Towards Effective Integration and Positive Impact of Automated Writing Evaluation in L2 Writing

Elena Cotos
Iowa State University, ecotos@iastate.edu

Follow this and additional works at: http://lib.dr.iastate.edu/engl_pubs

Part of the Bilingual, Multilingual, and Multicultural Education Commons, Curriculum and Instruction Commons, Educational Assessment, Evaluation, and Research Commons, and the Educational Methods Commons

The complete bibliographic information for this item can be found at http://lib.dr.iastate.edu/engl_pubs/60. For information on how to cite this item, please visit http://lib.dr.iastate.edu/howtocite.html.

This Book Chapter is brought to you for free and open access by the English at Iowa State University Digital Repository. It has been accepted for inclusion in English Publications by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Towards Effective Integration and Positive Impact of Automated Writing Evaluation in L2 Writing

Abstract
The increasing dominance of English has elevated the need to develop an ability to effectively communicate in writing, and this has put a strain on second language education programs worldwide. Faced with time-consuming and copious commenting on student drafts and inspired by the promise of computerized writing assessment, many “educational technology enthusiasts are looking to A WE [automated writing evaluation] as a silver bullet for language and literacy development” (Warschauer & Ware, 2006, p. 175). This chapter reviews what AWE offers for learners and teachers and raises a number of controversies regarding AWE effectiveness with the underlying message that clear milestone targets need to be set with respect to AWE development, implementation, and evaluation in order to ensure positive impact of this technology on L2 writing. In support of this message, the chapter introduces an example-lADE, a prototype of contextbased AWE conceptualized and operationalized to address latent issues through a synthesis of theoretical premises and learning needs. Multifaceted empirical evaluation of lADE further provides insights into processes triggered by interaction with AWE technology and foregrounds a call for future research needed to inform effective application of AWE in L2 writing classrooms.

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

Comments
This is a book chapter from Technology Across Writing Contexts and Tasks (2012): 81. Posted with permission.
Chapter 5

Towards Effective Integration and Positive Impact of Automated Writing Evaluation in L2 Writing

ELENA COTOS
Iowa State University

Abstract
The increasing dominance of English has elevated the need to develop an ability to effectively communicate in writing, and this has put a strain on second language education programs worldwide. Faced with time-consuming and copious commenting on student drafts and inspired by the promise of computerized writing assessment, many “educational technology enthusiasts are looking to AWE [automated writing evaluation] as a silver bullet for language and literacy development” (Warschauer & Ware, 2006, p. 175). This chapter reviews what AWE offers for learners and teachers and raises a number of controversies regarding AWE effectiveness with the underlying message that clear milestone targets need to be set with respect to AWE development, implementation, and evaluation in order to ensure positive impact of this technology on L2 writing. In support of this message, the chapter introduces an example—IADE, a prototype of context-based AWE conceptualized and operationalized to address latent issues through a synthesis of theoretical premises and learning needs. Multifaceted empirical evaluation of IADE further provides insights into processes triggered by interaction with AWE technology and foregrounds a call for future research needed to inform effective application of AWE in L2 writing classrooms.

1. Automated Writing Evaluation for Learning and Teaching

Automated writing evaluation (AWE), defined as “the ability of computer technology to evaluate and score written prose” (Shermis & Burstein, 2003, p. xiii), is informed by educational measurement, computational linguistics as well as cognitive science and pedagogy. In other words, psychometric evaluations of reliability and validity, considerations about intelligent operational systems and their functionality as well as models that reflect thought processes and factors considered to be most beneficial for learners have all contributed to the development of AWE systems.
This technology originated from automated scoring engines and was initially referred to as computerized essay scoring, computer essay grading, computer-assisted writing assessment, or machine scoring of essays. The first scoring system, Project Essay Grade (PEG), was developed in the 1960s and employed multiple regression analysis to predict scores based on measurable variables in the form of surface linguistic features (Page, 1994). The systems following PEG—Intelligent Essay Assessor (IEA), Electronic Essay Rater (e-rater), Conceptual Rater (c-rater), Schema Extract Analyze and Report (SEAR), Paperless School freetext Marking Engine (PS-ME), Automark, and AntMover—assess written constructed responses using natural language processing (NLP) in combination with statistical techniques that analyze a wide range of aspects of writing constructs such as grammar, syntactic complexity, mechanics, style, topical content, content development, deviance, and so on. Before long, the systems were reconfigured to generate intelligent feedback on all these features.

These systems led to the development of AWE products like the pioneering Writer’s Workbench (MacDonald, Frase, Gingrich, & Keenan, 1982) and leading-edge programs like Criterion, MY Access! (for further information on this writing program see Ware & Rivas’s chapter in this volume), WriteToLearn, Summary Street, and Holt Online Essay Assessor. These commercially available products are being increasingly used in writing classrooms, shifting the role of AWE from pure assessment to assistance for learning (Chen & Cheng, 2008; Warschauer & Grimes, 2008) by offering both automated feedback and a wide range of complementary tools and features intended to help students (Burstein, Chodorow & Leacock, 2004; Pearson Education, 2007; Vantage Learning, 2007). In Criterion, for example, the students can solicit and receive feedback from their teacher through the program’s interface. This helps them to focus not only on automatically detectable errors, but also on other, more subtle, aspects of writing identified by the teacher. Students can also view their performance summary, which includes a holistic score, the number of errors, and links to detailed feedback on each error category. In addition, Criterion has a context-sensitive Writer’s Handbook that provides additional definitions and lessons. To assist students in their planning process, this program also offers a ‘Make a plan’ tool with a choice of eight templates for planning strategies. MY Access!, in turn, has an online writing coach which evaluates student writing and provides revision goals and remediation activities for each of the writing traits, as well as an editor which highlights errors and provides editing suggestions. It also offers a writer’s checklist for guidance, scoring rubrics for self-assessment, word banks for appropriate vocabulary use, and graphical prewriting tools for better formulation and organization of ideas. WriteToLearn has similar options; plus, it allows students to hear the text in read-


2 NLP is a branch of Artificial Intelligence.

3 For comprehensive reviews of these automated scoring systems see Chapelle and Chung (2010), Dikli (2006), Phillips (2007), and Valenti, Neri, and Cucchiarelli (2003).
ing passages through text-to-speech technologies and to see the translation of words and their dictionary definitions in on-demand pop-up windows.

A number of different options have been designed for teachers as well. Teachers can set certain parameters for their assignments and either select topics from the system's library or create their own. In MY Access!, teachers can choose from a large pool of over 700 prompts, including narrative, persuasive, informative, literary, and expository essay topics. WriteToLearn has a comprehensive database of more than 300 reading passages in science, history, language arts, social studies, and fiction, targeted at students in grades 4 through 12. Additionally, these programs have powerful data-analysis and reporting features that allow for organizing student essays, storing them for quick retrieval, and generating error reports. The error reports can be set at the level of performance summary, group comparison, student history, and so on, thereby allowing teachers to monitor student progress individually and as a class. Structured online training modules are available to ensure better instructor experience.

2. Panacea or Fallacy?

Despite the appeal of automated feedback on a wealth of writing traits and of the plethora of additional features, the use of AWE is being heavily debated. Supporters argue that the ability of AWE programs to assess and respond to student writing as well as humans in a much more time and cost-effective way is an immense advantage (Attali & Burstein, 2006; Pearson Education, 2007; Vantage Learning, 2007). AWE is believed to motivate and guide student revision and to foster learner autonomy. The feedback provided by AWE programs is meant to support process-writing approaches that emphasize the value of multiple drafting through scaffolding suggestions and explanations. The integration of AWE programs into the curriculum is said to also be consistent with the drive toward individualized assessment and instruction. The vendors of these programs promote them as instructional supplements to process-writing instruction and as vehicles of consistent writing and evaluation across the curriculum.

The truth of these claims has been questioned by an opposing camp in the academic community, and a great deal of skepticism has been expressed especially with respect to the consequences of AWE classroom implementations (Ericsson & Haswell, 2006). Cheville (2004), for instance, takes a very critical stance towards AWE. She is concerned that "early acculturation to such a program might undermine the language and learning of students" (p. 48). She also fears that "automated scoring technologies make it possible to eliminate the evaluative influence of teachers altogether," and that "what we know and what we do in the classroom is in jeopardy" (p. 49). Another apprehension is that, although it is unlikely that students can trick AWE software, they may consciously or unconsciously adjust their writing to meet its assessment criteria. Moreover, teachers may feel pressured to support such adjustment in an attempt to raise test scores (Conference on College Composition and Communication, 2006) which would in turn impair their practice. In addition, questions have been raised from a theoretical point of view.
Some contend that the social and communicative dimensions of writing are not supported in AWE systems since they are grounded in a cognitive information-processing model (Ericsson, 2006).

The overall arguments for or against implementing AWE in writing instruction express significant concerns related to impact, that is, whether AWE may cause positive or negative effects on teaching and learning. However, not much consideration is given to ecology, a concept used to describe phenomena in their context and to understand both the context and the interactions that create that context (Garner & Borg, 2005). Another problem is that this debate centers on hypothetical claims which are sound and reasonable but not fully supported empirically. Much research has been decontextualized and psychometrically driven, reporting on AWE validity determined by comparing automated and human scores (see Keith, 2003; Warschauer & Ware, 2006)—a necessary but not only prerequisite attribute for valid interpretations of AWE assessment outcomes. Scholars have called for contextualized studies on the “how” rather than the “if” of AWE integration (see Chen & Cheng, 2008; Grimes & Warschauer, 2010; Warschauer & Ware, 2006).

3. Still a Conundrum

To date, few studies have responded to the call to investigate how context-related factors may evoke one or another kind of AWE effect or how they may influence teachers and learners to use AWE in ways that promote or inhibit learning. Most AWE research has examined effectiveness through the lens of perceptions and outcomes. Because of the paucity of research on AWE in L2 contexts, discussion here includes work with both native and nonnative speakers of English.

3.1 AWE with English Speakers

A number of studies focused on whether AWE programs can facilitate improvement in L1 writing. Having conducted four studies on the use of MY Access! by 5th-11th graders over a period of 6 weeks to a full academic year, Elliot and Mikulas (2004) reported that student writing skills, as measured by performance on statewide writing assessments, were significantly improved after submitting on average four essays and revising them 2-5 times. Their survey results indicated that the students were highly satisfied with the automated feedback and judged it as both helpful and accurate. In Foltz, Laham, and Landauer (1999), students used WriteToLearn to revise their essays in an undergraduate course, and their improvement in scores ranged from 0 to 33 points over an average of three revisions. Attali (2004) investigated how Criterion was used nationwide by 6th-12th graders throughout a school year and found an increase in scores for essays submitted more than once and a significant decrease in error rates, although the revisions were made mainly in spelling and grammar and less so in structure. Similarly, in Leah Rock’s (2007) study, 9th graders who used Criterion for 4 weeks received higher analytic scores on their essays written at the end of the study period and improved the mechanical aspects of their writing. Criterion also reportedly facili-
tated the writing improvement of criminal justice students studied by Schroeder, Grohe, and Pogue (2008) and helped them understand their errors. Here, the final grades of two out of three experimental groups were significantly higher than those of the comparison group. The results of a regression analysis in the study suggested that the participants' grades improved by one-third of a point with each submission to Criterion.

Warschauer and Grimes (2008), on the other hand, presented evidence that is not so supportive of the effectiveness of Criterion and MY Access! Although both of these programs had a positive impact on students by motivating them to write and revise, no significant improvement was found, and the revisions focused mainly on spelling, punctuation, and grammatical errors, similar to Attali's (2004) findings. When Shermis, Burstein, and Bliss (2004) compared the performance of high school students who were randomly assigned to either a treatment group or a control group, they also found no significant differences between the two groups.

3.2 AWE with English Learners

Research on the use of AWE with L2 learners can be considered in its infancy. The few existing studies are narrower in focus and less sound in the quality of their research methodology, targeting small samples of participants and concentrating mostly on perceptions. For instance, Fang (2010) reported somewhat ambiguous findings regarding the attitudes of 45 low-proficiency EFL learners towards MY Access!'s feedback. While the learners perceived automated feedback as having a positive effect on their writing skill development, most indicated that they would still prefer teacher feedback. Along the same lines, Yang (2004) reported that while most of her 300 participants valued the instant feedback of this program, they complained about it being fixed, repetitive, and unspecific compared to teacher feedback. Repetitiveness as well as complexity of various functions of MY Access! was also marked as an issue by Lai (2010).

Two other studies attempted to evaluate AWE effectiveness though comparisons with other computer software. Tsou (2008) compared MY Access! with Microsoft Word based on posttest scores, a student questionnaire, and teacher interviews; and Yeh, Liou, and Yu (2007) used a questionnaire and self-reported checklists to compare improvement in student writing completed with MY Access! and with a bilingual concordancer. Given that these writing tools are intended for different purposes—MY Access! for the use of automated feedback, the concordancer for queries of word usage, and Microsoft Word for word processing—claims of helpfulness of AWE feedback in these studies are difficult to justify, to say the least.

Another comparative study is that of Lai (2010). Although the data were limited to 22 questionnaires and interviews, Lai's study is more insightful because it compared the effectiveness of automated feedback by MY Access! to that of peer feedback, which alternated over a period of 16 weeks. Lai found that Taiwanese college learners of English made more revisions on their descriptive and narrative essays based on peer feedback, paying special attention to the categories of focus and meaning. Overall, peer feedback was preferred over the automated feedback.
and led to greater improvement. Interestingly, the participants indicated that they would have been more motivated had the peer feedback been scored. Perhaps that is why they used the automated feedback more frequently to revise content and development, but this question remains to be investigated.

Similar findings were presented by Chen and Cheng (2008), who, like the others reviewed above, examined the use of MY Access! with Taiwanese college learners of English. Their research employed a more complex methodology, using questionnaires, focus group interviews with the students, individual interviews with the instructors, students' writing samples, and MY Access!’s scores and feedback data. Their findings highlight major differences in students’ perceptions of AWE usefulness and instructors’ implementation choices. Most importantly, though, their findings reveal the role of the context and the human factor, showing how pedagogical practices can affect students’ attitudes towards AWE and, consequently, its perceived effectiveness.

An ongoing longitudinal mixed-methods study of Criterion by Hegelheimer, Dursun, Li, Saricaoglu, and Yang (2011) reports positive preliminary findings about both students’ and teachers’ perspectives. The students expressed confidence in using the program and satisfaction with its feedback, explaining that it helped them better understand their weaknesses in writing. At the same time, they had negative comments on fixed feedback or unclear suggestions, which is a drawback that has also been noted for MY Access! (Lai, 2010; Yang, 2004). Hegelheimer et al. are also examining metalinguistic development since their students self-reported better ability to find and correct errors after having used Criterion. Their teachers’ use of Criterion varied in scope and approach. Some used it more extensively for a variety of activities (e.g., prewriting strategy training, peer review, grammar lecturing, and grading) while others used it mainly for error correction. Classroom and teacher observations indicate that Criterion may influence instructors’ teaching approach.

4. Potential Effectiveness Issues

While previous research has been ambivalent, not lending clear support to the effectiveness of AWE and not directly addressing the influence of ecology-related factors, it does provide some insights that allow for reflection on potential issues likely to surface when this technology is used in different L2 contexts. These issues may include inappropriate AWE implementation, users’ different levels of technology and language skills, and, quite importantly, unfounded transfer of automated scoring across contexts and purposes.

It is absolutely imperative to distinguish between AWE and “its scoring technology ... because they serve different purposes” (Grimes &Warschauer, 2010, p. 5). Automated scoring engines are used in summative assessment to measure writing proficiency for the purpose of grading, certification, or accountability. The evaluation capabilities of AWE programs, on the other hand, are intended for formative use during instruction to help identify learners’ difficulties and monitor the learning process. The problem is that AWE has been implemented based merely
on availability, functionality, or practicality without regard to these significant differences in purpose. The scoring ability of AWE technology has been transferred to learning contexts with the assumption that it is suitable for high-quality formative assessment because of its ability to inexhaustibly provide feedback in response to student writing. However, it was initially developed for summative assessment rather than for instructional purposes, and, as a result, the feedback is nothing but a direct output of the scoring system. Such feedback does not necessarily suit all pedagogical purposes (Warschauer & Ware, 2006).

Due to such unfounded transfer, AWE programs are often misused, as tangentially documented in much of the research discussed above. For instance, some teachers used the AWE program as an explicit form of test preparation (Warschauer & Grimes, 2008). One particular teacher simply adapted the program to her usual nonprocess-writing approach for its scoring capabilities, not its feedback capabilities. She was convinced, based on her experience, that “it doesn’t matter if they [students] get a lot of feedback” (Warschauer & Grimes, 2008, p. 32), thus obviating the need for much revision at all. Warschauer and Grimes also uncovered a few paradoxes. First, their participating teachers valued revision, but scheduled little time for it. It is not surprising, therefore, that students focused on grammar and mechanics; to revise at more complex levels, the students need to become more deeply cognitively engaged which requires sufficient time to make revisions with the AWE tool. The authors explain that providing limited revision time is consistent with more general practices in public schools, where revision “invariably focuses on a quick correction pointed by the teacher or peer” (p. 29).

Second, teachers rarely used AWE in their classes, although they expressed very positive views about the programs. The teachers explained that this was because much of the curriculum was in reading and language arts and not in composition. Another reason was that some writing tasks did not fall within the range of the genres that come with the program (e.g., newspaper articles, brochures, letters, etc.).

Considering this, Warschauer and Ware (2006) argue that “any classroom innovation, and especially those using technology, will likely have its best effect if it is fully integrated into instruction” (p. 169). “No matter how much teachers claim that they like a type of software ... if they find various reasons not to use the software, it cannot be expected to have much impact” (Warschauer & Grimes, 2008, p. 28). Indeed, Shermis et al. (2004) speculated that their discouraging results were partly due to poor implementation and high attrition, with only 112 of the 537 treatment students completing all the essays. They also estimated that if students had completed five more writing assignments each, their performance would have significantly improved. This may also be the case in Attali (2004) and Warschauer and Grimes (2008), where the students did not exploit the revision capabilities of the AWE programs; 71% and 72% of the students, respectively, submitted their essay to the system only one time and made no revisions.

In contexts where AWE implementation is more rigorous, the technology appears to benefit both students and teachers. Chen and Cheng (2008) suggest that when MY Access! is used at earlier stages of drafting and revision, with teacher
and peer feedback provided at a later point of the writing process, the students perceive it more favorably. Hegelheimer et al. (2011) are finding that, when the teachers adopt Criterion for formative assessment, they see it as a suitable tool despite differences in the way they integrate it. Moreover, some teachers explain that Criterion's detailed feedback on grammar and mechanics saves them class time and allows them to focus on more complex issues like content development. The effectiveness of AWE may be affected not only by teachers' practices and attitudes, but also by the users' technological literacy and knowledge of specific AWE features. For instance, according to Warschauer and Grimes (2008), teachers with relatively little computer experience were more reluctant to use AWE programs in their classes, while those with better computer skills were more enthusiastic. According to Hegelheimer et al. (2011), lack of familiarity with certain program settings increased the workload of one teacher, who did not know how to save comments on previous writing assignments for students' new submissions. Although learner and teacher training prior to AWE use is a key step towards increased effectiveness, few publications address this issue or at best only briefly mention it.

Furthermore, student characteristics such as literacy and language proficiency may be an additional issue. Unlike higher level students, students performing below grade level find it difficult to understand the automated feedback other than its most basic aspects (Chen & Cheng, 2008). That is why Chen and Cheng argue that implementing AWE with minimum teacher facilitation can result in frustration and negatively affect learning. Also, learners with different levels of language proficiency may benefit differently from interaction with AWE. Yang (2004) concludes that more advanced L2 learners are likely to see less value in automated feedback compared to less advanced students.

Finally, but perhaps most importantly, the conceptual design of AWE programs needs to be revisited because a number of limitations of the programs are bound to affect their effectiveness. First, if AWE programs are to be used by L2 learners, their design should integrate theoretical perspectives about second language acquisition (SLA) as well as targeted learner needs. It is arguably a problem that AWE programs, and their scoring systems in particular, were initially designed for native speakers and then used with language learners without grounding in SLA. Second, it is of concern that the machine analyzers are calibrated to static compositional features and formulaic expressions, potentially subordinating meaning (Cheville, 2004), and that the form-focused feedback is not meant to direct learners' attention to the meaning of the written discourse. AWE programs need to extend their analysis and feedback capabilities to include a substantial emphasis on the contextual richness and functional meanings of the discourse. Third, the analysis and feedback generated by AWE systems are limited to one genre—the prototypical five-paragraph essay—which is not always the goal of writing. In different educational contexts, L2 writers are expected to produce other genres; to master those genres, they would benefit from genre-specific feedback pointing to rhetorical shifts that are more sophisticated and subtle than thesis and topic sentences.
5. Towards Effective AWE: The Case of IADE

This section exemplifies how the issues enumerated above, as well as some aspects of the AWE debate, can be approached through context-specific AWE development and empirical evaluation. It describes the needs of a targeted group of L2 writers in a specific instructional context to further explain how the design of the AWE program presented here accounts for potential effectiveness caveats by integrating relevant theoretical perspectives with teaching and learning goals.

5.1 Targeted Learners and Learning Needs

Similar to many universities in English-speaking countries, Iowa State University (ISU) is home to more than 2,400 international students from more than 100 countries. For them, becoming successful academic writers is crucial, and mastering a particular genre—the research report in English—is a precursor to their academic success, admission to their disciplinary discourse communities, and potential contribution to the body of theoretical and empirical knowledge in their field. Writing up research requires a specialized literacy that consists of the ability to use the discipline’s symbolic resources for developing a scientific argument by routinely connecting rhetorical purposes with features of the text (Berkenkotter, Huckin, & Ackerman, 1991; Hyland, 2000). However, acquiring such literacy is a challenging process for L2 learners due to their complex writing difficulties, the most consequential ones being caused by inappropriate rhetorical choices (Jenkins, Jordan, & Weiland, 1993; Thompson, 1999) and lack of genre awareness (Dong, 1998). An even greater challenge is writing for the discipline, that is, following discipline-specific writing norms expected of the genre.

5.2 The Specific L2 Writing Context

Coping with genre and discipline specificity is an area that falls beyond the expertise of most L2 writing teachers. In an advanced graduate-level academic writing course at ISU, we have been addressing this difficulty through a corpus-based pedagogy, employing computerized tools for identifying the structural and linguistic features of research articles. Over the years, the course has become largely learner oriented, with students conducting corpus analyses in order to identify and learn about the writing conventions of this genre in their particular disciplines. The class work has been organized in a way that combines top-down and bottom-up approaches to corpus analysis (see Cortes, 2007). The top-down approach is realized through deductive activities based on analyses of texts for moves (or rhetorical units) and steps (or rhetorical functions) reflective of the genre (see Swales 1981, 1990, 2004, 2011). The bottom-up approach relies on inductive activities with a concordance that require the students to observe the lexicogrammatical features used to convey certain functional meanings. The end product of the course is a research paper of publishable quality that follows genre-writing norms for structural and rhetorical development in the student’s field.

Despite the indisputable value of disciplinary corpora, this corpus approach appeared to have an inherent drawback—the corpus-based activities, exploratory in
nature—supported only recognition. While extensive exposure to corpora did lead to self-discovery and increased awareness about the genre and the writing in the discipline, the approach did not create substantial opportunities for the production of writing similar to that observed, and the students were not able to transfer their corpus observations to their own writing. This task is, indeed, very challenging. It requires sufficient practice on the part of the student and guided discipline-specific feedback on the part of the instructor. To satisfy both these conditions, the course has been enhanced with a "custom-made" pilot AWE program—Intelligent Academic Discourse Evaluator (IADE)—trained to evaluate the rhetorical development of research article Introduction sections and to generate instant, individualized, genre, and discipline-specific feedback.

5.3 Theoretical Framework

IADE's design is based on a theoretical framework developed through the synthesis of three perspectives identified as informative for the writing pedagogy in the context described here: interactionist approach (IA) in SLA, skill acquisition theory (SAT), and systemic functional linguistics (SFL). Following is a summary of the most essential tenets relevant to the pedagogical practice and a rationale explaining how the synthesis of the three perspectives seeks to address the AWE issues outlined earlier.

The first set of teaching goals and learner needs listed in Table 1 are assessment for learning by means of formative feedback to students.

Table 1
Goals, Theories, and Issues Reflected in the Design of IADE

<table>
<thead>
<tr>
<th>Teaching goal</th>
<th>Learner need</th>
<th>Theory</th>
<th>AWE issue addressed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Formative assessment</td>
<td>Formative feedback</td>
<td>SLA (input, interaction, output)</td>
<td>Unfounded transfer</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Misuse</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Language proficiency</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Adjustment to scoring</td>
</tr>
<tr>
<td>Genre writing conventions</td>
<td>Genre and disciplinary writing norm</td>
<td>SAT (declarative knowledge)</td>
<td>Lack of acquired literacy</td>
</tr>
<tr>
<td>Written production</td>
<td>Extensive guided practice</td>
<td>SAT (procedural knowledge)</td>
<td>Lack of knowledge of AWE features</td>
</tr>
<tr>
<td>Written production of genre writing conventions</td>
<td>Genre and disciplinary writing with appropriate rhetoric</td>
<td>SFL (functional language)</td>
<td>Misuse</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Lack of understanding of AWE feedback and features</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Lack of integrated communicative dimension</td>
</tr>
</tbody>
</table>

4 For a detailed description of IADE, see Cotos (2009, 2010).
Educators working in the field of second language writing see feedback as being crucial for encouraging and consolidating learning. While the importance of feedback was first articulated in the early days of behaviorism (Thorndike, 1913), then tackled in social cognitivism, and later emphasized with the emergence of learner-centered approaches to L2 writing, it was also embraced by IA, according to which learning occurs “through the learner’s exposure to language, production of language, and feedback on that production” (Gass & Mackey, 2007, p. 176). In other words, such constructs as input, interaction, feedback, and output play a significant role in second language acquisition.

The corpus-based approach employed in the writing course only partially conformed with the IA perspective. It exposed the students to large amounts of input by having them read and analyze texts and language use in the given genre. However, the elements of interaction and feedback considered essential in the chain of acquiring language knowledge, were largely in deficit. This prevented the students from noticing discourse-related infelicities in their output and correcting these problems as a result of engaging in meaningful interaction. IADE was developed to fill in this gap by providing anytime feedback on rhetorical development in response to individual learner output.

The second and third sets of teaching goals and learner needs focus on learning about genre and disciplinary writing conventions and developing an ability to use them. Relevant to these sets of goals and needs is SAT rooted in Psychology (Newell & Rosenbloom, 1981), according to which “learning a wide variety of skills shows remarkable similarity in development from initial representation of knowledge through initial changes in behavior to eventual fluent, spontaneous, largely effortless, and highly skilled behavior” (DeKeyser, 2007, p. 97). Byrne (1986) postulated three stages of skill development: presentation, practice, and production. These stages were elsewhere referred to as cognitive, associative, and autonomous (Fitts & Posner, 1967) and declarative, procedural, and automatic (Anderson, 1982). At the first stage, learners acquire some knowledge about a certain skill without attempting to apply that knowledge in practice. Acting on this knowledge and transforming it into a behavior, or, in DeKeyser’s words “turning ‘knowledge that’ into ‘knowledge how’” (2007, p. 98), occurs at the second stage once the declarative knowledge is well acquired. Considerable practice is needed, though, before the specific behavior becomes automatic at the third stage.

How do these theoretical insights apply to the teaching of academic writing? In this writing course, the students acquired the declarative knowledge through teaching and guided observation of the rhetorical moves in the corpus. The problem was that learners were missing the second stage of the skill development—acquiring procedural knowledge through extensive practice. Consequently, IADE was introduced to provide sufficient practice opportunities through its multiple resubmission feature with the use of its feedback aimed at stimulating qualitative change in learners’ cognitive mechanisms activated during the writing task. Practice is also key to output production, which according to IA enhances learning. The fourth set of teaching goals and learner needs is related to the preceding ones, but is more distinct. To meet the genre-writing expectations in a particular
field, learners need to know what rhetorical choices are typical of their field and how those choices are expressed linguistically. This is where SFL comes into play. Halliday and Hasan (1985) affirm that familiarity with different genres is crucial and that such familiarity "does not grow automatically" (p. 68), advising that learners need to be exposed to genres, "particularly those that are actively required in the educational process" (p. 69). The corpus-based approach in the writing course provided the necessary exposure to the research article genre, and students' work with texts was fundamentally important since texts are viewed in SFL as the most appropriate units for analysis because they are semantic units that exhibit internal cohesiveness and contextual consistency (Halliday & Hasan, 1976). However, SFL posits that texts are both products of choices made from the meaning potential available in particular contexts and processes of meaning exchanges between the participants (here, writers and readers). Therefore, texts have an interactive nature and represent social events that unfold linguistically (Hasan, 1978). The corpus-based pedagogy, then, was both advantageous and faulty. Advantageous was the fact that the students worked with professional texts in their own disciplines in order to learn to engage in social interaction between themselves and their academic audiences by producing texts that model those they were exposed to when working with their discipline-specific corpus. The disadvantage was that the corpus exploration tasks were completed largely by means of identifying the structural components of texts (or, in SFL terminology, the schematic structures) and overlooking functional meanings. This was a considerable weakness since "every structural feature has its origin in the semantics, that is, it has some function in the expression of meaning" (Halliday, 1982, p. 8). IADE was trained to tackle this problem. It sustains the analysis of learners' production, and its feedback highlights specific rhetorical moves to draw their attention to the functional meaning of the discourse. Additionally, through its discipline-specific feedback based on a comparison of student drafts with a corpus annotated for rhetorical functions in a respective discipline, IADE is designed to establish relations between learners' academic texts and the professional practice they realize.

Applying IA, SAT, and SFL concepts to the goals and needs of the targeted instructional context at ISU informed the design of IADE and served to prevent potential problems that are likely to surface upon implementation of AWE programs. Being custom made for specific formative goals, IADE does not generate

---

5 According to Halliday (1985), "a theory of language [is] essentially consumer-oriented" and "the value of a theory lies in the use that can be made of it" (p. 7). He explains that "systemic theory is designed not so much to prove things as to do things" (p. 11). Therefore, in the academic writing context under discussion here, SFL does not serve as the background theory for studying the language in order to understand how it works and what people do with it. Rather, relevant SFL concepts are used to inform and improve teaching practice.

6 Numerical feedback is provided in the form of percentages showing the distribution of each move in the student draft and an average of the moves in the corpus of the student's discipline. It also includes percentages for the minimum and maximum occurrence of each move in the corpus.
scores but evaluative comments indicating how the rhetorical moves in a student draft approximate the norm in his or her discipline. This evaluation approach eliminates the problem of transferring and misusing an automated scoring tool, as well as the concerns of those members of academia who fear that students and teachers might adjust their writing to the assessment criteria of the software. Misuse is also prevented by implementation at the right point in the teaching and learning process, after the declarative knowledge about rhetorical functions is acquired. Plus, the influential role of the teacher and the social dimension of the human audience are acknowledged by including peer and teacher feedback in the revision process and implementing IADE as a supplementary tool that compensates for what the instructor and fellow students cannot offer. Next, the issues of insufficient technical literacy and understanding of AWE functionality are also addressed since IADE's feedback and features are based on specific course objectives and learned declarative knowledge. Its analysis engine is not trained on static compositional features; on the contrary, the feedback highlights functional meaning and is based on the communicative dimensions of disciplinary writing. Most importantly, IADE is designed with L2 learners in mind, centered on the constructs of input, interaction, output, and salience. The program analyzes learners' output and then uses it to generate color-coded feedback that is returned to the learners as modified input. The color codes serve as input enhancement designed to encourage noticing and focus on discourse form. IADE's feedback is intended to stimulate learner-computer interaction during the writing process and to trigger testing linguistic hypotheses with respect to functional meaning. Finally, the proficiency level of learners is also taken into account because the feedback is meant to help advanced learners of English, not in terms of grammar and mechanics—which they have largely mastered—but in terms of more global aspects of writing such as developing a scientific argument—which they still need to grasp and are motivated to learn.

AWE design that is grounded in theory and responsive to pedagogical practice and learning needs is an important preemptive measure for facilitating effective use; however, it is not a measure of effectiveness. An AWE system must be validated in the application context (Chung & Baker, 2003), and validation approaches should encompass evaluation of AWE effectiveness from multiple perspectives relevant to the purpose of implementation. Therefore, IADE was evaluated in terms of two major dimensions: system-centric (i.e., performance of the system and prototype testing) and user-centric (i.e., user-system interaction) (see Chodorow, Gamon, & Tetreault, 2010). Performance measures of accuracy, precision, and recall were previously introduced in Pendar and Cotos (2008), and calculations of reliability between IADE and human raters were reported in Cotos (2010). The user-centric dimension focused on investigating the effectiveness of IADE with L2 learners and was guided by a conceptual computer-assisted language learning (CALL) evaluation framework rooted in instructed SLA (see

IADE has a set of Help Options, which includes color-coded corpora annotated for rhetorical functions, definitions and examples of moves and steps, and tips on how to revise a draft with the program more effectively.
Chapelle, 2001), concentrating on the theoretical tenets operationalized in the design of IADE. This framework contains critical CALL qualities such as Language Learning Potential, Meaning Focus, Learner Fit, and Impact. The following section elaborates on the evaluation of Impact and presents evidence of cognitive and socioaffective processes that unfolded during students’ interaction with IADE.

5.4 Empirical Evaluation of Impact

IADE was implemented in the writing course discussed above with 105 international students from 11 countries, of which 37 were Masters and 68 were PhD students specializing in one of 34 disciplines. Different types of quantitative and qualitative data were collected for evidence of Impact, that is, the effects of learners’ interactions with IADE and its feedback. The quantitative data consisted of Likert-scale and yes/no survey responses. The qualitative data included open-ended survey responses, computer screen capturing of participants’ interaction with IADE, think-aloud protocols, observations, and semistructured interviews. This mixed-methods study adopted a process-product approach (Warschauer & Ware, 2006) and employed a concurrent transformative strategy (Creswell, 2003) by integrating quantitative and qualitative data that were obtained concomitantly and then integrated during the analysis and interpretation stage.

All the data were initially analyzed by source (see Table 2).

Table 2
Impact Data Analysis

<table>
<thead>
<tr>
<th>Data source</th>
<th># part.</th>
<th>Extenta</th>
<th>Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quantitative</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Yes/No questions</td>
<td>83</td>
<td>5</td>
<td>Comparison of response %</td>
</tr>
<tr>
<td>2. Likert-scale</td>
<td>88</td>
<td>4</td>
<td>Comparison of response %</td>
</tr>
<tr>
<td>Qualitative</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Open-ended</td>
<td>83</td>
<td>5</td>
<td>Manual analysis for themes</td>
</tr>
<tr>
<td>2. Think-aloud</td>
<td>16</td>
<td>25</td>
<td>Coding; descriptive statistics</td>
</tr>
<tr>
<td>3. Camtasia</td>
<td>16</td>
<td>25-38 minutes; 32 drafts</td>
<td>Coding; descriptive statistics; manual analysis of output modifications</td>
</tr>
<tr>
<td>4. Observations</td>
<td>16</td>
<td>25-38 minutes</td>
<td>Coding; descriptive statistics</td>
</tr>
<tr>
<td>5. Semistructured</td>
<td>16</td>
<td>10-15 minutes</td>
<td>Coding; descriptive statistics</td>
</tr>
</tbody>
</table>

The surveys used in the larger study contained 25 Yes/No questions, 25 open-ended questions, and 19 Likert-scale questions. This table shows the number of questions used to elicit data for Impact.

The scale questions offered four choices: a lot or very well, somewhat or well, a little, and not at all. Participants’ answers on individual scale questions were converted to percentages. Also, for the ultimate purpose of overall evaluation of Impact is capitalized in this chapter to indicate that it is a CALL quality from Chapelle’s (2001) framework targeted in the evaluation. The other CALL qualities (Language Learning Potential, Meaning Focus, and Learner Fit) are capitalized in the text for the same reason.
Impact, the four choices were equaled to a certain degree of evidence strength. In other words, a lot or very well was considered as excellent evidence, somewhat or well as good evidence, a little as weak evidence, and not at all as poor evidence. The open-ended response data were analyzed in terms of emerging themes reflecting participants’ perceptions of the types and consequences of Impact, which were then quantified in terms of percentages of students who mentioned them. Other qualitative data were transcribed and segmented into idea units (Kroll, 1977) in Transana. The transcripts were coded using a coding taxonomy developed on the basis of the results from a pilot study (coder reliability $k = .886$), which included the following categories: positive learning experience, negative learning experience, motivation, cognitive involvement, affective influence, strategy development, and influence on the revision process. Descriptive statistics were calculated for each coded category. The results of these quantitative and qualitative analyses were then integrated and triangulated.

5.4.1 Overall evidence of Impact

Evidence from both quantitative and qualitative data provided an understanding of the nature of the learner experience with IADE. Table 3 summarizes the extent of positive and negative evidence obtained. Triangulation of data allows for the inference that, at different stages, IADE’s feedback had different effects on the revision process, influencing the participants at affective, intrinsic, pragmatic, and cognitive levels.

<table>
<thead>
<tr>
<th>Evidence of Impact</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Data source</th>
<th># part</th>
<th>Positive evidence</th>
<th>Negative evidence</th>
<th>No evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likert-scale</td>
<td>Q 14 [motivation]</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>questions</td>
<td>Q 17 [affect]</td>
<td>35%</td>
<td>65%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>Q 18 [positive experience]</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>Q 19 [future AWE use]</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Yes/No and open-</td>
<td>Q 24 [motivation]</td>
<td>89%</td>
<td>0%</td>
<td>11%</td>
</tr>
<tr>
<td>ended questions</td>
<td>Q 25 [cognition]</td>
<td>57%</td>
<td>0%</td>
<td>43%</td>
</tr>
<tr>
<td></td>
<td>Q 26 [affect]</td>
<td>92%</td>
<td>0%</td>
<td>8%</td>
</tr>
<tr>
<td></td>
<td>Q 27 [positive experience]</td>
<td>90%</td>
<td>0%</td>
<td>10%</td>
</tr>
<tr>
<td></td>
<td>Q 28 [pragmatics]</td>
<td>90%</td>
<td>0%</td>
<td>10%</td>
</tr>
</tbody>
</table>

Think-alouds/Camtasia

<table>
<thead>
<tr>
<th>Interviews</th>
<th>16</th>
<th>367 of 1227 idea units</th>
<th>3 of 1227 idea units</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Observations</th>
<th>16</th>
<th>40 of 233 idea units</th>
<th>1 of 233 idea units</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Interviews</th>
<th>16</th>
<th>174 of 460 idea units</th>
<th>7 of 460 idea units</th>
</tr>
</thead>
</table>
5.4.2 Learner perceptions on Impact

Impact was regarded as positive by the majority of participants, as seen in the Yes/No survey data (see Figure 1).

Figure 1
Strength of Positive Impact in Yes/No Survey Data ($N = 83$)

The Likert-scale responses also largely pointed to learners' positive experience with IADE, averaging to 47% of excellent, 40% of good, 11% of weak, and 1% of poor evidence (see Figure 2a).

Figure 2a
Strength of Positive Impact in Likert-scale Data ($N = 88$)

According to the more detailed survey findings summarized in Figure 2b, most of the participants were motivated to improve, had a positive experience, felt excited about their improvement and sometimes frustrated about the lack of it, and expressed willingness to work with IADE in the future.
Participants' open-ended responses revealed why interaction with IADE was perceived as motivational. It appeared that the feedback was a major factor since it was mentioned by 41% of the students who felt stimulated to improve their drafts when they received “positive,” “negative,” or “guiding” feedback, that is, “feedback constantly direct[ing them] to[wards] improvement” (Student 67). As Student 4 put it, “the feedback really gave me some positive power for the every revision.” The disciplinary orientation of IADE was motivating for 28%, and the opportunity for iterative resubmission for another 18%. The remaining 13% did not elaborate on how IADE motivated them, having only stated that it did.

The same data source suggests affective Impact. Most of the respondents (92%) noted that they were excited to see improvement when the feedback on their modifications was returned. For instance, Student 38 wrote, “When I saw my improve feedback, I feel good” and Student 65 remarked that he “did get excited, not like a child ;-) but as a happy student.” Many of the students (63%) explained the causes of their excitement. For 30% it was the feeling of accomplishment, which was oftentimes not easy to achieve. In other words, having gone through a series of trial and error attempts, the learners felt like their hard work “paid off” (Student 97) when they saw a positive change in IADE’s feedback. For 24% it was the proof of improvement displayed in IADE’s feedback that had this effect. A few students (9%) experienced disappointment due to multiple unsuccessful output modifications, but they felt like their excitement was stronger after that point. Some participants made it clear that their experience with IADE was not entirely positive. Responding to a Likert-scale question, 3% indicated that they were very frustrated during the interaction with the program, and 25% were somewhat frustrated (the rest of the participants described their experience as either a little frustrating (36%) or not at all (35%). Two explanations were predominant: their particular discipline was not represented in IADE, and the color-coded feedback did not always match their communicative intent.

A pragmatic effect became evident as the participants (90%) acknowledged that the feedback influenced their usual revision process and indicated modified or newly acquired revision strategies (see Table 4).
One of those strategies was learners' reliance on the guidance of IADE's feedback (19%). In other words, they took the feedback almost as a directive. If there was some negative evidence about a move (especially numerical), they would try to improve on that particular move; if the move was in the average range compared to their discipline, they would not work on it anymore. A similar strategy, which was articulated by 11% of the students, was to set the goal of reaching the standards of the discipline as presented by the average percentages in the feedback. While setting goals is a positive choice, it is still reliance on the numerical feedback that leads to accomplishing this type of goal with IADE, which is why employing this strategy by itself is somewhat confining. Additionally, focus on discourse form became part of revision for some participants (9%) who mentioned that they consistently paid attention to the distribution of the color-coded moves. A larger group of participants (27%) developed their focus on meaning into a strategy by which they carefully thought of the functional meaning of individual sentences. For example, Student 89 explained, "I decided to check every sentence and think what step it function, so I checked if the color was right—then good, if not—then I know I have to think about what I say and how to make it sound like what should be." It must be mentioned here that, for 8%, the focus on meaning strategy took the form of "pay[ing] much attention to wording" (Student 31) and searching for move-specific vocabulary to then use in their text modifications.

The emerging strategy of focusing on the meaning and form of the discourse during the revision process is an encouraging finding, especially considering that these types of focus are believed to contribute to language learning (Ellis, 2005; Long, 1996; Schmidt, 1994). However, some learners' heavy reliance on the program's automated feedback to reach average percentages in their field and too much attention to move-specific phraseology may not be as plausible (though the latter does have an advantage that will be commented on later). Perhaps this is why a number of students (24%) mentioned that they did not think much about the quality of their revisions because they "just tried to meet the feedback" (Student 33). These strategies may also be limiting in that they might constrain autonomous learning and the development of strategies needed for independent writing that are not be facilitated by IADE or other CALL applications. Moreover, it is

---

**Table 4**

Perceptions on Strategy Development

<table>
<thead>
<tr>
<th>Revision strategies</th>
<th>Participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reliance on the guidance of feedback</td>
<td>19%</td>
</tr>
<tr>
<td>Setting the goal to reach disciplinary standards</td>
<td>11%</td>
</tr>
<tr>
<td>Focus on discourse form</td>
<td>9%</td>
</tr>
<tr>
<td>Focus on functional meaning</td>
<td>27%</td>
</tr>
<tr>
<td>(Move-specific lexical output modifications)</td>
<td>(8%)</td>
</tr>
<tr>
<td>No previous strategy</td>
<td>6%</td>
</tr>
<tr>
<td>No comments</td>
<td>28%</td>
</tr>
</tbody>
</table>

---
also lamentable that learners may not evaluate their own work products; rather, they may be more likely to take the automated feedback as the only indication of the quality of their writing, trying to address only the salient negative evidence in the feedback and finish revising as soon as the feedback displayed percentages close to the average in their discipline.

Nevertheless, the responses of 57% of the students who said that they did think about how to improve their writing shed some useful light on this issue. Interestingly, their explanations were directly related to the revision strategies that included conscious focus on discourse meaning and form. It seems that those learners who employed these focus-directed strategies found themselves thinking more often and more deeply than they normally would have. Some of them believed that the more they revised, the more profoundly they became engaged cognitively. In Student 72's words, "Actually not at the beginning, not when I began to revise, but a lot later after I figured my idea didn't match with colors." Others explained that it was necessary to think a lot on their own "because it helps to compare my thought with what is in IADE" (Student 30) and "because if I don't think by myself, the feedback from IADE is not so meaningful to me" (Student 4).

5.4.3 Introspective insights on Impact

Data from the think-aloud/Camtasia protocols, observations, and semistructured interviews substantiated the survey findings with evidence indicating Impact at different levels: cognitive, pragmatic, affective, and intrinsic (see Table 5).

Table 5
Themes Indicative of Impact Identified in the Transcripts

<table>
<thead>
<tr>
<th>Impact levels and themes that emerged</th>
<th>Think alouds/ Camtasia</th>
<th>Observations</th>
<th>Interviews</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cognitive</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cognitive involvement</td>
<td>38.65% (143 idea units)</td>
<td>25.41% (46 idea units)</td>
<td>30.23% (13 idea units)</td>
</tr>
<tr>
<td>Pragmatic</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Strategy development</td>
<td>34.32% (127 idea units)</td>
<td>31.49% (57 idea units)</td>
<td>11.63% (5 idea units)</td>
</tr>
<tr>
<td>Influence on revision process</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Affective</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Emotional involvement</td>
<td>7.84% (29 idea units)</td>
<td>6.63% (12 idea units)</td>
<td></td>
</tr>
<tr>
<td>Positive learning experience</td>
<td>7.84% (29 idea units)</td>
<td>12.71% (23 idea units)</td>
<td>37.21% (16 idea units)</td>
</tr>
<tr>
<td>Negative learning experience</td>
<td>0.81% (3 idea units)</td>
<td>3.87% (7 idea units)</td>
<td>2.33% (1 idea units)</td>
</tr>
<tr>
<td>Intrinsic</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Motivation</td>
<td>10.54% (39 idea units)</td>
<td>19.89% (36 idea units)</td>
<td>9.3% (4 idea units)</td>
</tr>
<tr>
<td>Impact idea units</td>
<td>370</td>
<td>181</td>
<td>43</td>
</tr>
<tr>
<td>Total idea units</td>
<td>1227</td>
<td>460</td>
<td>233</td>
</tr>
</tbody>
</table>
The students engaged in self-reflection, analyzed their own output, explained their communicative intent, confirmed or rejected their own hypotheses, and drew conclusions based on IADE's feedback. Such cognitive involvement, in addition to being overtly expressed in students' utterances, was covertly noticeable in frequent interjections like "hm," long pauses, and body language as well as when they highlighted and reread parts of their writing on the computer screen. What is interesting, although not surprising, is that thinking was triggered particularly when the learners noticed negative evidence related to their intended functional meaning and the form of their discourse.

In terms of pragmatic influence, the transcripts attested to learners' development of new revision strategies during the interaction with IADE. The idea units coded as strategy development were second in frequency after the cognitive involvement category. The emerging strategies theme here was very similar to that mentioned by the participants in their open-ended survey responses—taking action prompted by the feedback and focusing on discourse meaning and form. To a certain extent, this was determined by the program's features, which offered immediate feedback of numerical and color-coded types, opportunities for resubmission, and Help Options. While making use of these features was common for all the observed participants, the sequence in which they were accessed changed gradually as the revision process unfolded. Reliance on feedback and resubmission were more frequent at the beginning as opposed to cognitive involvement and consultation of help that became more prominent later on.

All the students began their revisions by trying to improve on the move that was the farthest from the average in their discipline. For example, Student 39 said, "OK. Because average is about 55% for move1, I have 61% roughly, so it's ok for me. And for move2, the average is about 14, and I have 14%! And for move3 the average is 30%, and I have so little ... It's ok. I think I just have to add more move3" (think-aloud/Camtasia transcript). Later, the students developed a more thoughtful approach and acted more based on their self-reflection, constantly accessing the Help Options and comparing the examples of moves and steps in the annotated corpus to those in their own draft. For some, reaching this stage took longer than for others because they appeared to continue focusing mainly on the numerical feedback, taking actions only when it displayed some negative evidence and attending to the problem by making quick minor modifications, which were resubmitted multiple times.

Although the participants did not finish revising their drafts when being observed, towards the end of the session many of them (73%) reached a stage where their revision process took the form of very detailed self-verification. They began checking the move colors in the feedback against the intended functional meaning sentence by sentence. Thus, focus on functional meaning seemed to be a determining factor in the development of a new, more effective revision approach. As shown in Figure 3, it appeared that, at the initial revision stage, when the learners paid attention mostly to the numerical feedback, their modifications (e.g., change of placement, deletions, and substitutions of words) were sporadic, inconsistent,
and not very successful. As a result, they experienced frustration and did not seem motivated.

Figure 3
Participants’ Revision Process

Once the learners began focusing on functional meaning, they tended to reflect on and revise sentence by sentence, or, when they focused more on the form of their discourse, they tended to revise move by move. In both these cases, the revision process became more organized and resulted in more successful output modifications that were accompanied by bursts of positive emotions. In many instances, the learners verbalized satisfaction when receiving positive feedback upon resubmission or when self-confirming successful output modifications, indications of a positive learning experience. On the other hand, their dissatisfaction expressed when seeing negative feedback, which was often returned for repeated attempts to modify the same piece of writing, was indicative of a negative learning experience. Positive experience appeared to be more frequent compared to negative (see Table 5 above), possibly because the learners realized that improvement “does not come easy” and that “it’s normal that I don’t get right the first time” (Students 43 and 86, semistructured interview).

Next, it appeared that the learners gradually developed an intrinsic desire to improve and that their motivation was driven by IADE’s both negative and positive feedback. Negative feedback motivated them to address their problems, while positive feedback motivated them to continue revising. However, if negative feedback was too frequent, meaning that the modifications made were repeatedly unsuccessful, the participants were more likely to become frustrated and either give up or decide to return to that modification later. What also seemed to have a motivational effect was the content in IADE’s Help Options, which the participants accessed when struggling with finding a way to approach certain issues in
their drafts. The interviews revealed that characteristics of the feedback such as iterativeness and instancy were also motivational, driving the students towards evaluating the effectiveness of their modifications. Student 32 clarified, “It helps because what I change is still fresh and it makes sense when I have feedback right away. When I get some back from the teacher, I forget and don’t worry about every comment too much” (interview transcript). It also seemed that the more cognitively engaged the participants were and the more effective revision strategies they employed, the more motivated they became.

Given these insights it can be inferred that, during revision with IADE, Impact was exerted at different levels and seemed to have had certain effects on one another (see Figure 4).

Specifically, the depth of cognitive involvement conditioned the use of new revision strategies, thus influencing the pragmatic aspect of the revision process. The pragmatic choices, in turn, had either positive or negative effects on learners at affective and motivational levels. Further, it can be assumed that the two types of feedback, numerical and color coded, both perceived as beneficial and motivational by the learners, may have had opposite effects on their approach to revision and on the type of Impact (negative or positive) they experienced. Numerical feedback may have inhibited the effectiveness of the revision strategies the learners employed, possibly encouraging them to limit revision to approximating percentages. Color-coded feedback, on the other hand, seemed to enhance this process by facilitating meaning focus. This is not to say, however, that one or the other type of feedback is clearly disadvantageous or beneficial. Most likely, it is separate use of individual strategies that induced negative Impact. Developing and combining strategies that integrate information from both the numerical and
6. Discussion

Unlike much prior AWE research, this study supports perceived and observed effectiveness with introspective evidence. In short, interaction with IADE resulted in both positive and negative Impact. The Impact was likely to be negative when learners relied only on numerical feedback; in this case their cognitive involvement was low, which led to ineffective revision strategies and sporadic output modifications and resulted in frustration and disappointment. Positive Impact, on the other hand, occurred when the learners focused on color-coded feedback, in which case the degree of cognitive involvement increased. They employed more effective revision strategies and made consistent and successful changes to their texts, consequently experiencing positive affective Impact. Negative Impact was common at the beginning of the revision process; however, the learning curve changed its direction towards positive Impact as the students reached a turning point when the color-coded feedback drew their attention to functional meaning.

These findings support explicit claims about the interaction between IADE use and learning outcomes and the processes triggered by this interaction.

- Automated feedback generated from learner output and provided as enhanced (color-coded) input can trigger focus on the functional meaning of discourse and stimulate cognitive involvement, which is a prerequisite for positive pragmatic, affective, and intrinsic Impact.
- Cognitive involvement with genre/meaning-based AWE may enhance the revision process and learners' positive experience.
- Automated feedback generated from learner output and provided as discipline-specific comments can boost motivation to learn and conform with disciplinary genre conventions.

Based on these claims, it can be inferred that the factors likely to contribute to AWE effectiveness are: creating cognitive conditions for intrapersonal interaction, emphasizing focus on meaning, and providing sufficient opportunities for practice guided by learner-fit feedback—all of which is contingent upon previously acquired declarative knowledge. Taking these claims further, it can be hypothesized that if such conditions are created through the use of AWE, a more productive revision process will culminate in learning and improvement, which is commonly recounted in AWE literature (Attali, 2004; Elliot & Mikulas, 2004; Foltz et al., 1999; Leah Rock, 2007; Schroeder et al., 2008). In this study, for instance, indirect evidence supporting this hypothesis was found in participants' output modifications captured with Camtasia, which improved as their revision process moved towards a more thoughtful approach to evaluating their writing. Indeed, as reported by Cotos (2011), students who used IADE improved the rhetorical quality of their drafts from first to last (measured by the program and by human raters) and exhibited learning gains (assessed through pre- and posttests). Improvement was explained in terms of a qualitatively observed sequential learn-
ing cycle that occurred repeatedly during revision with IADE, with the head of the cycle being focus on discourse form and other stages being noticing negative evidence, enhanced understanding, and output modification. A similar cycle could have been in place for Schroeder et al.'s (2008) and Hegelheimer et al.'s (2011) students who thought that Criterion helped them understand the nature of their errors and identify the weaknesses in their writing.

The findings of this study indicate that cognitive involvement was a key factor contributing to the effectiveness of IADE and advanced our understanding the role of cognition in L2 writing facilitated by AWE. This is particularly valuable especially because thought processes are emphasized in classical cognitive models of writing (Bereiter & Scardamalia, 1987; Hayes, 1996), which agree that expert writing involves the use of expressive, receptive, and reflective modes of thought that support fluent production, self-monitoring and revision, and planning and evaluation, respectively. Writing skills are also conceptualized in terms of such types of cognitive representations as social, conceptual, textual, and verbal. Lines of tangency can be drawn between these representations and specific IADE Impact evidence, supporting the assumption that effective interaction with automated feedback may lead to better writing outcomes. First, the social element presupposes conscious awareness of the "rhetorical problem space" (Bereiter & Scardamalia, 1987) or, in other words, the rhetorical relation between the author and the audience. The participants in this study appeared to develop this rhetorical awareness through their focus on the functions of the moves and steps in the discourse. Second, the conceptual element involves knowledge and reasoning, which was repeatedly observed as the learners reflected on and clarified their understanding of the moves by consulting definitions and examples provided in the Help Options, consequently analyzing their own modified output and confirming or rejecting self-generated hypotheses based on IADE’s feedback. Third, the textual element concerns the production of a coherent and well organized text. Learners’ focus on discourse form with the help of the color-coded feedback (which provides a visual image of the text’s structural development) accounts for this type of cognitive contribution. And fourth, the verbal element requires appropriate use of linguistic representations of sentences and the rhetorical messages they encode. As mentioned earlier, the participants discovered a connection between move-specific phraseology and translated that connection into a strategy of using lexical items indicative of certain functional meanings.

Learners’ progress towards effective strategic use of genre/meaning-based AWE is another significant insight because skillful strategies are essential for a developing writer (Hayes & Flower, 1980). During interaction with IADE, the learners employed both effective and ineffective strategies. One may argue that training learners to use effective strategies is called for, as has been suggested by research on the acquisition of complex skills (see Block & Parris, 2008; Graham, MacArthur, Graham, & Fitzgerald, 2006). Indeed, learners, especially those using AWE outside instruction, would benefit from strategic tutorials. However, the lack of such training may also have its value, as it occurred with the users of IADE who came to realize which strategies were helpful and which were not. For
them, the discovery of effective strategies increased positive affective and intrinsic impact. Moreover, they comprehended how strategic revision coheres with the specific writing task in a meaningful way, which is of paramount importance. As Deane (2011) explains, "writers are likely to be ill-served if they learn strategies piecemeal, without understanding how to connect them to meaningful purposes" (p. 21).

On the other hand, the instructional context probably prepared learners for reaching the turning point from ineffective to effective strategy use. Context-specific implementation of IADE may thus account for learners' positive experience with the tool since they had acquired the necessary rhetorical literacy in the classroom prior to using IADE and easily mastered its features, which facilitated their interaction with it, and allowed them to focus on what the program was intended to do rather than trying to decipher the meaning of the color-coded feedback or the functionality of the Help Options. Therefore, context and pedagogical practices are indeed potential effectiveness factors, as previously adduced by Chen and Cheng (2008). Similarly, the instructional design of IADE may have steered learners' cognitive involvement and promoted consolidation of newly acquired declarative knowledge of the rhetorical shifts.

IADE's context-specific design and implementation may also help explain personal reactions like intrinsic motivation and affective behavior, especially positive reactions. It is worth recalling here that the discipline-specific feedback was numerical, and while it did not contain a score, it did provide a range of percentages and evaluative statements to inform the learners how far or how close they were to change the averages of their moves. That motivated them to revise the content of their drafts, time and again, in order to improve rhetorical effectiveness and even inclined them towards opting for future use of IADE or analogous AWE programs. It seems that motivation can be increased with guiding evaluative comments, not necessarily by scored feedback, as suggested by Lai (2010). As for affect, IADE's integration as a complementary formative assessment tool along with teacher and peer feedback may have reduced the level of frustration and negative attitudes. Perhaps it is the lack of context-related evidence that produced ambiguous learner perception results in the studies of Fang (2010), Lai (2010), and Yang (2004). Finally, the time of exercising AWE-based formative evaluation may be another circumstance affecting AWE effectiveness. Since L2 students may not have good strategies for revision with AWE to start with, as was found in this study, AWE use is warranted at early stages of drafting, which is a deduction that Chen and Cheng (2008) made as well.

7. Conclusion

The overall case of IADE supplies a more in-depth understanding of how AWE technology can be conceptualized, how it can be operationalized through princi-
plied use, and how it can be empirically evaluated for effectiveness in a given context. While the findings about IADE cannot be generalized across contexts, they are informative and have direct implications for classroom uses of other AWE programs, particularly in terms of matching certain program features with specific learning goals and using them as postulated by relevant theories.

This chapter raised issues that may affect AWE effectiveness, attempting to tackle their causes through evidence and inferences from previous research. AWE, in principle, has empowering potential to enhance L2 teaching and learning processes provided it is designed to address specific learning goals in a way that is supported by relevant theoretical knowledge. IADE was introduced here as an example of how it might be best, if not ideal, to avert problems undermining the potency of AWE. This example makes explicit the importance of the relationship between principled design and learning objectives for understanding and attaining specific outcomes. However, custom-made AWE design is not always a realistic or feasible endeavor. The underlying idea here is that L2 writing practice should be interwoven with theory and guided by empirical evidence so that the potential of AWE is exploited to its fullest and so that latent caveats are anticipated and prevented.

Like any other technologies, AWE may unavoidably bring both desired and undesired effects. However, it should not be viewed only as exerting an influence; rather, it should be validated as "reshap[ing] the broader ecology of the classroom" (Warschauer & Ware, 2006, p. 175) which is determined by context-specific policies, instructional strategies, learners' needs, and so forth. To move forward towards effective AWE use, we need to explore the abstruse land of teaching and learning processes in order to understand what may lead to improvement or regression in outcomes. This is an area wide open for empirical investigation. Researchers are encouraged to pursue this enterprise and build an AWE evaluation research agenda that would guide L2 writing pedagogy, and practitioners are encouraged to join and help conduct classroom-based action research with Criterion, MY Access!, WriteToLearn, and other AWE programs. Collaborative undertakings will put AWE to the test, reveal strengths and weaknesses, and inform mindful use, at the same time benefiting developers as they strive to improve the design and functionality of their products. Ultimately, only a synergy of efforts by theoreticians, researchers, practitioners, and software designers will equip L2 writing with automated evaluation capable of transforming challenging writing tasks into substantive learning experiences.

References


ELENACOTOS 109


