classroom-based assessment, a means of organizing and prioritizing research questions and data collection is essential. Third, it provides a means of organizing the presentation of research to allow for critical review about the points of weakness as well as the strength of evidence supporting interpretations and use of AWE in particular contexts. If applied linguists are to increase their understanding of the factors involved in successful classroom assessments, a means of accumulating knowledge from various research contexts and synthesizing that knowledge is needed. Interpretive arguments that state explicitly the inferences, warrants, assumptions, and backing appear to offer promise for achieving this goal. It is a goal relevant for a growing number of diagnostic assessment that are available to teachers because of the widespread use of technology for enhancing other classroom materials (e.g., Jamieson, Grgurovic, & Becker, 2008).

References

Bennett, R. E., & Bejar, I. I. (1998). Validity and automated scoring: It's not only the scoring. *Educational Measurement: Issues and Practice*, Winter, 9–17.

Beuningen, C. V. (2010). Corrective feedback in L2 writing: Theoretical perspectives, empirical insights, and future directions. *International Journal of English Studies, 10*(2), 1-27.

Carroll, S., & Swain, M. (1993). Explicit and implicit negative feedback: An empirical study of the learning of linguistic generalizations. *Studies in Second Language Acquisition*, *15*, 357–366.

Chapelle, C. A., & Douglas, D. (2006). Assessing language through computer technology. Cambridge: Cambridge University Press.

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

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

Cotos, E. (2014). *Genre-based automated writing evaluation for L2 research writing: From design to evaluation and enhancement*. Basingstoke, UK: Palgrave Macmillan.

Cotos, E., & Pendar, N. (2008). Automated diagnostic writing test: Why? How? In C. A. Chapelle, Y.-R. Chung & J. Xu (Eds.), *Towards adaptive CALL: Natural language processing for diagnostic language assessment* (pp. 65-81). Ames: Iowa State University.

Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning, and Assessment*, *5*(1), 3-35.

Ferris, D. R. (1997). The influence of teacher commentary on student revision. *TESOL Quarterly, 31*, 315–339.

Fisher, S. L., & Ford, J. K. (1998). Differential effects of learner effort and goal orientation on two learning outcomes. *Personnel Psychology, 51*, 397–420.

Han, N., Chodorow, M., & Leacock, C. (2006). Detecting errors in English article usage by non-native speakers. *Natural Language Engineering, 12*(2), 115-129.

Heift, T. (2003). Multiple Learner Errors and Meaningful Feedback: A Challenge for ICALL Systems. *CALICO, 20*(3), 533-549.

Hyland, F. (1998). The impact of teacher written feedback on individual writers. *Journal of Second Language Writing*, *7*(3), 255-286.

Hyland, K., & Hyland, F. (Eds.). (2006). *Feedback in Second Language Writing: Contexts and issues.* New York: Cambridge University Press.

Jamieson, J., Grgurovic, M., & Becker, T. (2008). Using diagnostic information to adapt tradition-al textbookbased instruction. In C. A. Chapelle, Y.R. Chung, & J. Xu (Eds.), Towards adaptive CALL: Natural language processing for diagnostic language assessment (pp. 25-39). Ames, IA: Iowa State University.

Kroll, B. (1977). Combining ideas in written and spoken English: A look at subordination and coordination. In E. Ochs & T. Bennett (Eds.), *Discourse across time and space* (Southern California Occasional Papers in Linguistics, 5). Los Angeles, University of Southern California.

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

Lee, J., & Hegelheimer, V. (2012, August). *A hybrid use of Criterion® and teacher feedback in process writing.* Paper presented at EUROCALL conference, Gothenburg, Sweden.

Long, M. (1996). The role of the linguistic environment in second language acquisition. In W. Ritchie & T. K. Bhatia (Eds.), *Handbook of language acquisition: Vol. 2. Second language acquisition* (pp. 413-468). San Diego, CA: Academic Press.

Lyster, R. (1998). Negotiation of form, recasts, and explicit correction in relation to error types and learner repair in immersion classrooms. *Language Learning*, *48*, 183–218.

Mitchell, R. & F. Myles (1998). *Second Language Learning Theories*. London: Arnold.

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2001). Stumping e-rater: Challenging the validity of automated scoring. (GRE No. 98–08). Princeton, NJ: Educational Testing Service.

Rosa, E., & Leow, R. (2004). Awareness, different learning conditions, and second language development. *Applied Psycholinguistics*, *25*, 269–292.

van der Linden, E. (1993). Does feedback enhance computer-assisted language learning. *Computers & Education, 21*(1-2), 61-65.

Warschauer, M., & Ware, P. (2006). Automated writing evaluation: defining the classroom research agenda. *Language Teaching Research, 10*(2), 1–24.