2011

Potential of Automated Writing Evaluation Feedback

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, 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/58. For information on how to cite this item, please visit http://lib.dr.iastate.edu/howtocite.html.
Potential of Automated Writing Evaluation Feedback

Abstract
This paper presents an empirical evaluation of automated writing evaluation (AWE) feedback used for L2 academic writing teaching and learning. It introduces the Intelligent Academic Discourse Evaluator (IADE), a new web-based AWE program that analyzes the introduction section to research articles and generates immediate, individualized, and discipline-specific feedback. The purpose of the study was to investigate the potential of IADE's feedback. A mixed-methods approach with a concurrent transformative strategy was employed. Quantitative data consisted of responses to Likert-scale, yes/no, and open-ended survey questions; automated and human scores for first and final drafts; and pre-/posttest scores. Qualitative data contained students' first and final drafts as well as transcripts of think-aloud protocols and Camtasia computer screen recordings, observations, and semistructured interviews. The findings indicate that IADE's colorcoded and numerical feedback possesses potential for facilitating language learning, a claim supported by evidence of focus on discourse form, noticing of negative evidence, improved rhetorical quality of writing, and increased learning gains.

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

Comments
Potential of Automated Writing Evaluation Feedback

ELENA COTOS
Iowa State University

ABSTRACT
This paper presents an empirical evaluation of automated writing evaluation (AWE) feedback used for L2 academic writing teaching and learning. It introduces the Intelligent Academic Discourse Evaluator (IADE), a new web-based AWE program that analyzes the introduction section to research articles and generates immediate, individualized, and discipline-specific feedback. The purpose of the study was to investigate the potential of IADE’s feedback. A mixed-methods approach with a concurrent transformative strategy was employed. Quantitative data consisted of responses to Likert-scale, yes/no, and open-ended survey questions; automated and human scores for first and final drafts; and pre-/posttest scores. Qualitative data contained students’ first and final drafts as well as transcripts of think-aloud protocols and Camtasia computer screen recordings, observations, and semistructured interviews. The findings indicate that IADE’s color-coded and numerical feedback possesses potential for facilitating language learning, a claim supported by evidence of focus on discourse form, noticing of negative evidence, improved rhetorical quality of writing, and increased learning gains.

INTRODUCTION
Over the past decade, automated writing evaluation (AWE) has witnessed an increasing interest in the field of L2 writing (Chen & Cheng, 2008; Warschauer & Ware, 2006; Yang, 2004). Arguably, the most promising point of contact between the areas of AWE and L2 writing is the move toward designing AWE programs that are modeled on pedagogical principles of second language acquisition (SLA). Such feedback checks, clarification requests, comprehension checks, or recasts (see Gass, Mackey, & Ross-Williams, 2007)—to a model (Block, 2003; Ramirez, 2005) or paradigm (Byrnes, 2005). Carroll (1999) even calls it the interaction theory. Gass and Mackey (2007) refer to it as the interaction approach (IA), explaining that it is a model in the sense that it describes the processes involved when the learners encounter input, are involved in interaction, and receive feedback and produce output. However, it is moving towards the status of a theory in the sense that it also attempts to explain why interaction and learning can be linked, using cognitive concepts … such as noticing, working memory, and attention.

The major constructs of the IA are input, interaction, feedback, and output. Input, or the target language to which learners are exposed, assumes a central role in any SLA theory and is “perhaps the single most important concept of second language acquisition” (Gass, 1997, p. 1). An underlying tenet of the IA is that the input to the learner coupled with the learner’s manipulation of the input through interaction forms a basis for language development” (Gass, 1997, p. 87). Long (1996) argues that the input has to provide both positive evidence, that is, “target-like models” (Mackey, 2006, p. 406), and negative evidence, that is, “direct or indirect information about what is ungrammatical and/or unacceptable” (Gor & Long, 2009, p. 445). Such evidence becomes available during interaction.

Interaction is the context in which the language is used. During interaction, the learners’ attention is drawn to problematic aspects of their language use. They may notice a gap (Schmidt & Frota, 1986), or a mismatch between the input and their own organization of the target language” (Gass & Mackey, 2007, p. 184), and that “providing feedback in a linguistic form, conversational structure, message content … until an acceptable level of understanding is achieved” (Long, 1996, p. 418). In other words, the learners engage in negotiation of meaning, during which they can receive feedback that either confirms their communicative success or points to failure in their production.

Feedback is an essential aspect of interaction, and it is generally viewed as a form of negative evidence that can help the learners notice the mismatch between the target language and their own interlanguage form (Mackey, 2006). Interactional feedback can be explicit, provided in the form of corrections and metalinguistic explanations, and implicit such as confirmation checks, clarification requests, comprehension checks, or recasts (see Gass, Mackey, & Ross-Feldman, 2005; Mackey, Gass, & McDonough, 2000; Oliver & Mackey, 2003). Such feedback is valuable in that it can stimulate learners to generate hypotheses concerning the nature of their linguistic problem.

Feedback that took the form of a learning cycle, the elements of which were: focus on form, noticing of negative evidence, enhanced understanding of functional meaning, and output modification. As a result of such interaction, the quality of learners’ written products improved significantly, modifications being made mostly at the level of content, very little at the level of structure and less in grammar and mechanics. Consequently, the findings of this study attest to the value of the interactionist concepts for the implementation of AWE in second language (L2) writing contexts and have direct implications for AWE design.
ABSTRACT
This paper presents an empirical evaluation of automated writing evaluation (AWE) feedback used for L2 academic writing teaching and learning. It introduces the Intelligent Academic Discourse Evaluator (IADE), a new web-based AWE program that analyzes the introduction section to research articles and generates immediate, individualized, and discipline-specific feedback. The purpose of the study was to investigate the potential of IADE’s feedback. A mixed-methods approach with a concurrent transformative strategy was employed. Qualitative data consisted of Likert-scale, yes/no, and open-ended survey questions; automated and human scores for first and final drafts; and pre-/posttest scores. Qualitative data contained students’ first and final drafts as well as transcripts of think-aloud protocols and Camtasia computer screen recordings, observations, and semi-structured interviews. The findings indicate that IADE’s color-coded and numerical feedback possesses potential for facilitating language learning, a claim supported by evidence of focus on discourse form, noticing of negative evidence, improved rhetorical quality of writing, and increased learning gains.

INTRODUCTION
Over the past decade, automated writing evaluation (AWE) has witnessed an increasing interest in the field of L2 writing (Chen & Cheng, 2008; Warschauer & Ware, 2006; Yang, 2004). Arguably, the most promising point of contact between the areas of AWE and L2 writing is the interaction approach (IA), explaining that it is perhaps the single most important concept of second language acquisition (Gass, 1997, p. 1). An underlying tenet of the IA is that “the input to the learner coupled with the learner’s manipulation of the input through interaction forms a basis for language development” (Gass, 1997, p. 87). Long (1996) argues that the input has to provide both positive evidence, that is, “target-like models” (Mackey, 2006, p. 406), and negative evidence, that is, “direct or indirect information about what is ungrammatical and/or unacceptable” (Gor & Long, 2009, p. 445). Such evidence becomes available during interaction.

The major constructs of the IA are input, interaction, feedback, and output. Input, or the target language to which learners are exposed, assumes a central role in any SLA theory and is perhaps the single most important concept of second language acquisition (Gass, 1997, p. 1). An underlying tenet of the IA is that “the input to the learner coupled with the learner’s manipulation of the input through interaction forms a basis for language development” (Gass, 1997, p. 87). Long (1996) argues that the input has to provide both positive evidence, that is, “target-like models” (Mackey, 2006, p. 406), and negative evidence, that is, “direct or indirect information about what is ungrammatical and/or unacceptable” (Gor & Long, 2009, p. 445). Such evidence becomes available during interaction.

Interaction is the context in which the language is used. During interaction, the learners’ attention is drawn to problematic aspects of their language use. They may notice a gap (Schmidt & Frolta, 1986), or a “mismatch between the input and their own organization of the task at hand” (Gass & Mackey, 2007, p. 184), and that “providing input in the linguistic form, conversational structure, message content ... until an acceptable level of understanding is achieved” (Long, 1996, p. 418). In other words, the learners engage in negotiation of meaning, during which they can receive feedback that either confirms their communicative success or points to failure in their production.

Feedback is an essential aspect of interaction, and it is generally viewed as a form of negative evidence that can help the learners notice the mismatch between the target language and their own interlanguage form (Mackey, 2006). Interactional feedback can be explicit, provided in the form of corrections and metalinguistic explanations, and implicit such as confirmation checks, clarification requests, comprehension checks, or recasts (see Gass, Mackey, & Ross-Feldman, 2005; Mackey, Gass, & McDonough, 2000; Oliver & Mackey, 2003). Such feedback is valuable in that it can stimulate learners to generate hypotheses concerning the nature of their linguistic problem.

The major constructs of the IA are input, interaction, feedback, and output. Input, or the target language to which learners are exposed, assumes a central role in any SLA theory and is perhaps the single most important concept of second language acquisition (Gass, 1997, p. 1). An underlying tenet of the IA is that “the input to the learner coupled with the learner’s manipulation of the input through interaction forms a basis for language development” (Gass, 1997, p. 87). Long (1996) argues that the input has to provide both positive evidence, that is, “target-like models” (Mackey, 2006, p. 406), and negative evidence, that is, “direct or indirect information about what is ungrammatical and/or unacceptable” (Gor & Long, 2009, p. 445). Such evidence becomes available during interaction.

Interaction is the context in which the language is used. During interaction, the learners’ attention is drawn to problematic aspects of their language use. They may notice a gap (Schmidt & Frolta, 1986), or a “mismatch between the input and their own organization of the task at hand” (Gass & Mackey, 2007, p. 184), and that “providing input in the linguistic form, conversational structure, message content ... until an acceptable level of understanding is achieved” (Long, 1996, p. 418). In other words, the learners engage in negotiation of meaning, during which they can receive feedback that either confirms their communicative success or points to failure in their production.

Feedback is an essential aspect of interaction, and it is generally viewed as a form of negative evidence that can help the learners notice the mismatch between the target language and their own interlanguage form (Mackey, 2006). Interactional feedback can be explicit, provided in the form of corrections and metalinguistic explanations, and implicit such as confirmation checks, clarification requests, comprehension checks, or recasts (see Gass, Mackey, & Ross-Feldman, 2005; Mackey, Gass, & McDonough, 2000; Oliver & Mackey, 2003). Such feedback is valuable in that it can stimulate learners to generate hypotheses concerning the nature of their linguistic problem.
Output gives learners the opportunity to produce language and “serves as a means of hypothesis testing” (Gass, 1997, p. 7). After having used a language form unsuccessfully and after having created a certain hypothesis based on the received feedback, learners are pushed to modify their linguistic form and produce more precise and appropriate output—comprehensible output (Swain, 1985). Modified output is useful provided that learners see the connection between their erroneous form, the feedback, and the revised output (Carroll, 2001; Gass & Mackey, 2006). Continued production of output is important because, in the long run, it leads to automaticity.

Traditionally, interaction has referred to learners’ engagement in conversations with interlocutors. Ellis (1999) expanded the idea of interaction from interpersonal level to that of intrapersonal level; “interaction that can occur in our minds, ... and, more covertly, when different modules of the mind interact to construct an understanding of or a response to some phenomenon” (p.i.). Later, Chapelle (1998, 2001, 2007) connected these concepts to learner-computer interaction, showing how IA constructs can be enhanced to facilitate language learning with the help of computers. Specifically, the linguistic features in the input can be made salient through highlighting, glosses, hyperlinks, pictorial or video representations, and so forth. The output can be repeatedly challenged (Ferris, 1999, 2002, 2003; Hyland & Hyland, 2006; Polio 1997). Ferris (1991; Krashen, 1984; Truscott, 1996) asserted that “[g]rammar correction should be abandoned” (p. 328) and that error correction in general can be ineffective and can even have deleterious effects because it is capable of responding to multiple problematic aspects of language use that may occur in learner’s production.

The concerns regarding ICALL feedback resonate with the debate on corrective feedback. Most of the existing ICALL programs target the development of learners’ grammatical competence, revolve around the functionality of a parser, and provide immediate intelligent feedback. Their feedback features vary in their degree of specificity and explicitness as well as in their ability to adapt to individual learners (see list of ICALL programs and their feedback features in Appendix A). The question, however, is not what kind of feedback ICALL can generate; rather, it is what kind of feedback ICALL should generate. Nagata (1995) argues that “[i]f we use an intelligent system, we should examine carefully what kind of error messages should be provided ... and how effective they are” (p. 49).

Research on ICALL feedback is still scarce (see Heft & Schulze, 2007), but the relatively few pieces of empirical evidence suggest positive effects of explicit intelligent feedback. When comparing ICALL feedback with CALL feedback, Nagata (1993, 1995) found that the former is more effective than the latter. The more detailed the intelligent feedback is, the better the learning outcomes are (Heft, 2001, 2002, 2004, 2005). Learners appear to show significantly more uptake over time with a more error-specific feedback type (Heft, 2008; Heft & Rimrott, 2008); Yang and Akahori’s (1999) findings indicate that feedback that corresponds to the input created by the learner is superior to feedback displaying the correct answer in a multiple-selection method in that it enhances self-correction. In terms of correction, van der Linden (1993), while examining the strategies learners employed when interacting with different levels of feedback, observed that learners felt motivated to self-correct when they received feedback about the type of error committed. In sum, intelligent feedback is claimed to be effective if (a) it is specific to a learner input, (b) points to the error type, (c) explicitly explains the error, and (d) leads to self-correction.

Even less is known about the potential of automated writing evaluation (AWE) feedback, which is why the controversy in that area is also very prominent. AWE is perceived as a perfect solution by some and as a threat by others. The supporters of AWE use in the classroom argue that the immense advantages of such AWE programs as Criterion by Educational Testing Service (ETS) and the automated essay writing system, Vantage Learning, are their ability to assess and respond to student writing as well as humans do (Attali & Burstein, 2006; Pearson Education, 2007; Vantage Learning, 2007) and to do so in a much more time- and cost-effective way. Theoretically, AWE may be able to motivate and guide student revision and to foster learner autonomy (Chen & Cheng, 2008). It is meant to support process writing approaches that emphasize the value of multiple drafting through scaffolding suggestions and explanations. The integration of AWE into the curriculum is said to be also consistent with the drive toward individualized assessment and instruction (Burstein, Marcu, & Knight, 2003). The developers of these programs promote them as instructional supplements to process writing instruction and as vehicles of consistent writing and evaluation across the curriculum.
Output gives learners the opportunity to produce language and “serves as a means of hypothesis testing” (Gass, 1997, p. 7). After having used a language form unsuccessfully and after having created a certain hypothesis based on the received feedback, learners are pushed to modify their linguistic form and produce more precise and appropriate output—comprehensible output (Swain, 1985). Modified output is useful provided that learners see the connection between their erroneous form, the feedback, and the revised output (Carroll, 2001; Gass & Mackey, 2006). Continued production of output is important because, in the long run, it leads to automaticity.

Traditionally, interaction has referred to learners’ engagement in conversations with interlocutors. Ellis (1999) expanded the idea of interaction from interpersonal level to that of intrapersonal level; “interaction that can occur in our minds, ... and, more covertly, when different modules of the mind interact to construct an understanding of or a response to some phenomenon” (p.1). Later, Chapelle (1998, 2001, 2007) connected these concepts to learner-computer interaction, showing how IA constructs can be enhanced to facilitate language learning with the help of computers. Specifically, the linguistic features in the input can be made salient through highlighting, glosses, hyperlinks, pictorial or video representations, and so forth. The output of AWE is automatically adapted or modified through mouse clicks and hypertext links which provide learners with opportunities to request help on demand and offer them immediate assistance during the learning process in the form of explanations, examples, definitions, concordancing lines, and so on. AWE is a specific example of technology that can be used to operationalize the key IA constructs and thus to encourage learner-computer interaction. AWE systems generate instant feedback that can vary in the degree of specificity and explicitness as well as in the ability to adapt to individual learners. Such feedback may direct learners’ attention to their linguistic infelicities and motivate them to produce more comprehensible output, but this assumption still needs empirical validation, as does the role of feedback in language learning in general.

FEEDBACK ISSUES IN SLA, I/CALL, AND AWE

Although the importance of feedback was articulated in behaviorism (see “knowledge of results” in Thorndike, 1913) long before it was embraced by SLA and learner-centered approaches to L2 writing, it has automatic spell checking is the subject of heated debate. The issues discussed range from providing feedback that corrects all errors (Lalande, 1982) to selective feedback (Bates, Lane, & Lange, 1993; Ferris, 1995) and even to calls for elimination of any degree of feedback that provides negative evidence or corrective feedback (Cook, 1991; Krashen, 1984; Truscott, 1996). Truscott (1996) asserted that “[g]rammar correction should be abandoned” (p. 328) and that error correction in general can be ineffective and can even have deleterious effects on the quality of students’ writing (Truscott, 2004). The empirical evidence on corrective feedback, however, are conflicting and far from being conclusive, and Truscott’s claims have been challenged by Hylten (2005) and Vantage Learning (Ferris, 1999, 2002, 2007; Hyland & Hyland, 1999; Hyland & Pearson, 2004; van der Linden, 2004) who argue that “existing research predicts ... positive effects for written error correction” (p. 50), and Russell and Spada (2006) conclude that such feedback is beneficial for the acquisition of L2 grammar. The incomparability of findings in corrective feedback research may be due to inconsistencies in research design (Ferris, 1999, 2004, 2006; Guenette, 2007) and to the degree of implicitness or explicitness of the information provided in the feedback (Russell & Spada, 2006). Considering the need for well designed experimental and descriptive studies, several pedagogical recommendations, Ellis (2009) outlines a typology of written feedback types, placing them on a continuum between implicit and explicit (see also Ellis, Loewen, & Erelam, 2006).

Computer-assisted language learning (CALL) and intelligent computer-assisted language learning (ICALL) employ various types of feedback that is provided more or less directly. Garrett (1987) classifies it into four categories: (1) only the correct answer is presented, (2) the location of errors is indicated based on a letter-by-letter comparison of the learner’s input with the machine-stored correct version, (3) error messages associated with possible errors are stored in the computer and are presented if the learner’s response matches those possible errors based on an analysis of the anticipated incorrect answers, and (4) problematic or missing items are computed on a linguistic analysis of the learner’s response compared to an analysis derived from relevant grammar rules and lexicon of the target language. The fourth type of feedback, known as intelligent feedback, is much more sophisticated than the pattern-marker and error-anticipation techniques used in other conventional types of CALL feedback because it is capable of responding to multiple problematic aspects of language use that may occur in learner’s production.

The concerns regarding ICALL feedback resonate with the debate on corrective feedback. Most of the existing ICALL programs target the development of learners’ grammatical competence, revolve around the functionality of a parser, and provide immediate intelligent feedback. Their feedback features vary in their degree of specificity and explicitness as well as in their ability to adapt to individual learners (see list of ICALL programs and their feedback features in Appendix A). The question, however, is not what kind of feedback ICALL can generate; rather, it is what kind of feedback ICALL should generate. Nagata (1995) argues that “[i]f we use an intelligent system, we should examine carefully what kind of error messages should be provided ... and how effective they are” (p. 49).

Research on ICALL feedback is still scarce (see Heft & Schulze, 2007), but the relatively few pieces of empirical evidence suggest positive effects of explicit intelligent feedback. When comparing ICALL feedback with CALL feedback, Nagata (1993, 1995) found that the former is more effective than the latter. The more detailed the intelligent feedback is, the better the learning outcomes are (Heft, 2001, 2002, 2004, 2005). Learners appear to show significantly more uptake over time with a more error-specific feedback type (Heft, 2008; Heft & Rimrott, 2008). Yang and Akahori’s (1999) findings indicate that feedback that corresponds to the input created by the learner is superior to feedback displaying the correct answer in a multiple-choice format. It enhances self-correction. In terms of correction, van der Linden (1993), while examining the strategies learners employed when interacting with different levels of feedback, observed that learners felt motivated to self-correct when they received feedback about the type of error committed. In sum, intelligent feedback is claimed to be effective if (a) it is no specific learner input, (b) points to the error type, (c) explicitly explains the error, and (d) leads to self-correction.

Even less is known about the potential of automated writing evaluation (AWE) feedback, which is why the controversy in that area is also very prominent. AWE is perceived as a perfect solution by some and as a threat by others. The supporters of AWE use in the classroom argue that the immense advantages of such AWE programs as Criterion by Educational Testing Service (ETS) and Voyant (2005) and Vantage Learning (2006) make it possible to assess and respond to student writing as well as humans do (Attali & Burstein, 2006; Pearson Education, 2007; Vantage Learning, 2007) and to do so in a much more time- and cost-effective way. Theoretically, AWE may be able to motivate and guide student revision and to foster learner autonomy (Chen & Cheng, 2008). It is meant to support process writing approaches that emphasize the value of multiple drafting through scaffolding suggestions and explanations. The integration of AWE into the curriculum is said to be also consistent with the drive toward individualized assessment and instruction (Burstein, Marcu, & Knight, 2005). The developers of these programs promote them as instructional supplements to process writing instruction and as vehicles of consistent writing and evaluation across the curriculum.
Many of these claims have been questioned by some members of the academic community. L1-writing scholars are particularly skeptical when it comes to implementations of AWE in the classroom (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) and that the machine analyzer, and therefore the feedback it generates, is calibrated to static compositional features and formulaic expressions and heavily subordinates meaning. Therefore, Cheville suspects, AWE use may encourage students to pay more attention to the surface features that are more easily detected by AWE systems than to the construction of meaning for communicative purposes. Questions have also been raised from the 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). Student essays are evaluated automatically against generic writing traits, eliminating the value of human audiences in real-world contexts. "While they [AWE programs] promise consistency, they distort the very nature of writing as a complex and context-rich interaction between people" (Conference on College Composition and Communication, 2006).

AWE use in L2 writing brings about a number of issues as well. First, the actual impact of automated feedback on the development of writing skills has not yet been understood (Warschauer & Ware, 2006). Second, AWE programs were not originally developed for nonnative speakers. Third, existing research studies, of which only a few investigated AWE in L2 contexts (e.g., Chen & Cheng, 2008; Yang, 2004), have focused mainly on outcomes lacking a focus on the learning process and therefore shedding no light on how automated feedback may shape learning to write. Their results either indicate improvement in students' performance or point to superficial revisions. Given the inconclusive nature of the empirical evidence, researcher-teacher dyads usually use AWE programs as only supplements to writing instruction (Shermis & Burstein, 2003; Ware, 2005; Warschauer & Ware, 2006) more than justifiably. This recommendation is hardly satisfactory, however. To benefit the stakeholders and the profession, the design of AWE software needs to draw directly from the empirical evidence of how language is acquired, and AWE research needs to be rooted in SLA theory. The IA is a theoretical framework that can help conceptualize the design of AWE applications for L2 learners and frame AWE research in general. The following is a description of one such application that was designed and evaluated from an IA perspective.

### INTELLIGENT ACADEMIC DISCOURSE EVALUATOR (IADE)

The Intelligent Academic Discourse Evaluator (IADE) is a web-based automated writing evaluation program that can analyze the introduction section of research articles at the level of discourse elements in the student's academic discipline. It is based on automated analysis, it provides learners with individual feedback when they submit their drafts. The characteristics of the targeted instructional context (L2 graduate level academic writing) and considerations as to how IA constructs can be most informative vis-à-vis the development of academic writing were central to the decisions regarding IADE's design (Cotos, 2009).

The targeted instructional context employed a corpus-based approach to teaching academic writing (see Cortes, 2007; Cotos, 2010) in which a corpus of 1,000 research articles in students' disciplines exposed the students to large amounts of input through reading and analysis of the genre. Conducting corpus analyses was meant to help the students notice the characteristics of the academic writing conventions in their field. Then, the students were given the opportunity to produce their own written output in the form of sections of research articles modeled on the patterns observed in the input corpora. Feedback, however, although considered essential by the IA, was limited to in-class teacher-student group explanatory exchanges and occasional comments.

Given that the focus of instruction is on discourse conventions, IADE provides feedback at the level of rhetorical moves in the introductions to research articles. The approach to teaching how to write introductions is based on Swales' (1981, 1990, 2004) genre analyses of work in which the Create-a-Research-Space model was proposed. In this model, introduction sections consist of three moves: Move 1—Establishing a territory, Move 2—Establishing a niche, and Move 3—Occupying the niche. Each move contains a number of steps, and these steps express a particular functional meaning, which may be obligatory and/or optional depending on the norms adopted by field-specific discourse communities (see list of moves and steps in Appendix B).

Although the moves and their possible steps are clear cut, move identification is not as transparent as it may seem. Nwogu (1990) claims that it is more of a bottom-up process, which is at the same time influenced by one's schemata about the structuring of text type and genres. Swales (2004), however, states that certain lexico-grammatical features can indicate certain moves. For instance, the present continuous tense can invoke recency in statements of centrality in Move 1, lexical units with negative connotations can indicate a gap or a problem, and deixis and personal pronouns can signal the onset of Move 3. In other cases, the placement of a discourse piece can help to interpret its function. These insights determined the choice of automated analysis approach applied in IADE.

IADE's analysis module performs automatic identification of introduction discourse moves, approaching this task as a classification problem. The classifier analyzes and classifies each sentence of the text as belonging to a particular move. This classification is done by means of identifying the lexical features that are indicative of a certain move. Then, with the help of preprogrammed scripts, percentages for the move distribution in the student's draft are automatically calculated and compared with the distribution of moves in the corpus of his/her field (see Pendar & Cotos, 2008). The classification into moves and the information about the distribution of moves, both in the student draft and in the corpus, are the sources of two forms of feedback—color-coded and numerical (see sample feedback in Appendix C). IADE's feedback combines a number of characteristics. It is

- immediate (provided immediately, in less than 60 seconds from the time of submission),
- intelligent (generated automatically by a natural language processing based engine),
- specific to the individual (provided to the individual student based on his/ her submission and on its comparison to the respective discipline),
- metalinguistic (provided in definitional terms, as information and comments about the well-formedness of the student's discourse; e.g., "... of your sentences belong to Move 1"),
- short (concise in that it briefly presents the descriptive percentages representing the distribution of Moves in the students' draft and in the introductions of his/her discipline; e.g., "This is below average compared to Move 1 in your discipline, where the minimum is 45.455%, the average is 65.799%, and the maximum is 87.097%"),
- negative (points to drawbacks in the student's discourse; e.g., "This is below average (or above average) compared to Move 1 in your discipline..."),
articles modeled on the patterns observed in the input corpora. Feedback, however, although considered essential by the IA, was limited to in-class teacher-student group explanatory exchanges and occasional comments.

Given that the focus of instruction is on discourse conventions, IADE provides feedback at the level of rhetorical moves in the introductions to research articles. The approach to teaching how to write Introductions is based on Swales’ (1981, 1990, 2004) genre analysis work in which the Create-a-Research-Space model was proposed. In this model, introduction sections consist of three moves: Move 1—Establishing a territory, Move 2—Establishing a niche, and Move 3—Occupying the niche. Each move contains a number of steps, and these steps express a particular functional meaning, which may be obligatory and/or optional depending on the norms adopted by field-specific discourse communities (see list of moves and steps in Appendix B).

Although the moves and their possible steps are clear cut, move identification is not as transparent as it may seem. Nwogu (1990) claims that it is more of a bottom-up process, which is at the same time influenced by one’s schemata about the structuring of text type and genres. Swales (2004), however, states that certain lexicogrammatical features can indicate certain moves. For instance, the present continuous tense can invoke recency in statements of centrality in Move 1, lexical units with negative connotations can indicate a gap or a problem, and deictics and personal pronouns can signal the onset of Move 3. In other cases, the placement of a discourse piece can help interpret its function. These insights determined the choice of automated analysis approach applied in IADE.

IADE’s analysis module performs automatic identification of introduction discourse moves, approaching this task as a classification problem. The classifier analyzes and classifies each sentence of the submitted text as belonging to a particular move. This classification is done by means of identifying the lexical features that are indicative of a certain move. Then, with the help of preprogrammed scripts, percentages for the move distribution in the student’s draft are automatically calculated and compared with the distribution of moves in the corpus of his/her academic field (see Pendar & Cotos, 2008). The classification into moves and the information about the distribution of moves, both in the student draft and in the corpus, are the sources of two forms of feedback—color-coded and numerical (see sample feedback in Appendix C). IADE’s feedback combines a number of characteristics. It is:

- immediate (provided immediately, in less than 60 seconds from the time of submission);
- intelligent (generated automatically by a natural language processing based engine);
- specific to the individual (provided to the individual student based on his/ her submission and on its comparison to the respective discipline);
- metalinguistic (provided in definitional terms, as information and comments about the well-formedness of the student’s discourse; i.e., “... of your sentences belong to Move 1”);
- short (concise in that it briefly presents the descriptive percentages representing the distribution of Moves in the students’ draft and in the introductions of his/her discipline; e.g., “This is below average compared to Move 1 in your discipline, where the minimum is 45.455%, the average is 65.799%, and the maximum is 87.097%);
- negative (points to drawbacks in the student’s discourse; e.g., “This is below average (or above average) compared to Move 1 in your discipline...”).
The process of feedback generation was conceptualized with a focus on IA constructs. IADE analyzes learners’ output and then uses it to generate feedback that is returned to them as modified input. The color codes serve as input enhancement designed to encourage noticing and focus on discourse form. Intended to stimulate learner-computer interaction during the writing process, IADE’s feedback either confirms learners’ communicative success or points to shortcomings in their production that is meant to trigger testing hypotheses that learners may generate with regards to the nature of their linguistic problem.

THE STUDY

The study reported here is part of a larger project aimed at evaluating the effectiveness of IADE and its automated feedback. The conceptual framework that overarched this work is Chapelle’s (2001) CALL evaluation framework, which draws “from theory and research on conditions for instructed SLA” (p. 54). This framework puts forth six CALL qualities, or criteria, for evaluation: language learning potential, meaning focus, learner fit, impact, authenticity, and practicality. This article focuses on evidence of language learning potential (LLP). Chapelle (2001) defines LLP as “the extent to which the task promotes beneficial focus on form” (p. 55), which is believed to be one of the most important conditions for language acquisition. Here, LLP is defined as the potential of the feedback to promote beneficial focus on discourse form and to enhance learning. Discourse form is referred to as the rhetorical moves on which the feedback is provided; evidence of learning is sought in the learning gains (i.e., how well the participants acquired the moves after having revised their writing with IADE) and the improvement in learners’ written performance. The point of interest was to find whether IADE’s feedback can draw learners’ attention to the moves and whether the focus on discourse form would lead to intrapersonal interaction resulting in the learning of moves. Two research questions were posed:

1. Is there evidence that IADE’s feedback can stimulate students’ noticing of and focus on discourse form?
2. If yes, is there evidence suggesting that students acquire the target discourse forms that were focused on during the interaction with IADE’s feedback?

Methods

To answer the two questions, the study pursued a process-product research direction that encompasses “the interaction between [technology] use and outcome” (Warschauer & Ware, 2006, p.10). It employed a mixed-methods approach with a concurrent transformative strategy guided by theory (Creswell, 2003). This approach is diagrammed in Figure 1. Under the umbrella of Chapelle’s (2001) CALL evaluation framework representing instructed SLA, the quantitative and qualitative data, enumerated in the first two oval shapes, were collected concurrently and integrated during the analysis and interpretation stage, shown in the circle on the right. The data were triangulated to best converge the information containing evidence for LLP.

Figure 1

The Concurrent Transformative Strategy Applied in this Study

Participants

The participants were 105 international graduate students (59 male and 46 female) who were enrolled in a graduate academic writing course at a large US university. Thirty-nine were Masters students, and 66 were Ph.D. students specializing in a total of 34 disciplines. Table 1 summarizes the characteristics of participating students (number of students in each category written in parentheses).

Table 1

Participants' Characteristics

<table>
<thead>
<tr>
<th>Age</th>
<th>L1</th>
<th>Other languages</th>
<th># semesters at US university</th>
<th>TOEFL scores (test type)</th>
<th>Wrote research articles in L1</th>
<th>Wrote research articles in English</th>
</tr>
</thead>
<tbody>
<tr>
<td>22 (9)</td>
<td>Chinese (72)</td>
<td>1 (79)</td>
<td>1 (67)</td>
<td>80-104 (IBT)</td>
<td>43</td>
<td>Yes</td>
</tr>
<tr>
<td>23 (8)</td>
<td>Korean (11)</td>
<td>2 (24)</td>
<td>2 (19)</td>
<td>243-255 (CBT)</td>
<td>22 not (14 not published)</td>
<td></td>
</tr>
<tr>
<td>24 (10)</td>
<td>Spanish (7)</td>
<td>3 (2)</td>
<td>3 (10)</td>
<td>520-667 (PBT)</td>
<td>62</td>
<td></td>
</tr>
<tr>
<td>25 (15)</td>
<td>Turkish (4)</td>
<td>4 (4)</td>
<td>520-667 (PBT)</td>
<td>19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26 (14)</td>
<td>Arabic (3)</td>
<td>5 (5)</td>
<td>44</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27 (6)</td>
<td>Portuguese (2)</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28 (7)</td>
<td>Tigrigna (2)</td>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29 (5)</td>
<td>Filipino (1)</td>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30 (4)</td>
<td>Greek (1)</td>
<td>67</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>31 (9)</td>
<td>Italian (1)</td>
<td>64</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32 (2)</td>
<td>Thai (1)</td>
<td>64</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>33 (4)</td>
<td>64</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>34 (3)</td>
<td>64</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>35 (2)</td>
<td>64</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>36 (1)</td>
<td>64</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>37 (3)</td>
<td>64</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>38 (1)</td>
<td>64</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>39 (1)</td>
<td>64</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40 (1)</td>
<td>64</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>41 (1)</td>
<td>64</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The process of feedback generation was conceptualized with a focus on IA constructs. IADE analyzes learners’ output and then uses it to generate feedback that is returned to them as modified input. The color codes serve as input enhancement designed to encourage noticing and focus on discourse form. Intended to stimulate learner-computer interaction during the writing process, IADE’s feedback either confirms learners’ communicative success or points to shortcomings in their production that is meant to trigger testing hypotheses that learners may generate with regards to the nature of their linguistic problem.4

THE STUDY

The study reported here is part of a larger project aimed at evaluating the effectiveness of IADE and its automated feedback. The conceptual framework that overarched this work is Chapelle’s (2001) CALL evaluation framework, which draws “from theory and research on conditions for instructed SLA” (p. 54). This framework puts forth six CALL qualities, or criteria, for evaluation: language learning potential, meaning focus, learner fit, impact, authenticity, and practicality. This article focuses on evidence of language learning potential (LLP). Chapelle (2001) defines LLP as “the extent to which the task promotes beneficial focus on form” (p. 55), which is believed to be one of the most important conditions for language acquisition. Here, LLP is defined as the potential of the feedback to promote beneficial focus on discourse form and to enhance learning. Discourse form is referred to as the rhetorical moves on which the feedback is provided; evidence of learning is sought in the learning gains (i.e., how well the participants acquired the moves after having revised their writing with IADE) and the improvement in learners’ written performance. The point of interest was to find whether IADE’s feedback can draw learners’ attention to the moves and whether the focus on discourse form would lead to intrapersonal interaction resulting in the learning of moves. Two research questions were posed:

1. Is there evidence that IADE’s feedback can stimulate students’ noticing of and focus on discourse form?
2. If yes, is there evidence suggesting that students acquire the target discourse forms that were focused on during the interaction with IADE’s feedback?

Methods

To answer the two questions, the study pursued a process-product research direction that encompasses “the interaction between [technology] use and outcome” (Warschauer & Ware, 2006, p.10). It employed a mixed-methods approach with a concurrent transformative strategy guided by theory (Creswell, 2003). This approach is diagrammed in Figure 1. Under the umbrella of Chapelle’s (2001) CALL evaluation framework representing instructed SLA, the quantitative and qualitative data, enumerated in the first two oval shapes, were collected concurrently and integrated during the analysis and interpretation stage, shown in the circle on the right. The data were triangulated to best converge the information containing evidence for LLP.

Table 1

<table>
<thead>
<tr>
<th>Participants’ Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
</tr>
<tr>
<td>22 (9)</td>
</tr>
<tr>
<td>23 (8)</td>
</tr>
<tr>
<td>24 (10)</td>
</tr>
<tr>
<td>25 (15)</td>
</tr>
<tr>
<td>26 (14)</td>
</tr>
<tr>
<td>27 (6)</td>
</tr>
<tr>
<td>28 (7)</td>
</tr>
<tr>
<td>29 (5)</td>
</tr>
<tr>
<td>30 (4)</td>
</tr>
<tr>
<td>31 (9)</td>
</tr>
<tr>
<td>32 (2)</td>
</tr>
<tr>
<td>33 (4)</td>
</tr>
<tr>
<td>34 (3)</td>
</tr>
<tr>
<td>35 (2)</td>
</tr>
<tr>
<td>37 (3)</td>
</tr>
<tr>
<td>38 (1)</td>
</tr>
<tr>
<td>39 (1)</td>
</tr>
<tr>
<td>40 (1)</td>
</tr>
<tr>
<td>41 (1)</td>
</tr>
</tbody>
</table>

Figure 1
The Concurrent Transformative Strategy Applied in this Study

Descriptive statistics, t tests, quasistatistics (105 participants)
Manual analysis, coding (16 participants)

Participants
The participants were 105 international graduate students (59 male and 46 female) who were enrolled in a graduate academic writing course at a large US university. Thirty-nine were Masters students, and 66 were Ph.D. students specializing in a total of 34 disciplines. Table 1 summarizes the characteristics of participating students (number of students in each category written in parentheses).
Instruments and Materials

IADE
The IADE program was utilized both as the core material used by students to complete the revision task and as a data collection instrument. Once the students signed in, its database stored the following information:

1. students’ first and last names,
2. students’ automatically assigned ID numbers,¹
3. students’ degree program (MA or Ph.D.),
4. students’ academic disciplines,
5. all the drafts submitted by each student,
6. the number of drafts submitted by each student,
7. the automated analysis and feedback generated for every draft, and
8. the date and time of draft submission.

Pre-/posttests
To measure the students’ knowledge of the moves and functions (or steps in Swales’ terms) before and after their interaction with IADE’s feedback, a pretest and a posttest were developed. Both tests consisted of two tasks. The first task required the students to name the moves and the steps in a number of given decontextualized examples. The second task focused on annotating each sentence of a research article introduction in terms of moves and steps by using the ‘Insert comment’ function in Word documents. Because students had the entire text, they could make their judgments about the function of each sentence based on the context. The number of sentences in both tasks of the pre- and posttests was the same. To ensure that the texts were not burdened by discipline-specific terminology, they were selected after being pilot-ed with a group of 17 international graduate students prior to this study.

Survey questions: yes/no, open-ended, and Likert scale
The survey contained eight questions eliciting information about the students’ characteristics such as age, gender, first language, knowledge of other languages, period of study at a US university prior to taking the academic writing course, TOEFL score, and research article writing experience. Eight yes/no and open-ended follow-up questions elicited information related to the LLP of IADE’s feedback by inquiring about students’ focus on discourse form as well as about their perceptions of learning and self-improvement.

The first Likert-scale question asked the students to assess their general level of English language proficiency on a scale of excellent, very good, good, fair, and poor. The other five LLP-related questions offered a choice of four answers, which, depending on the question, were: a lot or very well, somewhat or well, a little, and not at all (see survey questions in Appendix D).

Think-aloud protocols
For the purpose of introspective data collection, 16 participants were asked to think aloud (see Ericsson & Simon, 1987) while revising their drafts with IADE. A short demonstration of a think-aloud protocol was performed for each participant. During the think-aloud sessions, which ranged from 25 to 37 minutes, the researcher also prompted the students to verbalize their thoughts whenever it was necessary. The audio recording function of Camtasia Studio 5 software by TechSmith was used to record the participants.

Screen recording
Camtasia’s screen recording function was also used to capture the data on participants’ interaction with IADE in the form of files containing a video record of all the actions visible on the participant’s computer screen.

Observations
The researcher conducted observations of the same 16 students who participated in the think-aloud sessions. The researcher sat to the right of the participant at a distance from which she could see both the student and the computer screen. Notes about each participant’s behavior (e.g., cursor movements, verbal reactions, and body language) during the interaction with IADE were made on paper, and question marks were put next to the entries that required further clarification. The length of the observation notes ranged from one to two pages of 12-point Times New Roman single-spaced text.

Semistructured interviews
The semistructured interviews contained questions about participants’ actions and/or utterances that were marked with a question mark in the observation notes. Those were potential points of interest that could give a better introspective insight into the nature of observed instances.

Procedure
IADE was implemented in the classroom as a revision tool. First, the participants received instruction on the writing conventions of research article introductions based on Swales’ (1981, 1990, 2004) move schema. One class period was devoted to studying each move and to corpus-based work on a given move. Then, the students took the pretest, which was delivered at this time as opposed to prior to instruction because the intent was to measure not the learning gains after instruction, but rather after revision with IADE’s feedback. After the pretest, the students were required to write a draft of the introduction section for their own research articles as homework.

The next class period, the instructor introduced IADE and modeled how to interact with it. The interaction consisted of submitting the draft for automated evaluation, receiving immediate individualized feedback, making revisions, and resubmitting the new draft to the system. This was an iterative process that began in class and ended outside of class and that allowed the students to spend additional time on practice and revision. During the following class session, when the final introduction draft was due for submission, the posttest was given. After the posttest, the participants answered the survey questions.²

¹ Students did not have access to their automatically assigned ID numbers.
² Students returned the surveys to the research assistant in a sealed envelope at the beginning of the first class session after reading the consent form and ethics information.
Instruments and Materials
IADE
The IADE program was utilized both as the core material used by students to complete the revision task and as a data collection instrument. Once the students signed in, its database stored the following information:

1. students’ first and last names,
2. students’ automatically assigned ID numbers,
3. students’ degree program (MA or Ph.D.),
4. students’ academic disciplines,
5. all the drafts submitted by each student,
6. the number of drafts submitted by each student,
7. the automated analysis and feedback generated for every draft, and
8. the date and time of draft submission.

Pre-/posttests
To measure the students’ knowledge of the moves and functions (or steps in Swales’ terms) before and after their interaction with IADE’s feedback, a pretest and a posttest were developed. Both tests consisted of two tasks. The first task required the students to name the moves and the steps in a number of given decontextualized examples. The second task focused on annotating each sentence of a research article introduction in terms of moves and steps by using the ‘Insert comment’ function in Word documents. Because students had the entire text, they could make their judgments about the function of each sentence based on the context. The number of sentences in both tasks of the pre- and posttests was the same. To ensure that the texts were not burdened by discipline-specific terminology, they were selected after being pilot-ed with a group of 17 international graduate students prior to this study.

Survey questions: yes/no, open-ended, and Likert scale
The survey contained eight questions eliciting information about the students’ characteristics such as age, gender, first language, knowledge of other languages, period of study at a US university prior to taking the academic writing course, TOEFL score, and research article writing experience. Eight yes/no and open-ended follow-up questions elicited information related to the LLP of IADE’s feedback by inquiring about students’ focus on discourse form as well as about their perceptions of learning and self-improvement.

The first Likert-scale question asked the students to assess their general level of English language proficiency on a scale of excellent, very good, good, fair, and poor. The other five LLP-related questions offered a choice of four answers, which, depending on the question, were: a lot or very well, somewhat or well, a little, and not at all (see survey questions in Appendix D).

Think-aloud protocols
For the purpose of introspective data collection, 16 participants were asked to think aloud (see Ericsson & Simon, 1987) while revising their drafts with IADE. A short demonstration of a think-aloud protocol was performed for each participant. During the think-aloud sessions, which ranged from 25 to 37 minutes, the researcher also prompted the students to verbalize their thoughts whenever it was necessary. The audio recording function of Camtasia Studio 5 software by TechSmith was used to record the participants.

Screen recording
Camtasia’s screen recording function was also used to capture the data on participants’ interaction with IADE in the form of files containing a video record of all the actions visible on the participant’s computer screen.

Observations
The researcher conducted observations of the same 16 students who participated in the think-aloud sessions. The researcher sat to the right of the participant at a distance from which she could see both the student and the computer screen. Notes about each participant’s behavior (e.g., cursor movements, verbal reactions, and body language) during the interaction with IADE were made on paper, and question marks were put next to the entries that required further clarification. The length of the observation notes ranged from one to two pages of 12-point Times New Roman single-spaced text.

Semistructured interviews
The semistructured interviews contained questions about participants’ actions and/or utterances that were marked with a question mark in the observation notes. Those were potential points of interest that could give a better introspective insight into the nature of observed instances.

Procedure
IADE was implemented in the classroom as a revision tool. First, the participants received instruction on the writing conventions of research article introductions based on Swales’ (1981, 1990, 2004) move schema. One class period was devoted to studying each move and to corpus-based work on a given move. Then, the students took the pretest, which was delivered at this time as opposed to prior to instruction because the intent was to measure not the learning gains after instruction, but rather after revision with IADE’s feedback. After the pretest, the students were required to write a draft of the introduction section for their own research articles as homework.

The next class period, the instructor introduced IADE and modeled how to interact with it. The interaction consisted of submitting the draft for automated evaluation, receiving immediate individualized feedback, making revisions, and resubmitting the new draft to the system. This was an iterative process that began in class and ended outside of class and that allowed the students to spend additional time on practice and revision. During the following class session, when the final introduction draft was due for submission, the posttest was given. After the posttest, the participants answered the survey questions.
Sixteen participants volunteered to complete the revision task with IADE in an experimental computer lab setting under observable conditions and in the presence of the researcher, a process that allowed for the collection of concurrent-revision data in the form of think-aloud protocols, Camtasia screen recordings, observations, and semistructured interviews.

The data are shown in Table 2. It should be clarified here that the 16 first and 16 final drafts analyzed manually were written by the same 16 participants who volunteered to use IADE in an experimental setting. The writing of these particular participants was chosen so that the results of the analysis could be triangulated with their think-aloud, observation, interview, and screen-captured data. Also, the same drafts were scored by human raters, who had no knowledge of which drafts were first and which were final.

Table 2
Summary of Data

<table>
<thead>
<tr>
<th>Data source</th>
<th># participants</th>
<th>Extent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quantitative</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Pretest and posttest</td>
<td>104</td>
<td>2 tasks (17 sentences each)</td>
</tr>
<tr>
<td>2. Yes/no survey questions</td>
<td>83</td>
<td>8</td>
</tr>
<tr>
<td>3. Likert-scale survey questions</td>
<td>88</td>
<td>5</td>
</tr>
<tr>
<td>4. Automated evaluation of first and final drafts</td>
<td>105</td>
<td>% and comment comparison to discipline for 210 drafts</td>
</tr>
<tr>
<td>5. Human evaluation of first and final drafts</td>
<td>16</td>
<td>32 drafts</td>
</tr>
<tr>
<td>Qualitative</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Open-ended survey questions</td>
<td>83</td>
<td>12</td>
</tr>
<tr>
<td>2. Think-aloud protocols</td>
<td>16</td>
<td>25-38 minutes</td>
</tr>
<tr>
<td>3. Camtasia recordings</td>
<td>16</td>
<td>25-38 minutes</td>
</tr>
<tr>
<td>4. Observations</td>
<td>16</td>
<td>25-38 minutes</td>
</tr>
<tr>
<td>5. Semistructured interviews</td>
<td>16</td>
<td>10-15 minutes</td>
</tr>
</tbody>
</table>

Table 3 summarizes what analyses were performed and what claims could be made regarding the potential of IADE’s feedback to enhance focus on and learning of discourse forms.

Table 3
Summary of Data Analysis

<table>
<thead>
<tr>
<th>Claims</th>
<th>Data and analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Automated feedback stimulated focus on discourse form and noticing of negative evidence</td>
<td>Quantitative: yes/no responses; Likert-scale responses; comparison of response percentages</td>
</tr>
<tr>
<td>Qualitative</td>
<td>Qualitative: open-ended survey responses; think-aloud protocols; observations; semistructured interviews</td>
</tr>
<tr>
<td>Focus on discourse form triggered by automated feedback lead to learning gains</td>
<td>Quantitative: yes/no responses; Likert-scale responses; pre- and posttest scores; descriptive statistics; t tests for pre- and posttest scores</td>
</tr>
<tr>
<td>Qualitative</td>
<td>Qualitative: open-ended survey responses; manual analysis of responses; quasistatistics of emerging themes</td>
</tr>
<tr>
<td>Focus on discourse form triggered by automated feedback lead to improvement in the rhetorical quality of writing</td>
<td>Quantitative: yes/no responses; Likert-scale responses; automated evaluation of all first and final drafts converted to scores; human rater evaluation of 16 first and final drafts; manual analysis of responses; coding; manual analysis of output modifications in 16 first and 16 final drafts; quasistatistics of emerging themes</td>
</tr>
</tbody>
</table>

Survey questions

Percentages for yes/no and Likert-scale responses were calculated and compared. The four Likert-scale response choices were interpreted as follows: 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. Participants’ responses to the open-ended questions were analyzed by identifying emerging themes which were then quantified in terms of the percentage of students who mentioned them.

Pre-/posttests

The pre- and posttests were scored for every sentence in each of the two tasks. The decontextualized sentences in task one had unambiguous and clearly expressed functions realized
Sixteen participants volunteered to complete the revision task with IADE in an experimental computer lab setting under observable conditions and in the presence of the researcher, a process that allowed for the collection of concurrent-revision data in the form of think-aloud protocols, Camtasia screen recordings, observations, and semistructured interviews.

The data are shown in Table 2. It should be clarified here that the 16 first and 16 final drafts analyzed manually were written by the same 16 participants who volunteered to use IADE in an experimental setting. The writing of these particular participants was chosen so that the results of the analysis could be triangulated with their think-aloud, observation, interview, and screen-captured data. Also, the same drafts were scored by human raters, who had no knowledge of which drafts were first and which were final.

Table 2 Summary of Data

<table>
<thead>
<tr>
<th>Data source</th>
<th># participants</th>
<th>Extent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quantitative</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Pretest and posttest</td>
<td>104</td>
<td>2 tasks (17 sentences each)</td>
</tr>
<tr>
<td>2. Yes/no survey questions</td>
<td>83</td>
<td>8</td>
</tr>
<tr>
<td>3. Likert-scale survey questions</td>
<td>88</td>
<td>5</td>
</tr>
<tr>
<td>4. Automated evaluation of first</td>
<td>105</td>
<td>% and comment comparison to</td>
</tr>
<tr>
<td>and final drafts</td>
<td></td>
<td>discipline for 210 drafts</td>
</tr>
<tr>
<td>5. Human evaluation of first and</td>
<td>16</td>
<td>32 drafts</td>
</tr>
<tr>
<td>final drafts</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Qualitative</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Open-ended survey questions</td>
<td>83</td>
<td>12</td>
</tr>
<tr>
<td>2. Think-aloud protocols</td>
<td>16</td>
<td>25-38 minutes</td>
</tr>
<tr>
<td>3. Camtasia recordings</td>
<td>16</td>
<td>25-38 minutes</td>
</tr>
<tr>
<td>4. Observations</td>
<td>16</td>
<td>25-38 minutes</td>
</tr>
<tr>
<td>5. Semistructured interviews</td>
<td>16</td>
<td>10-15 minutes</td>
</tr>
</tbody>
</table>

Data Analysis

Table 3 summarizes what analyses were performed and what claims could be made regarding the potential of IADE’s feedback to enhance focus on and learning of discourse forms.

Table 3 Summary of Data Analysis

<table>
<thead>
<tr>
<th>Claims</th>
<th>Data and analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Automated feedback stimulated</td>
<td>Quantitative: yes/no responses; Likert-scale responses</td>
</tr>
<tr>
<td>focus on discourse form</td>
<td>comparison of response percentages</td>
</tr>
<tr>
<td>and noticing of negative evidence</td>
<td>Qualitative: open-ended survey responses; think-aloud</td>
</tr>
<tr>
<td></td>
<td>protocols; observations; semistructured interviews</td>
</tr>
<tr>
<td>Focus on discourse form</td>
<td>Quantitative: yes/no responses; Likert-scale responses;</td>
</tr>
<tr>
<td>triggered by automated</td>
<td>pre- and posttest scores</td>
</tr>
<tr>
<td>feedback lead to learning gains</td>
<td>manual analysis of responses; coding; quasistatistics of</td>
</tr>
<tr>
<td></td>
<td>emerging themes</td>
</tr>
<tr>
<td>Focus on discourse form</td>
<td>Quantitative: yes/no responses; Likert-scale responses;</td>
</tr>
<tr>
<td>triggered by automated</td>
<td>automated evaluation of all first and final drafts</td>
</tr>
<tr>
<td>feedback lead to improvement in</td>
<td>human rater evaluation of 16 first and final drafts</td>
</tr>
<tr>
<td>the rhetorical quality of writing</td>
<td>Qualitative: open-ended survey responses; think-aloud/</td>
</tr>
<tr>
<td></td>
<td>Camtasia transcripts; student first and final drafts</td>
</tr>
</tbody>
</table>
| Survey questions                 | Percentages for yes/no and Likert-scale responses were calculated and compared. The four Likert-scale response choices were interpreted as follows: 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. Participants' responses to the open-ended questions were analyzed by identifying emerging themes which were then quantified in terms of the percentage of students who mentioned them.

Pre-/posttests

The pre- and posttests were scored for every sentence in each of the two tasks. The decontextualized sentences in task one had unambiguous and clearly expressed functions realized...
through vocabulary that signaled a certain move and step. Similarly, the texts chosen for the second task had a clear rhetorical development signposted by functional lexical items. Only one correct answer was possible for each sentence in both test tasks. A score of 2 was assigned for a correct move and a correct step; 1 for an incorrect move but a correct step; or for a correct move but an incorrect step; and 0 for an incorrect move and an incorrect step. Descriptive statistics and t tests were calculated for each task as well as for overall test scores.

Automated evaluation

Because IADE does not give scores, the comments in the feedback were used to assign the following scores, which helped determine improvement from first to final draft.

IADE comment: “about average” → score: 3
IADE comment: “below average” or “above average” → score: 2
IADE comment: “way below average” or “way above average” → score: 1

The score of 1 was the lowest score, and 3 was the highest since it represented a range closest to the discipline average. Then, descriptive statistics were calculated, and the mean scores for Move 1, Move 2, Move 3, and draft length were compared through t tests. The scores were also used to classify improvement into four categories as shown in Table 4.

Table 4
Categories of Improvement in Rhetorical Moves and Draft Length

<table>
<thead>
<tr>
<th>First draft scores</th>
<th>Final draft scores</th>
<th>Improvement category</th>
</tr>
</thead>
<tbody>
<tr>
<td>1, 2</td>
<td>3</td>
<td>Considerable improve</td>
</tr>
<tr>
<td>1, 2 or 3, 2</td>
<td>2, 3 (respectfully)</td>
<td>Noticeable improve</td>
</tr>
<tr>
<td>1, 2 or 3, 2</td>
<td>1, 2 or 1, 2 (respectively)</td>
<td>No improvement</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>No need for improve</td>
</tr>
</tbody>
</table>

Human rater evaluation

To find whether improvement also occurred in rhetorical quality, the first and final drafts of the 16 volunteering students were scored by two raters who were trained to use the rubric given in Appendix E. Table 5 shows the agreement between raters on each move, which resulted from calculations of Cohen’s kappa coefficient (κ). All the coefficients indicate a good level of agreement between the raters. In cases of disagreement, the author acted as a third rater.

Table 5
 Interrater Reliability

<table>
<thead>
<tr>
<th>Rater 1 - Rater 2</th>
<th>N agreed</th>
<th>Move 1</th>
<th>Move 2</th>
<th>Move 3</th>
<th>All moves</th>
</tr>
</thead>
<tbody>
<tr>
<td>k</td>
<td>.633</td>
<td>.719</td>
<td>.757</td>
<td>.714</td>
<td></td>
</tr>
</tbody>
</table>

Consistent with the improvement analysis based on IADE’s evaluation, descriptive statistics and t tests with the scores assigned by human raters to the 16 students’ moves in first and final drafts were also calculated. Then, the scores were classified into the improvement categories shown in Table 4 above.

RESULTS AND DISCUSSION

Focus on Discourse Form and Noticing

Evidence that automated feedback stimulated focus on discourse form and noticing of negative evidence were obtained from multiple sources (see Table 6). The think-aloud transcripts contained a total of 1,227 idea units, the interview transcripts 233 idea units, and the observation transcripts 460 idea units. Of these, 484, 63, and 118 idea units, respectively, contained evidence of LLP.

Table 6
Overall Evidence of Focus on Discourse Form and Noticing

<table>
<thead>
<tr>
<th>Data source</th>
<th># participants</th>
<th>Responses to</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likert-scale questions</td>
<td>88</td>
<td>Q-n 1 92.05%</td>
</tr>
<tr>
<td>Q-n 2 100.00%</td>
<td>Q-n 3 90.91%</td>
<td>9.09%</td>
</tr>
<tr>
<td>Yes/no and open-ended questions</td>
<td>83</td>
<td>Q-n 1 100.00%</td>
</tr>
<tr>
<td>Q-n 2 89.16%</td>
<td></td>
<td>10.84%</td>
</tr>
<tr>
<td>Think-aloud protocols/Camtasia recordings</td>
<td>16</td>
<td>Transcripts of 16 think-aloud/Camtasia recordings</td>
</tr>
<tr>
<td>Semistructured interviews</td>
<td>16</td>
<td>16 interviews</td>
</tr>
<tr>
<td>Observations</td>
<td>16</td>
<td>16 observations</td>
</tr>
</tbody>
</table>

For survey questions, % stands for percentage of participants in whose responses evidence was found.

Think-aloud protocols, observations, and semistructured interviews

All these data were transcribed, and the analysis was done according to a coding taxonomy developed for LLP considering IA constructs and based on the results of the pilot conducted prior to this study. The coding categories were

- focus on discourse form,
- noticing of negative evidence,
- output modification, and
- enhanced understanding.

For coding, data were segmented into semantic units, more precisely, “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 required extensive training; however, to ensure the reliability of coding, the author coded the think-aloud protocols, interviews, and observations from the pilot study data twice with an interval of eight months, which helped confirm and refine the initial coding categories (Cohen’s κ = .886). Quasistatistics were calculated for each coding category. Also, analytic induction (Katz, 1983, 2001) was employed to formulate procedural hypotheses for all instances of the observed phenomena, and logical analysis (Miles & Huberman, 1994) was used to generalize causation.
through vocabulary that signaled a certain move and step. Similarly, the texts chosen for the second task had a clear rhetorical development signposted by functional lexical items. Only one correct answer was possible for each sentence in both test tasks. A score of 2 was assigned for a correct move and a correct step; 1 for an incorrect move but a correct step, or for a correct move but an incorrect step; and 0 for an incorrect move and an incorrect step. Descriptive statistics and t-tests were calculated for each task as well as for overall test scores.

Automated evaluation

Because IADE does not give scores, the comments in the feedback were used to assign the following scores, which helped determine improvement from first to final draft.

- IADE comment: "about average" → score: 3
- IADE comment: "below average" or "above average" → score: 2
- IADE comment: "way below average" or "way above average" → score: 1

The score of 1 was the lowest score, and 3 was the highest since it represented a range closest to the discipline average. Then, descriptive statistics were calculated, and the mean scores for Move 1, Move 2, Move 3, and draft length were compared through t-tests. The scores were also used to classify improvement into four categories as shown in Table 4:

### Table 4: Categories of Improvement in Rhetorical Moves and Draft Length

<table>
<thead>
<tr>
<th>First draft scores</th>
<th>Final draft scores</th>
<th>Improvement category</th>
</tr>
</thead>
<tbody>
<tr>
<td>1, 2</td>
<td>3</td>
<td>Considerable improvement</td>
</tr>
<tr>
<td>1, 2 or 3, 2</td>
<td>1, 2 or 2, 1 (respectively)</td>
<td>Noticeable improvement</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>No improvement</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>No need for improvement</td>
</tr>
</tbody>
</table>

Human rater evaluation

To find whether improvement also occurred in rhetorical quality, the first and final drafts of the 16 volunteering students were scored by two raters who were trained to use the rubric given in Appendix E. Table 5 shows the agreement between raters on each move, which resulted from calculations of Cohen’s kappa coefficient (κ). The coefficients indicate a good level of agreement between the raters. In cases of disagreement, the author acted as a third rater.

### Table 5: Interrater Reliability

<table>
<thead>
<tr>
<th>Move</th>
<th>All moves</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater 1 - Rater 2</td>
<td>N agreed</td>
</tr>
<tr>
<td></td>
<td>k</td>
</tr>
<tr>
<td>Move 1</td>
<td>25</td>
</tr>
<tr>
<td>Move 2</td>
<td>26</td>
</tr>
<tr>
<td>Move 3</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>All moves</td>
</tr>
</tbody>
</table>

Consistent with the improvement analysis based on IADE’s evaluation, descriptive statistics and t-tests with the scores assigned by human raters to the 16 students’ moves in first and final drafts were also calculated. Then, the scores were classified into the improvement categories shown in Table 4 above.
Positive evidence of focus on discourse form in the yes/no survey data amounted to 87%. In the Likert-scale responses, as detailed in Table 7, excellent evidence of such focus averaged 44.3%, good 47.7%, weak 6.8%, and poor 1.1%. The participants also positively self-evaluated the degree to which they noticed negative evidence in their own writing, noticed vocabulary indicative of a particular move, and engaged in interactional modification (Ellis, 1999) by changing their written output to address the negative evidence, which constitutes additional positive evidence of noticing and focus on form.

Table 7
Evidence of Noticing and Focus on Discourse Form in Likert-Scale Data (N = 88)

<table>
<thead>
<tr>
<th>Not at all (poor)</th>
<th>A little (weak)</th>
<th>Much (good)</th>
<th>Very much (excellent)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Focused on discourse form</td>
<td>1.10%</td>
<td>6.80%</td>
<td>47.70%</td>
</tr>
<tr>
<td>Noticed negative evidence</td>
<td>1.14%</td>
<td>7.95%</td>
<td>50.00%</td>
</tr>
<tr>
<td>Noted vocabulary indicative of moves</td>
<td>0.00%</td>
<td>9.09%</td>
<td>52.27%</td>
</tr>
<tr>
<td>Engaged in modified interaction</td>
<td>5.68%</td>
<td>44.32%</td>
<td>26.14%</td>
</tr>
<tr>
<td>Average</td>
<td>1.98%</td>
<td>17.05%</td>
<td>44.03%</td>
</tr>
</tbody>
</table>

Table 8 lists a summary of the themes that emerged in participants' open-ended survey responses.

Table 8
Evidence of Focus on Discourse Form in the Open-ended Survey Data (N = 83)

<table>
<thead>
<tr>
<th>Evidence</th>
<th>No evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Focused on discourse form</td>
<td>100.00%</td>
</tr>
<tr>
<td>Feedback triggered focus on discourse form</td>
<td>95.19%</td>
</tr>
<tr>
<td>- color-coded feedback</td>
<td>59.04%</td>
</tr>
<tr>
<td>- numerical feedback</td>
<td>12.05%</td>
</tr>
<tr>
<td>- color-coded and numerical feedback</td>
<td>24.10%</td>
</tr>
<tr>
<td>Noticed peculiarities of discourse form</td>
<td>89.16%</td>
</tr>
<tr>
<td>Engaged in modified interaction</td>
<td>78.30%</td>
</tr>
<tr>
<td>Average</td>
<td>90.70%</td>
</tr>
</tbody>
</table>

The majority of respondents explained that it was the feedback that made them focus on the moves, clarifying that the color-coded feedback made them notice miscommunicated functional meaning and motivated them to revise the way they had expressed that meaning.

"I think it helped me focus on the moves by highlighting the different moves in colors." (Student 51, survey, question 1)

"The feedback is all about the moves by colors and % so that it makes me concentrate on moves which I was not concerned a lot." (Student 7, survey, question 1)

"For example, centrality in move 1. I know the sentence is move 1, but color say it's move 2. So I think and I make sure which step the sentence is, then I modify so it sounds right." (Student 44, survey, question 2)

In support of these findings, introspective and observational data from 16 participants indicated that they referred to the form of their discourse in one way or another. Of the LLP idea units identified in the think-aloud and interview transcripts, roughly half were coded for focus on discourse form (see Table 9).

Table 9
Evidence of Focus on Discourse Form in Transcript Data (N = 83)

<table>
<thead>
<tr>
<th>Idea units</th>
<th>Think-aloud protocols/ Camtasia</th>
<th>Observations</th>
<th>Interviews</th>
</tr>
</thead>
<tbody>
<tr>
<td>Focus on discourse form</td>
<td>53.72%</td>
<td>27.16%</td>
<td>43.13%</td>
</tr>
<tr>
<td>Noticing of negative evidence</td>
<td>24.17%</td>
<td>42.10%</td>
<td>31.07%</td>
</tr>
<tr>
<td>Output modification</td>
<td>14.05%</td>
<td>27.19%</td>
<td>15.80%</td>
</tr>
<tr>
<td>Enhanced understanding</td>
<td>8.06%</td>
<td>10.53%</td>
<td>10.00%</td>
</tr>
</tbody>
</table>

The participants also noticed negative evidence such as inappropriate distribution of their moves as well as miscommunication of the intended functional meaning. This prompted them to experiment with different changes (detailed in section on improvement in rhetorical development below), which helped them enhance their understanding of moves and their functions.

Focus on discourse form comments: "m2 is scattered inside m1." (Student 54, observations transcript)

Noticing of negative evidence: "... stating the value of present research. I think I missed that part. And this sentence is duplicating the previous one." (Student 43, think-aloud protocol and Camtasia transcript)

Output modification: "changes the problematic red sentence to begin with "the importance of... des-... a part of the sentence. Modifies more: "In fact, the importance of cement fineness has been elevated to the heat of hydration." (Student 54, think-aloud protocol and Camtasia transcript)

Enhanced understanding: "I realized that... in all the examples, in the m3, there's no previous research." (Student 27, interviews transcript)

These data also confirmed that the color-coded and numerical feedback generated by IADE was indeed the driving factor that made the students focus on the form of their discourse as noted by many participants in their open-ended survey answers.

"I see there is... the colors are... there are only two moves in my introduction." (Student 58, think-aloud protocol and Camtasia transcript)

"First of all, let me check the statistics with my introduction. [cursor over numerical feedback] ok." (Student 63, think-aloud protocol and Camtasia transcript)

"Looks carefully at colors, then at numbers (highlights what he's looking at)" (Student 28, observations transcript)
Positive evidence of focus on discourse form in the yes/no survey data amounted to 87%. In the Likert-scale responses, as detailed in Table 7, excellent evidence of such focus averaged 44.3%, good 47.7%, weak 6.8%, and poor 1.1%. The participants also positively self-evaluated the degree to which they noticed negative evidence in their own writing, noticed vocabulary indicative of a particular move, and engaged in interactional modification (Ellis, 1999) by changing their written output to address the negative evidence, which constitutes additional positive evidence of noticing and focus on form.

Table 7
Evidence of Noticing and Focus on Discourse Form in Likert-Scale Data (N = 88)

<table>
<thead>
<tr>
<th></th>
<th>Not at all (poor)</th>
<th>A little (weak)</th>
<th>Much (good)</th>
<th>Very much (excellent)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Focused on discourse form</td>
<td>1.10%</td>
<td>6.80%</td>
<td>47.70%</td>
<td>44.30%</td>
</tr>
<tr>
<td>Noticed negative evidence</td>
<td>1.14%</td>
<td>7.95%</td>
<td>50.00%</td>
<td>40.91%</td>
</tr>
<tr>
<td>Noticed vocabulary indicative of moves</td>
<td>0.00%</td>
<td>9.09%</td>
<td>52.27%</td>
<td>38.64%</td>
</tr>
<tr>
<td>Engaged in modified interaction</td>
<td>5.68%</td>
<td>44.32%</td>
<td>26.14%</td>
<td>23.86%</td>
</tr>
<tr>
<td>Average</td>
<td>1.98%</td>
<td>17.05%</td>
<td>44.03%</td>
<td>36.94%</td>
</tr>
</tbody>
</table>

Table 8 lists a summary of the themes that emerged in participants’ open-ended survey responses.

Table 8
Evidence of Focus on Discourse Form in the Open-ended Survey Data (N = 83)

<table>
<thead>
<tr>
<th></th>
<th>Evidence</th>
<th>No evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Focused on discourse form</td>
<td>100.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Feedback triggered focus on discourse form</td>
<td>95.19%</td>
<td>4.81%</td>
</tr>
<tr>
<td>- color-coded feedback</td>
<td>59.04%</td>
<td></td>
</tr>
<tr>
<td>- numerical feedback</td>
<td>12.05%</td>
<td></td>
</tr>
<tr>
<td>- color-coded and numerical feedback</td>
<td>24.10%</td>
<td></td>
</tr>
<tr>
<td>Noticed peculiarities of discourse form</td>
<td>89.16%</td>
<td>10.84%</td>
</tr>
<tr>
<td>Engaged in modified interaction</td>
<td>78.30%</td>
<td>21.69%</td>
</tr>
<tr>
<td>Average</td>
<td>90.70%</td>
<td>9.30%</td>
</tr>
</tbody>
</table>

The majority of respondents explained that it was the feedback that made them focus on the moves, clarifying that the color-coded feedback made them notice miscommunicated functional meaning and motivated them to revise the way they had expressed that meaning.

I think it helped me focus on the moves by highlighting the different moves in colors. (Student 51, survey, question 1)

The feedback is all about the moves by colors and % so that it makes me concentrate on moves which I was not concerned a lot. (Student 7, survey, question 1)

For example, centrality in move 1. I know the sentence is move 1, but color say it’s move 2. So I think and I make sure which step the sentence is, then I modify so it sounds right. (Student 44, survey, question 2)

In support of these findings, introspective and observational data from 16 participants indicated that they referred to the form of their discourse in one way or another. Of the LLP idea units identified in the think-aloud and interview transcripts, roughly half were coded for focus on discourse form (see Table 9).

Table 9
Evidence of Focus on Discourse Form in Transcript Data (N = 83)

<table>
<thead>
<tr>
<th></th>
<th>Think-aloud protocols/ Camtasia</th>
<th>Observations</th>
<th>Interviews</th>
</tr>
</thead>
<tbody>
<tr>
<td>Focus on discourse form</td>
<td>53.72%</td>
<td>27.16%</td>
<td>43.13%</td>
</tr>
<tr>
<td>Noticing of negative evidence</td>
<td>24.17%</td>
<td>42.10%</td>
<td>31.07%</td>
</tr>
<tr>
<td>Output modification</td>
<td>14.05%</td>
<td>27.19%</td>
<td>15.80%</td>
</tr>
<tr>
<td>Enhanced understanding</td>
<td>8.06%</td>
<td>10.53%</td>
<td>10.00%</td>
</tr>
</tbody>
</table>

The participants also noticed negative evidence such as inappropriate distribution of their moves as well as miscommunication of the intended functional meaning. This prompted them to experiment with different changes (detailed in section on improvement in rhetorical development below), which helped them enhance their understanding of moves and their functions.

Focus on discourse form comments that m2 is scattered inside m1. (Student 54, observations transcript)

Noticing of negative evidence... stating the value of present research. I think I missed that part. And this sentence is duplicating the previous one. (Student 43, think-aloud protocol and Camtasia transcript)

Output modification changes the problematic red sentence to begin with “the importance of”. deletes a part of the sentence. Modifies more: “In fact, the importance of cement fineness has been elevated to the heat of hydration.” (Student 54, think-aloud protocol and Camtasia transcript)

Enhanced understanding
I realized that... in all the examples, in the m3, there's no previous research. (Student 27, interviews transcript)

These data also confirmed that the color-coded and numerical feedback generated by IADE was indeed the driving factor that made the students focus on the form of their discourse as noted by many participants in their open-ended survey answers.

I see there is... the colors are... there are only two moves in my introduction. (Student 58, think-aloud protocol and Camtasia transcript)

First of all, let me check the statistics with my introduction. [cursor over numerical feedback] ok. (Student 63, think-aloud protocol and Camtasia transcript)

Looks carefully at colors, then at numbers (highlights what he's looking at) (Student 28, observations transcript)
Looks carefully at the color-coded text returned by IADE. Frowns. (Student 65, observations transcript)

Researcher (following up on the observation note [keeps highlighting sentences colored differently]): Why were you highlighting different colors?

Student 43: Because the colors helped me see how I structure my moves. (Student 64, interview transcript)

A close analysis of the students’ reported thoughts and actions based on analytic induction suggests that learning occurred through intrapersonal interaction that took the form of a cycle stimulated by IADE’s feedback. Focus on discourse form was the head of the cycle; the cycle began with learners’ focus on a certain discourse element of their text and ended with successful output modification of that element. Inside this cycle, there appeared to be another integrated mini-cycle, during which the learners noticed negative evidence in their work products and made multiple attempts to understand its nature and to make corrective changes (see Figure 2).

Figure 2
Revision Cycle Stimulated by IADE’s Feedback

It seems that the automated feedback triggered interactional adjustments on the part of the learners. When they noticed negative evidence in their production, they became more cognitively engaged and made new output modifications, resubmitting and negotiating for meaning with IADE. In Long’s (1996) words, the feedback appeared to “connect input, internal learner capacities, particularly selective attention, and output in productive ways” (Long, 1996, pp. 451-452).

For instance, as exemplified in an excerpt from a think-aloud/Camtasia transcript of participant 27 (see Appendix F), the feedback, both color coded and numerical, first prompted the learner’s focus to the distribution of the three moves in his introduction (lines 2-5). Then, he noticed the negative evidence clearly pointed to by the numerical feedback (lines 8-11), and then, while reading the color-coded text, realized that a move in his introduction was identified by IADE differently than he had intended (lines 11-13). Having noticed this negative
Looks carefully at the color-coded text returned by IADE. Frowns. (Student 65, observations transcript)

Researcher (following up on the observation note [keeps highlighting sentences colored differently]): Why were you highlighting different colors?

Student 43: Because the colors helped me see how I structure my moves. (Student 64, interview transcript)

A close analysis of the students’ reported thoughts and actions based on analytic induction suggests that learning occurred through intrapersonal interaction that took the form of a cycle stimulated by IADE’s feedback. Focus on discourse form was the head of the cycle; the cycle began with learners’ focus on a certain discourse element of their text and ended with successful output modification of that element. Inside this cycle, there appeared to be another integrated mini-cycle, during which the learners noticed negative evidence in their work products and made multiple attempts to understand its nature and to make corrective changes (see Figure 2).

Figure 2
Revision Cycle Stimulated by IADE’s Feedback

It seems that the automated feedback triggered interactional adjustments on the part of the learners. When they noticed negative evidence in their production, they became more cognitively engaged and made new output modifications, resubmitting and negotiating for meaning with IADE. In Long’s (1996) words, the feedback appeared to “connect input, internal learner capacities, particularly selective attention, and output in productive ways” (Long, 1996, pp. 451-452).

For instance, as exemplified in an excerpt from a think-aloud/Camtasia transcript of participant 27 (see Appendix F), the feedback, both color coded and numerical, first prompted the learner’s focus to the distribution of the three moves in his introduction (lines 2-5). Then, he noticed the negative evidence clearly pointed to by the numerical feedback (lines 8-11), and then, while reading the color-coded text, realized that a move in his introduction was identified by IADE differently than he had intended (lines 11-13). Having noticed this negative evidence, the learner made a change in his text (lines 48-19) based on a personal hypothesis (lines 14-15), and, upon resubmission, saw that his hypothesis was faulty (lines 22-23). This, thus, motivated him to think more (lines 26-30), which lead to an enhanced understanding of the discourse norms in his particular discipline (lines 30-32). With a better understanding of that, the student modified his output again (lines 37-43)—this time successfully, which he was able to see by focusing on form again (line 47). His revision process continued with another iteration of the cycle.

Evidence of this learning cycle was also found in the interview data. What was particularly interesting is that participants’ answers to questions about their actions during one stage of the cycle provided evidence about the effect of the preceding stage, which allowed for the inference that the relationship among the four elements of the identified cycle was sequential and causative (see example in Appendix G). That is to say, learner reports suggested that they often noticed negative evidence because of a focus on discourse form. They also reported that they acquired a better understanding of discourse conventions because they noticed negative evidence in their writing. Further, they indicated that they modified their output as a result of having noticed negative evidence and having acquired a better understanding of the rhetorical functions in the discourse. Finally, because they wanted to verify the quality of their modified output, they focused on the discourse form again.

Learning Gains
Focus on discourse form and going through the learning cycle contributed to learning gains, which is another aspect of LLP analyzed in this study. This claim is supported by evidence obtained from a Likert-scale question, five survey questions, and participants’ scores on pre- and posttests (see Table 10).

Table 10
Overall Evidence of Learning Gains

<table>
<thead>
<tr>
<th>Data source</th>
<th># participants</th>
<th>Question Evidence for LLP</th>
<th>No evidence for LLP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likert-scale questions</td>
<td>88</td>
<td>Responses to Q-n 4</td>
<td>98.86%</td>
</tr>
<tr>
<td>Yes/no and open-ended questions</td>
<td>83</td>
<td>Q-n 3</td>
<td>93.97%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Q-n 4</td>
<td>92.77%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Q-n 5</td>
<td>93.98%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Q-n 6</td>
<td>77.11%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Q-n 7</td>
<td>71.08%</td>
</tr>
<tr>
<td>Pre-/posttests</td>
<td>104</td>
<td>Scores for &lt; .001</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>104 pre-/posttests, task 1</td>
<td>104 pre-/posttests, task 2</td>
</tr>
</tbody>
</table>
The majority of the participants who completed the Likert-scale questionnaire believed that they learned the moves well (79.54%), very well (6.82%), a little (12.50%), not at all (1.14%). The themes that emerged in participants’ open-ended responses were causes of learning (i.e., why they thought they learned the moves) and perceived or observed outcomes (i.e., outcomes they could make a judgment about or that they could actually see). The participants thought that they learned the moves for one of two reasons: interaction with IADE through its feedback (21.56%) and their focus on discourse form (7.14%). The perceived outcome they mentioned was their enhanced understanding of the moves and steps (38.87%), and the observed outcome they named was improved quality of their final drafts (11.76%). The focus on form, enhanced understanding, and better quality of end products themes corroborate the evidence of the learning cycle presented above.

Considering the theoretically supported importance of focus on form and its hypothesized role in this study, a survey question asked the participants whether they thought they learned the moves as a result of having focused on the discourse form and why they thought so. The majority of respondents (93.98%) said yes, 2.4% were not sure, and 3.61% said no. The themes that emerged in the open-ended answers present justifiable interest because they were supportive of some stages of the learning cycle. In other words, it seems that the students realized that they noticed negative evidence, modified their output, and enhanced their understanding and knowledge of the moves because they had focused on the form of their introduction discourse.

Yes. I focused on the moves to revise my draft. If I lack one move, I add some sentences to represent this move. If I have too much of one move, I combine some sentences or delete some sentences to cut this move shorter. From this procedure, I learned what kind of sentences cannot express the moves correctly and how to revise them into explicit moves. (Student 50, survey question 4)

Yes, because focus on my moves helped me not only to recognize the moves but to understand them. (Student 43, survey question 4)

Additionally, most of the participants (77.11%) thought that they could transfer what they learned to their actual writing; 16.87% were not so confident, saying “I don’t know,” “I am not sure,” “I will try,” “possibly,” “maybe,” “I hope so,” or “kind of.” 6.02% did not think they were ready to apply the newly acquired knowledge to produce well rounded introductions, explaining that they may need more practice. Those who were more optimistic about their learning gains mostly justified their optimism by naming the actions that they would take to ensure the transfer of knowledge, among which were: paying attention to and analyzing their moves (64.06%), matching whether the functional meaning of their moves was successfully expressed through vocabulary (21.88%), and comparing their work to that of published professional texts (14.06%).

Evidence of participants’ learning gains was also obtained through t tests. Table 11 shows the results for task one, task two, and overall test scores. All the posttest means were significantly higher than the pretest means (p < .001) and reflect the substantial learning gains on the part of the participants.

### Table 11

<table>
<thead>
<tr>
<th>Mean Pre-/Posttest Scores for Task One, Task Two, and Overall (N = 104)</th>
<th>M</th>
<th>SD</th>
<th>T</th>
<th>DF</th>
<th>Effect size10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pretest task 1</td>
<td>20.50</td>
<td>4.942</td>
<td>18.26*</td>
<td>103</td>
<td>.77</td>
</tr>
<tr>
<td>Posttest task 1</td>
<td>29.65</td>
<td>1.909</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pretest task 2</td>
<td>21.05</td>
<td>6.750</td>
<td>13.75*</td>
<td>103</td>
<td>.68</td>
</tr>
<tr>
<td>Posttest task 2</td>
<td>30.16</td>
<td>1.596</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pretest overall</td>
<td>41.55</td>
<td>10.161</td>
<td>17.99*</td>
<td>103</td>
<td>.77</td>
</tr>
<tr>
<td>Posttest overall</td>
<td>59.82</td>
<td>2.988</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*p < .001

### Improvement in Rhetorical Development

Focus on discourse form triggered by automated feedback led to improvement in the rhetorical quality of student writing. Although the lack of a control group somewhat weakens this claim, evidence supporting the claim was found in multiple data, including successful output modifications from first to final drafts. The data sources are listed in Table 12.

### Table 12

Overall Evidence of Improvement

<table>
<thead>
<tr>
<th>Data source</th>
<th># participants</th>
<th>Data</th>
<th>Evidence for LLP</th>
<th>No evidence for LLP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likert-scale questions</td>
<td>88</td>
<td>Responses to Q-n 5</td>
<td>100.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Yes/no and open-ended questions</td>
<td>83</td>
<td>Q-n 8</td>
<td>92.80%</td>
<td>7.20%</td>
</tr>
<tr>
<td>Scores</td>
<td>IADE's automated evaluation of first and final drafts</td>
<td>105</td>
<td>Drafts</td>
<td>210 first and final drafts</td>
</tr>
<tr>
<td>Human scores of first and final drafts</td>
<td>16</td>
<td>32 first and final drafts</td>
<td>Move 1: p &lt; .001</td>
<td>Move 2: p &lt; .001</td>
</tr>
<tr>
<td>First and final drafts</td>
<td>16</td>
<td>32 first and final drafts</td>
<td>285 output modifications</td>
<td></td>
</tr>
<tr>
<td>Think-aloud protocols/Camtasia</td>
<td>16</td>
<td>Transcripts of 16 Think-aloud/Camtasia recordings</td>
<td>77 output modifications</td>
<td></td>
</tr>
</tbody>
</table>

According to the Likert-scale responses, all the participants believed that they improved their
The majority of the participants who completed the Likert-scale questionnaire believed that they learned the moves well (79.54%), very well (6.82%), a little (12.50%), not at all (1.14%). The themes that emerged in participants' open-ended responses were causes of learning (i.e., why they thought they learned the moves) and perceived or observed outcomes (i.e., outcomes they could make a judgment about or that they could actually see). The participants thought that they learned the moves for one of two reasons: interaction with IADE through its feedback (21.56%) and their focus on discourse form (7.14%). The perceived outcome they mentioned was their enhanced understanding of the moves and steps (38.87%), and the observed outcome they named was improved quality of their final drafts (11.76%).

The focus on form, enhanced understanding, and better quality of end products themes corroborate the evidence of the learning cycle presented above.

Considering the theoretically supported importance of focus on form and its hypothesized role in this study, a survey question asked the participants whether they thought they learned the moves as a result of having focused on the discourse form and why they thought so. The majority of respondents (93.98%) said yes, 2.4% were not sure, and 3.61% said no. The themes that emerged in the open-ended answers present justifiable interest because they were supportive of some stages of the learning cycle. In other words, it seems that the students realized that they noticed negative evidence, modified their output, and enhanced their understanding and knowledge of the moves because they had focused on the form of their introduction discourse.

Yes, I focused on the moves to revise my draft. If I lack one move, I add some sentences or delete some sentences to cut this move shorter. From this procedure, I learned what kind of sentences cannot express the moves correctly and how to revise them into explicit moves. (Student 50, survey question 4)

Yes, because focus on my moves helped me not only to recognize the moves but to understand them. (Student 43, survey question 4)

Additionally, most of the participants (77.11%) thought that they could transfer what they learned to their actual writing; 16.87% were not so confident, saying "I don't know," "I am not sure," "I will try," "possibly," "maybe," "I hope so," or "kind of;" 6.02% did not think they were ready to apply the newly acquired knowledge to produce well rounded introductions, explaining that they may need more practice. Those who were more optimistic about their learning gains mostly justified their optimism by naming the actions that they would take to ensure the transfer of knowledge, among which were: paying attention to and analyzing their moves (64.06%), matching whether the functional meaning of their moves was successfully expressed through vocabulary (21.88%), and comparing their work to that of published professional texts (14.06%).

Evidence of participants’ learning gains was also obtained through t tests. Table 11 shows the results for task one, task two, and overall test scores. All the posttest means were significantly higher than the pretest means (p < .001) and reflect the substantial leaning gains on all outcomes.

### Table 11

<table>
<thead>
<tr>
<th>Data source</th>
<th># participants</th>
<th>Evidence for LLP</th>
<th>No evidence for LLP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likert-scale questions</td>
<td>88</td>
<td>100.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Yes/no and open-ended questions</td>
<td>83</td>
<td>92.80%</td>
<td>7.20%</td>
</tr>
<tr>
<td>Scores IADE's automated evaluation of first and final drafts</td>
<td>105</td>
<td>t test</td>
<td>Move 1: p &lt; .001 Move 2: p &lt; .001 Move 3: p &lt; .001 Length: p &lt; .001</td>
</tr>
<tr>
<td>Human scores of first and final drafts</td>
<td>16</td>
<td>Move 1: p &lt; .001 Move 2: p &lt; .001 Move 3: p &lt; .001</td>
<td></td>
</tr>
<tr>
<td>First and final drafts</td>
<td>16</td>
<td>285 output modifications</td>
<td></td>
</tr>
<tr>
<td>Think-aloud protocols/ Camtasia</td>
<td>16</td>
<td>77 output modifications</td>
<td></td>
</tr>
</tbody>
</table>

* "I will try," "possibly," "maybe," "I hope so," or "kind of;"

The focus on discourse form triggered by automated feedback led to improvement in the rhetorical quality of student writing. Although the lack of a control group somewhat weakens this claim, evidence supporting the claim was found in multiple data, including successful output modifications from first to final drafts. The data sources are listed in Table 12.
skill in writing a research article introduction to some degree: 26% thought they improved a lot, 55% thought they improved somewhat, and 19% thought they improved a little. None of the participants perceived no improvement at all. Similarly, 92.8% of 83 respondents answered yes when asked if they thought they improved their writing skills, and 7.2% were uncertain. They appeared to judge improvement based on what they thought caused the improvement and based on perceived or observed outcomes. Among the causes of improvement, 48.05% mentioned the ability of the feedback to direct their attention to the discourse form of their draft and the opportunity for practice through multiple resubmissions. The outcomes that 41.56% believed were indicative of improvement in their skills were enhanced understanding and knowledge of the rhetorical conventions of introductions and better quality of their final drafts. These insights resonate with the themes that emerged in other survey data.

The number and percentage of the scores (1, 2, or 3) assigned to each move and to the length of students’ first and final drafts based on IADE’s automated evaluation (see section on automated evaluation above) are listed in Table 13.

Figure 3 shows the degree of improvement from first to final draft expressed as percentages for each move and length. The “noticeable improvement” and “no need for improvement” categories were the most prominent ones. On average, 39.5% of students improved their drafts noticeably and 18.5% considerably. Most noticeable and considerable improvements were made at the level of Move 3 (52.4% and 28.6%). Move 3 may have been easier for students to develop since, to accomplish the functions of this move, they had to describe their own work, which is something they knew very well. Length was the aspect that was least improved; however, 52.2% appeared to need no improvement. The percentages of students whose first draft discourse elements saw no improvement were the lowest (0.9%, 3.8%, 7.6%, and 14.3%). The “no need for improvement” category had a relatively high percentage, and there may be two possible explanations for this result. It may have been the case that some participants’ drafts were good to start with, or IADE’s automated analysis may have a limitation. Since IADE evaluates the discourse quality based on the unigrams, bigrams, and trigrams that have the highest probability of occurrence, the scores might have been higher when such vocabulary items were present in the text.

The figures in Table 13 show that the distribution of all these elements was better in final drafts than in first drafts; considerably more final drafts were evaluated with the highest score of 3 (move 1 97.1%, move 2 92.4%, move 3 87.6%, and length 83.8%). Overall, of 420 total possible scores, the highest score of 3 was assigned 159 times (37.9%) to first drafts and 379 times (90.2%) to final drafts. Conversely, considerably fewer scores of 1 and 2 were assigned to final drafts compared to first drafts.

### Table 13 Scores for First and Final Drafts Based on IADE’s Automated Evaluation

<table>
<thead>
<tr>
<th>Element</th>
<th>Score of 1</th>
<th>Score of 2</th>
<th>Score of 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Move 1</td>
<td>First draft</td>
<td>Final draft</td>
<td>First draft</td>
</tr>
<tr>
<td>Move 2</td>
<td>First draft</td>
<td>Final draft</td>
<td>First draft</td>
</tr>
<tr>
<td>Move 3</td>
<td>First draft</td>
<td>Final draft</td>
<td>First draft</td>
</tr>
<tr>
<td>Length</td>
<td>First draft</td>
<td>Final draft</td>
<td>First draft</td>
</tr>
</tbody>
</table>

The human scores for 32 drafts revealed that the first drafts were weaker in rhetorical quality and became stronger after revision. As shown in Table 15, a total of 47.9% of moves received a score of 1 in first drafts as compared with 0.0% in final drafts. The number of 2 scores also decreased from 45.8% in first drafts to 25.0% in final drafts. The number of the highest scores of 3 increased from 6.3% to 75.0%.

### Table 15 Scores for 32 First and Final Drafts Based on Human Ratings (N = 105)

<table>
<thead>
<tr>
<th>Element</th>
<th>Score of 1</th>
<th>Score of 2</th>
<th>Score of 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Move 1</td>
<td>First draft</td>
<td>Final draft</td>
<td>First draft</td>
</tr>
<tr>
<td>Move 2</td>
<td>First draft</td>
<td>Final draft</td>
<td>First draft</td>
</tr>
<tr>
<td>Move 3</td>
<td>First draft</td>
<td>Final draft</td>
<td>First draft</td>
</tr>
<tr>
<td>Length</td>
<td>First draft</td>
<td>Final draft</td>
<td>First draft</td>
</tr>
</tbody>
</table>

The evidence for improvement as judged by human raters is supported by t-test results comparing the first and final scores for each move (see Table 16). The means for all three moves significantly increased in final drafts (p < .001), similar to the results obtained when comparing the scores based on IADE’s automated analysis.
skill in writing a research article introduction to some degree: 26% thought they improved a lot, 55% thought they improved somewhat, and 19% thought they improved a little. None of the participants perceived no improvement at all. Similarly, 92.8% of 83 respondents answered yes when asked if they thought they improved their writing skills, and 7.2% were uncertain. They appeared to judge improvement based on what they thought caused the improvement and based on perceived or observed outcomes. Among the causes of improvement, 48.0% mentioned the ability of the feedback to direct their attention to the discourse form of their draft and the opportunity for practice through multiple resubmissions. The outcomes that 41.5% believed were indicative of improvement in their skills were enhanced understanding and knowledge of the rhetorical conventions of introductions and better quality of their final drafts. These insights resonate with the themes that emerged in other survey data.

The number and percentage of the scores (1, 2, or 3) assigned to each move and to the length of students' first and final drafts based on IADE's automated evaluation (see section on automated evaluation above) are listed in Table 13.

Table 13
Scores for First and Final Drafts Based on IADE’s Automated Evaluation

<table>
<thead>
<tr>
<th>Element</th>
<th>Score of 1</th>
<th>Score of 2</th>
<th>Score of 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Move 1</td>
<td>First draft</td>
<td>Final draft</td>
<td>First draft</td>
</tr>
<tr>
<td></td>
<td>16 (15.2%)</td>
<td>0 (0.0%)</td>
<td>43 (41%)</td>
</tr>
<tr>
<td>Move 2</td>
<td>31 (29.5%)</td>
<td>0 (0.0%)</td>
<td>36 (34.3%)</td>
</tr>
<tr>
<td>Move 3</td>
<td>35 (33.3%)</td>
<td>0 (0.0%)</td>
<td>55 (52.4%)</td>
</tr>
<tr>
<td>Length</td>
<td>8 (7.6%)</td>
<td>1 (0.9%)</td>
<td>37 (35.2%)</td>
</tr>
<tr>
<td>Total</td>
<td>90 (21.4%)</td>
<td>1 (0.2%)</td>
<td>171 (40.7%)</td>
</tr>
</tbody>
</table>

The figures in Table 13 show that the distribution of all these elements was better in final drafts than in first drafts; considerably more final drafts were evaluated with the highest score of 3. Most noticeable and considerable improvements were made at the level of Move 3 (52.4% and 28.6%). Move 3 may have been easier for students to develop since, to accomplish the functions of this move, they had to describe their own work, which is something they knew very well. Length was the aspect that was least improved; however, 55.2% appeared to need no improvement. The percentages of students whose first draft discourse elements saw no improvement were the lowest (0.9%, 3.8%, 7.6%, and 14.3%). The “no need for improvement” category had a relatively high percentage, and there may be two possible explanations for this result. It may have been the case that some participants’ drafts were good to start with, or IADE’s automated analysis may have a limitation. Since IADE evaluates the discourse quality based on the unigrams, bigrams, and trigrams that have the highest probability of occurrence, the scores might have been higher when such vocabulary items were present in the text.

Table 14
Mean Scores for Moves 1, 2, 3 and Length in First and Final Drafts Based on IADE’s Evaluation (N = 105)

<table>
<thead>
<tr>
<th>Element</th>
<th>First draft</th>
<th>Final draft</th>
<th>First draft</th>
<th>Final draft</th>
<th>First draft</th>
<th>Final draft</th>
</tr>
</thead>
<tbody>
<tr>
<td>Move 1</td>
<td>2.29</td>
<td>0.717</td>
<td>9.88*</td>
<td>104</td>
<td>.54</td>
<td></td>
</tr>
<tr>
<td>Move 2</td>
<td>2.97</td>
<td>0.167</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Move 3</td>
<td>2.04</td>
<td>0.808</td>
<td>11.34*</td>
<td>104</td>
<td>.59</td>
<td></td>
</tr>
<tr>
<td>Move 3</td>
<td>2.92</td>
<td>0.267</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Move 3</td>
<td>1.81</td>
<td>0.666</td>
<td>14.57*</td>
<td>104</td>
<td>.71</td>
<td></td>
</tr>
<tr>
<td>Move 3</td>
<td>2.88</td>
<td>0.331</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Length</td>
<td>2.46</td>
<td>0.665</td>
<td>6.25</td>
<td>104</td>
<td>.31</td>
<td></td>
</tr>
<tr>
<td>Length</td>
<td>2.83</td>
<td>0.403</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The evidence for improvement as judged by human raters is supported by t-test results comparing the first and final scores for each move (see Table 16). The means for all three moves significantly increased in final drafts (p < .001), similar to the results obtained when comparing the scores based on IADE’s automated analysis.
Improvement in students’ writing performance was also reflected in the modifications they made during revision, which were captured in Camtasia screen recordings and also identified in a direct comparison of their first and final drafts. Camtasia transcripts contained 14% idea units coded as output modification, containing overall 77 changes. Manual analysis of the 32 drafts revealed that 31 sentences remained unchanged (13%), and 200 sentences (87%) were modified and contained 285 changes from first to final draft. Those changes were made at different levels. The participants modified their output mostly in content, vocabulary, and structure and less in grammar and mechanics. This is not surprising given the nature of the feedback, which made the moves in students’ texts salient and triggered their thinking about what functional meaning they were trying to convey and how they could use language to do this more effectively. The frequencies of students’ output modifications identified in Camtasia transcripts and in students’ first and final drafts are presented in Table 17.

**Table 16**
Mean Scores for Moves 1, 2, 3 in 16 First and Final Drafts Based on Human Ratings Evaluation

<table>
<thead>
<tr>
<th>Move</th>
<th>SD</th>
<th>T</th>
<th>DF</th>
<th>Effect size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Move 1 first draft</td>
<td>1.75</td>
<td>0.447</td>
<td>7.41*</td>
<td>15</td>
</tr>
<tr>
<td>Move 1 final draft</td>
<td>2.61</td>
<td>0.403</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Move 2 first draft</td>
<td>1.31</td>
<td>0.479</td>
<td>8.88*</td>
<td>15</td>
</tr>
<tr>
<td>Move 2 final draft</td>
<td>2.69</td>
<td>0.479</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Move 3 first draft</td>
<td>1.69</td>
<td>0.793</td>
<td>5.51*</td>
<td>15</td>
</tr>
<tr>
<td>Move 3 final draft</td>
<td>2.75</td>
<td>0.447</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* p < .001

Table 17
Output Modifications in Camtasia and First to Final Drafts

<table>
<thead>
<tr>
<th>Level of output modification</th>
<th>Camtasia</th>
<th>First-final drafts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content (additions, deletions, modified ideas)</td>
<td>18 (23%)</td>
<td>80 (28%)</td>
</tr>
<tr>
<td>Lexical (move specific, nonmove specific)</td>
<td>38 (49%)</td>
<td>121 (42%)</td>
</tr>
<tr>
<td>Grammar (verb tense/form, SV agreement, plurals)</td>
<td>5 (7%)</td>
<td>14 (5%)</td>
</tr>
<tr>
<td>Structure (sentence, paragraph)</td>
<td>10 (13%)</td>
<td>60 (21%)</td>
</tr>
<tr>
<td>Mechanics (citation format, punctuation)</td>
<td>6 (8%)</td>
<td>10 (4%)</td>
</tr>
<tr>
<td>Total</td>
<td>77 (100%)</td>
<td>285 (100%)</td>
</tr>
</tbody>
</table>

These results add support to the small body of previous AWE research. Similar to the findings of this study, Elliot and Mikulas (2004) reported that student writing skills, as measured by performance on statewide writing assessments, were significantly improved by using MY Access! Their survey results also indicated that students were highly satisfied with the automated feedback on their essays. Foltz, Laham, and Landauer (1999) recorded an improvement in scores ranging from 0 to 33 points over an average of three revisions with WriteToLearn. Attali (2004) demonstrated that automated scores for essays submitted to Criterion more than once increased from first to final submission and that students significantly lowered their error rate. In their study, however, the revisions were made mainly at the level of spelling and grammar and not at the structural level. In Leah Rock’s (2007) study, 9th-grade students used Criterion for 4 weeks and, during this short period of time, received higher analytic scores on their essays written at the end of the study period and improved the mechanical aspects of their writing.

Unlike previous research findings, the evidence in this study not only suggests improvement, but also shows how and to what extent the rhetorical quality of student writing improved. Moreover, the evidence of the learning cycle prompted by the automated feedback is intriguing, especially because there appeared to exist certain sequential and causative relations among the elements of the cycle. Given that IADE’s functionality is very specific and that the current is the first study of this kind, it is premature to generalize its findings. Nevertheless, the results allow for the conclusion that automated feedback has potential to facilitate learning particularly if applied appropriately in targeted contexts.

**LIMITATIONS**
Inasmuch as the results reported here are encouraging, there were limitations in the research design. The fact that the study was conducted in an instructional setting in which students take writing courses in order to complete language proficiency required by the university made it difficult to employ a rigorous methodology that would rely on random sampling and assignment of participants from the population of interest. All the participants were subject to the same type of treatment; an experimental/control group design should be followed in future research. Examining the work products of learners who had and had not used IADE, not only for score comparison but also for the nature of the revision process and of the quality of final drafts, would have allowed for a better understanding of how learning to write academi-
In the case of these 32 drafts, as shown in Figure 4, the highest average percentage belongs to noticeable improvement (54.38%) and the lowest to no need for improvement (6.25%), a finding which is similar to the degrees of improvement that resulted from the automated evaluation.

Figure 4
Degrees of Improvement in Moves and Length Based on Human Ratings of 16 First and Final Drafts

Improvement in students’ writing performance was also reflected in the modifications they made during revision, which were captured in Camtasia screen recordings and also identified in a direct comparison of their first and final drafts. Camtasia transcripts contained 14% idea units coded as output modification, containing overall 77 changes. Manual analysis of the 32 drafts revealed that 31 sentences remained unchanged (13%), and 200 sentences (87%) were modified and contained 285 changes from first to final draft. Those changes were made at different levels. The participants modified their output mostly in content, vocabulary, and structure and less in grammar and mechanics. This is not surprising given the nature of the feedback, which made the moves in students’ texts salient and triggered their thinking about what functional meaning they were trying to convey and how they could use language to do this more effectively. The frequencies of students’ output modifications identified in Camtasia transcripts and in students’ first and final drafts are presented in Table 17.

These results add support to the small body of previous AWE research. Similar to the findings of this study, Elliot and Mikulas (2004) reported that student writing skills, as measured by performance on statewide writing assessments, were significantly improved by using MY Access!*. Their survey results also indicated that students were highly satisfied with the automated feedback on their essays. Foltz, Laham, and Landauer (1999) recorded an improvement in scores ranging from 0 to 33 points over an average of three revisions with WriteToLearn. Attali (2004) demonstrated that automated scores for essays submitted to Criterion more than once increased from first to final submission and that students significantly lowered their error rate. In their study, however, the revisions were made mainly at the level of spelling and grammar and less at the structural level. In Leah Rock’s (2007) study, 9th-grade students used Criterion for 4 weeks and, during this short period of time, received higher analytic scores on their essays written at the end of the study period and improved the mechanical aspects of their writing.

Unlike previous research findings, the evidence in this study not only suggests improvement, but also shows how and to what extent the rhetorical quality of student writing improved. Moreover, the evidence of the learning cycle prompted by the automated feedback is intriguing, especially because there appeared to exist certain sequential and causative relations among the elements of the cycle. Given that IADE’s functionality is very specific and that the current is the first study of this kind, it is premature to generalize its findings. Nevertheless, the results allow for the conclusion that automated feedback has potential to facilitate learning particularly if applied appropriately in targeted contexts.

LIMITATIONS

Inasmuch as the results reported here are encouraging, there were limitations in the research design. The fact that the study was conducted in an instructional setting in which students take writing courses in order to complete language proficiency required by the university made it difficult to employ a rigorous methodology that would rely on random sampling and assignment of participants from the population of interest. All the participants were subject to the same type of treatment; an experimental/control group design should be followed in future research. Examining the work products of learners who had and had not used IADE, not only for score comparison but also for the nature of the revision process and of the quality of final drafts, would have allowed for a better understanding of how learning to write academi-
The findings of this study have a number of implications. The interactionist perspective re-
has a relatively strong language learning potential. In other words, the results suggest that the program’s color-coded and numerical feedback
regulated and provided evidence that IADE’s automated feedback has the potential to trigger
the miscommunicated discourse function.

CONCLUSION
The purpose of this study was to investigate the effects of automated feedback applying the
phenomena and eventually produce better quality output. It was also interesting to see how
theoretical interactionist models can be more confidently extrapolated to computer-
served in this study are generalizable across contexts or are strictly context dependent. Nev-
learners with IADE’s feedback can be found in Cotos (2009, 2010).

not entirely surprising because of the reliance of IADE’s automated evaluation on the n-gram,
the annotated corpus representative of all disciplines covered by IADE. The main limitation of this lexical
misperceptions and eventually produce better quality output. It was also interesting to see how
The rest of the idea units were coded for other CALL qualities investigated in the larger study.
Very few of the responses (0.67%) expressed uncertainty and therefore were not informative.
All references to effect size statistics in the tables in this article reflect a moderate to large effect based on Cohen’s d values.

NOTES
1 The 1,000 text corpus contains smaller subcorpora for 50 academic disciplines (20 research articles per discipline) that were annotated and used to train IADE’s automated evaluation engine.
2 The lexical features that would allow for proper classification were identified by extracting word uni-
grams, bigrams, and trigrams (i.e., single words, two-word sequences, and three-word sequences) from
the annotated corpus representative of all disciplines covered by IADE. The main limitation of this lexical
approach is that it does not account for logical contextual meaning.
3 It must be mentioned that IADE-human reliability, or agreement between IADE and two human raters,
was estimated through calculations of the Cohen’s kappa coefficient (κ). The agreement was moder-
ate on evaluations of individual moves (.582 for Move 1, .529 for Move 2, and .607 for Move 3) and is
considered satisfactory for these data. The disagreement occurred mostly on first drafts; however, this
may not be entirely surprising because of the reliance of IADE’s automated evaluation on the n-gram,
lexical approach.
4 A detailed description of theoretical, practical, and technological perspectives that informed the deci-
sions regarding IADE’s feedback can be found in Cotos (2009, 2010).
5 These ID numbers were used during the analysis phase and will accompany the quotes provided to
exemplify the findings presented in the results section.
6 After revising their documents with IADE, the students submitted their drafts to their instructor and
received additional feedback before submitting their work for a grade.
7 The participants resubmitted their revisions to IADE for automated analysis and feedback multiple
times, but only the drafts that were submitted first and last (final drafts) were analyzed.
8 The rest of the idea units were coded for other CALL qualities investigated in the larger study.
9 Very few of the responses (0.67%) expressed uncertainty and therefore were not informative.
10 All references to effect size statistics in the tables in this article reflect a moderate to large effect based on Cohen’s d values.

REFERENCES
at the National Council on Measurement in Education Annual Meeting, San Diego, CA. Retrieved
ing, and Assessment, 4(3), 1-30.
& Heinle.
89, 297-298.
John Benjamins.

Pedagogical uses of AWE have been debated in the literature largely due to judgmental specu-
lations that lack empirical support. While this study does not prove the effectiveness of AWE applications in general, it provides a viable methodological model for future AWE evaluative research because it allows for the generation of complex and comprehensive results. It can be very appropriate for future studies on AWE feedback as well as for validations of AWE systems because it allows for capturing the complexities of both the processes and the products resulting from AWE use. Evidence yielded from such research can be used to build an empirical evaluation argument for the use of this technology.
The findings of this study have a number of implications. The interactionist perspective re-
cursively enhances a relatively strong language learning potential. In other words, the results suggest that the program’s color-coded and numerical feedback 
generated and provided evidence that IADE’s automated feedback has the potential to trigger 
CONCLUSION 
recordings would have offered more in-depth perceptions and more genuine reactions.

The purpose of this study was to investigate the effects of automated feedback applying the 
tenets of interactionist SLA. Multiple sources of quantitative and qualitative data were triang-
gulated and provided evidence that IADE’s automated feedback has the potential to trigger noticing and focus on discourse form, to enhance learning, and to contribute to improvement.

In other words, the results suggest that the program’s color-coded and numerical feedback has a relatively strong language learning potential.

The findings of this study have a number of implications. The interactionist perspective re-
ceived indirect support for the role of its constructs as applied to computer-assisted language 
Learning inferred from the qualitative insights about the learning cycle. Learners’ output, which was automatically processed and returned to them in the form of modified input, be-
came a productive source for their linguistic hypotheses. The automated feedback stimulated computer-learner interaction by confirming or disconfirming their hypotheses and prompted revised output. Focus on form and noticing of negative evidence lead to intra-personal in-
teraction, during which learners appeared to construct a better understanding of discourse phenomena and eventually produce better quality output. It was also interesting to see how potent input enhancement was. The color codes made the negative evidence more salient and therefore stimulated learners to test hypotheses they generated with regards to the nature of the miscommunicated discourse function.

Replication studies are needed in other contexts to investigate whether the phenomena ob-
served in this study are generalizable across contexts or are strictly context dependent. Nev-
ertheless, theoretical interactionist models can be more confidently extrapolated to computer-
ized language learning environments in general and to L2 academic writing in particular.

On the one hand, future studies investigating learners’ interaction with AWE applications can add support to SLA tenets and strengthen them with new empirical evidence. On the other hand, developers of these new learning technologies can rely on interactionist models in de-
signing and evaluating their products, which, while highly complex, are not always theoreti-
cally informed.

Pedagogical uses of AWE have been debated in the literature largely due to judgmental specu-
lations that lack empirical support. While this study does not prove the effectiveness of AWE applications in general, it provides a viable methodological model for future AWE evaluative research because it allows for the generation of complex and comprehensive results. It can be very appropriate for future studies on AWE feedback as well as for validations of AWE systems because it allows for capturing the complexities of both the processes and the products re-
sulting from AWE use. Evidence yielded from such research can be used to build an empirical evaluation argument for the use of this technology.

NOTES

1 The 1,000 text corpus contains smaller subcorpora for 50 academic disciplines (20 research articles per discipline) that were annotated and used to train IADE’s automated evaluation engine.

2 The lexical features that would allow for proper classification were identified by extracting word uni-
grams, bigrams, and trigrams (i.e., single words, two-word sequences, and three-word sequences) from the annotated corpus representative of all disciplines covered by IADE. The main limitation of this lexical approach to automated evaluation is that it does not account for logical contextual meaning.

3 It must be mentioned that IADE-human reliability, or agreement between IADE and two human raters, was estimated through calculations of the Cohen’s kappa coefficient (κ). The agreement was moder-
ate on evaluations of individual moves (.582 for Move 1, .529 for Move 2, and .607 for Move 3) and is considered satisfactory for these data. The disagreement occurred mostly on first drafts; however, this may not be entirely surprising because of the reliance of IADE’s automated evaluation on the n-gram, lexical approach.

4 A detailed description of theoretical, practical, and technological perspectives that informed the deci-
sions regarding IADE’s feedback can be found in Cotos (2009, 2010).

5 These ID numbers were used during the analysis phase and will accompany the quotes provided to exemplify the findings presented in the results section.

6 After revising their documents with IADE, the students submitted their drafts to their instructor and received additional feedback before submitting their work for a grade.

7 The participants resubmitted their revisions to IADE for automated analysis and feedback multiple times, but only the drafts that were submitted first and last (final drafts) were analyzed.

8 The rest of the idea units were coded for other CALL qualities investigated in the larger study.

9 Very few of the responses (0.67%) expressed uncertainty and therefore were not informative.

10 All references to effect size statistics in the tables in this article reflect a moderate to large effect based on Cohen’s d values.

REFERENCES


Potential of Automated Writing Evaluation Feedback

CALICO Journal, 28(2) Elena Cotos


---

**APPENDIX A**

Overview of ICALL systems and feedback

<table>
<thead>
<tr>
<th>ICALL system</th>
<th>System components</th>
<th>Language</th>
<th>Skills</th>
<th>Type of feedback</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labrie &amp; Singh (1991) - Miniprof</td>
<td>parser, error diagnosis, tutor</td>
<td>French</td>
<td>grammar</td>
<td>dialog, error-specific, explicit pointer to error, metalinguistic</td>
</tr>
<tr>
<td>Liou (1991)</td>
<td>pattern matching, parser, message generator</td>
<td>English</td>
<td>grammar</td>
<td>error-specific, explicit pointer to error, metalinguistic examples</td>
</tr>
<tr>
<td>Holland et al. (1993) - BRIDGE</td>
<td>parser, graphical aids, morphological parser, syntactic parser, core</td>
<td>German</td>
<td>grammar (spatial description)</td>
<td>right/wrong, error-specific, explicit pointer to error, metalinguistic examples</td>
</tr>
<tr>
<td>Nagata (1993) - NIHONGO-CALI</td>
<td>lexicon, morphological rules, grammar rules</td>
<td>Japanese</td>
<td>Grammar (passive structures); vocabulary</td>
<td>error-specific, explicit pointer to error, metalinguistic</td>
</tr>
<tr>
<td>Chen &amp; Tokuda (2003), Tokuda &amp; Chen (2001, 2004) - Azalea</td>
<td>template automaton structure, diagnostic engine, POS parser, parser-based learner model, visual interface</td>
<td>English</td>
<td>grammar, vocabulary</td>
<td>error-specific, explicit pointer to error, metalinguistic, correction</td>
</tr>
<tr>
<td>Reuer (2003)</td>
<td>LFG parser, unrestricted grammar and lexicon, linguistic knowledge</td>
<td>English</td>
<td>grammar</td>
<td>error-specific, explicit pointer to error, metalinguistic, tree-form</td>
</tr>
<tr>
<td>Shaalan (2005) - Arabic ICALL</td>
<td>user interface, course material, morphological analyzer, syntax parser, non/grammatical rules, lexicon, feedback module</td>
<td>Arabic</td>
<td>grammar</td>
<td>error-specific, explicit pointer to error, metalinguistic</td>
</tr>
<tr>
<td>Yang and Akahori (1998)</td>
<td>WWW interface, morpheme analyzer, syntax analyzer, analyzer; grammar rules, feedback system, knowledge database, students learning histories, dictionary</td>
<td>Japanese</td>
<td>grammar</td>
<td>error-specific, explicit pointer to error, metalinguistic, examples, source links, adaptive individualized</td>
</tr>
<tr>
<td>Heift (2002, 2003)</td>
<td>domain knowledge, analysis module, filtering module, Student Model</td>
<td>German</td>
<td>grammar</td>
<td>error-specific, level-tailored, explicit pointer to error, metalinguistic, error-prioritized, remediation exercises, adaptive individualized, error summary</td>
</tr>
</tbody>
</table>
APPENDIX A

Overview of ICALL systems and feedback

<table>
<thead>
<tr>
<th>ICALL system</th>
<th>System components</th>
<th>Language</th>
<th>Skills</th>
<th>Type of feedback</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labrie &amp; Singh (1991) - Miniprof</td>
<td>parser, error diagnosis, tutor</td>
<td>French</td>
<td>grammar</td>
<td>dialog, error-specific, explicit pointer to error, metalinguistic</td>
</tr>
<tr>
<td>Liou (1991)</td>
<td>pattern matching, parser, message generator</td>
<td>English</td>
<td>grammar</td>
<td>error-specific, explicit pointer to error, metalinguistic examples</td>
</tr>
<tr>
<td>Holland et al. (1993) - BRIDGE</td>
<td>parser, graphical aids morphological parser, syntactic parser, core lexicon, morphological rules, grammar rules</td>
<td>German</td>
<td>grammar (spatial description)</td>
<td>right/wrong, error-specific, explicit pointer to error, metalinguistic examples</td>
</tr>
<tr>
<td>Nagata (1993) - NIHONGO-CALI</td>
<td>template automaton structure, diagnostic engine, POS parser, parser-based learner model, visual interface</td>
<td>Japanese</td>
<td>Grammar (passive structures); vocabulary</td>
<td>error-specific, explicit pointer to error, metalinguistic</td>
</tr>
<tr>
<td>Chen &amp; Tokuda (2003), Tokuda &amp; Chen (2001, 2004) - Azatea</td>
<td>LFG parser, unrestricted grammar and lexicon, linguistic knowledge</td>
<td>English</td>
<td>grammar, vocabulary</td>
<td>error-specific, explicit pointer to error, metalinguistic, correction</td>
</tr>
<tr>
<td>Reuer (2003)</td>
<td>user interface, course material, morphological analyzer, syntax parser, non/grammatical rules, lexicon, feedback module</td>
<td>Arabic</td>
<td>grammar</td>
<td>error-specific, explicit pointer to error, metalinguistic, tree-form</td>
</tr>
<tr>
<td>Shaalan (2005) - Arabic ICALL</td>
<td>WWW interface, morpheme analyzer, syntax analyzer, analyzer; grammar rules, feedback system, knowledge database, students' learning histories, dictionary</td>
<td>Arabian</td>
<td>grammar</td>
<td>error-specific, explicit pointer to error, metalinguistic</td>
</tr>
<tr>
<td>Yang and Akahori (1998)</td>
<td>WWW interface, morpheme analyzer, syntax analyzer, analyzer; grammar rules, feedback system, knowledge database, students' learning histories, dictionary</td>
<td>Japanese</td>
<td>grammar</td>
<td>error-specific, explicit pointer to error, metalinguistic, examples, source links, adaptive individualized</td>
</tr>
<tr>
<td>Heift (2002, 2003)</td>
<td>domain knowledge, analysis module, filtering module, Student Model</td>
<td>German</td>
<td>grammar</td>
<td>error-specific, level-tailored, explicit pointer to error, metalinguistic, error-prioritized, remediation exercises, adaptive individualized, error summary</td>
</tr>
</tbody>
</table>
APPENDIX B

Move 1: Establishing a territory
Step 1 — Claiming centrality and/or
Step 2 — Making topic generalization/s and/or
Step 3 — Reviewing previous research

Move 2: Establishing a niche
Step 1A — Indicating a gap or
Step 1B — Highlighting a problem or
Step 1C — Question raising or
Step 1D — Hypothesizing or
Step 1E — Adding to what is known or
Step 1F — Presenting justification

Move 3: Occupying the niche
Step 1A — Announcing present research descriptively or
Step 1B — Announcing present research purposefully or
Step 2A — Present RQs and/or
Step 2B — Presenting hypotheses
Step 3 — Defining clarifications and/or
Step 4 — Summarizing methods and/or
Step 5 — Announcing principal outcomes and/or
Step 6 — Stating the value of the present research and/or
Step 7 — Outlining the structure of the paper

APPENDIX C
Sample color-coded IADE feedback

Sample numerical feedback generated by IADE
APPENDIX B

Move 1: Establishing a territory
Step 1 — Claiming centrality and/or
Step 2 — Making topic generalization/s and/or
Step 3 — Reviewing previous research

Move 2: Establishing a niche
Step 1A — Indicating a gap or
Step 1B — Highlighting a problem or
Step 1C — Question raising or
Step 1D — Hypothesizing or
Step 1E — Adding to what is known or
Step 1F — Presenting justification

Move 3: Occupying the niche
Step 1A — Announcing present research descriptively or
Step 1B — Announcing present research purposefully or
Step 2A — Present RQs and/or
Step 2B — Presenting hypotheses
Step 3 — Defining clarifications and/or
Step 4 — Summarizing methods and/or
Step 5 — Announcing principal outcomes and/or
Step 6 — Stating the value of the present research and/or
Step 7 — Outlining the structure of the paper

---

APPENDIX C
Sample color-coded IADE feedback

Sample numerical feedback generated by IADE
APPENDIX D
Survey questions

Likert-scale questions
(a lot or very well, somewhat or well, a little, and not at all)

Q-n 1: Did you focus on the moves?
Q-n 2: Did you notice any words/expressions that seemed to be characteristic of certain moves?
Q-n 3: Did the feedback provided by IADE help you see the weaknesses in your drafts?
Q-n 4: How well did you learn the moves?
Q-n 5: Did your research article Introduction writing skills improve?

Yes/no and open-ended questions

Q-n 1: When working with IADE, has the program’s feedback helped you focus on the moves? How? What exactly made you focus on them?
Q-n 2: Has it helped you notice anything about the moves/steps that you might not have paid much attention to before? How? What are these things?
Q-n 3: Did you learn the moves? What makes you say that?
Q-n 4: What do you think helped you learn the moves the most?
Q-n 5: Would you say that you learned the moves as a result of having focused on them? Why?
Q-n 6: Would you say that you can transfer this knowledge to your writing? Why?
Q-n 7: Has the program helped you learn new words/expressions that signal particular moves/steps? How? Please provide examples of the words/expressions you learned for each move/step.
Q-n 8: Do you think you improved your skill of writing a research article introduction? Why?

APPENDIX E
Research article introduction scoring rubric

<table>
<thead>
<tr>
<th>Rating criteria</th>
<th>Effectively</th>
<th>Satisfactorily</th>
<th>Poorly</th>
<th>Score points</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Establishes a territory (stressing interest, indicating importance, and/or showing topic prominence; describing previous research; making topic generalizations—Move 1)</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>2. Establishes a niche (indicating a gap and/or a problem; raising a question; hypothesizing; presenting a justification—Move 2)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Occupies the niche (stating the purpose and/or describing present research, its methodology, and/or its results; stating its value; outlining its structure—Move 3)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total score</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
</tbody>
</table>
**APPENDIX D**

Survey questions

- **Likert-scale questions**
  
  (a lot or very well, somewhat or well, a little, and not at all)

  - Q-n 1: Did you focus on the moves?
  - Q-n 2: Did you notice any words/expressions that seemed to be characteristic of certain moves?
  - Q-n 3: Did the feedback provided by IADE help you see the weaknesses in your drafts?
  - Q-n 4: How well did you learn the moves?
  - Q-n 5: Did your research article Introduction writing skills improve?

- **Yes/no and open-ended questions**

  - Q-n 1: When working with IADE, has the program’s feedback helped you focus on the moves? How? What exactly made you focus on them?
  - Q-n 2: Has it helped you notice anything about the moves/steps that you might not have paid much attention to before? How? What are these things?
  - Q-n 3: Did you learn the moves? What makes you say that?
  - Q-n 4: What do you think helped you learn the moves the most?
  - Q-n 5: Would you say that you learned the moves as a result of having focused on them? Why?
  - Q-n 6: Would you say that you can transfer this knowledge to your writing? Why?
  - Q-n 7: Has the program helped you learn new words/expressions that signal particular moves/steps? How? Please provide examples of the words/expressions you learned for each move/step.
  - Q-n 8: Do you think you improved your skill of writing a research article introduction? Why?

**APPENDIX E**

Research article introduction scoring rubric

<table>
<thead>
<tr>
<th>Rating criteria</th>
<th>Effectively</th>
<th>Satisfactorily</th>
<th>Poorly</th>
<th>Score points</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Establishes a territory (stressing interest, indicating importance, and/or showing topic prominence; describing previous research; making topic generalizations—Move 1)</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>x</td>
</tr>
<tr>
<td>2. Establishes a niche (indicating a gap and/or a problem; raising a question; hypothesizing; presenting a justification—Move 2)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Occupies the niche (stating the purpose and/or describing present research, its methodology, and/or its results; stating its value; outlining its structure—Move 3)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total score</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
APPENDIX F

Example of the learning cycle stimulated by IADE’s feedback (linearly unfolded excerpts from combined think-aloud protocol and Camtasia transcript, Student 27)

Focus on discourse form:
1. The program tells me that the blue is m1, red is m2, and green is m3. First, it’s good to know that I have all three moves. Let’s see what’s in each move. Discipline...
2. 66% belong to m1... above average. Yeah, so, that means in m1 I have about 20% above the average, but below the maximum.
3. Noticing of negative evidence:
4. [looks at the numerical feedback] So there is more room for improvement. [reads the feedback prompt - try to revise this move] OK, now I know what I’m going to do because I’m above the average for m1, I’m below, substantially below the average for m3. So, maybe now I need to shorten m1 and elaborate m3. [reads one of his sentences. highlights part of a sentence] Hmm... my previous research was identified as m3. Why it happened like this? If it appears at the beginning of this paragraph... so I guess maybe the program just sees m1 as it according to its location. Let’s see if I change the position, what will happen.
5. Output modification:
6. [goes down to the revision box. reads his text again. Copy-pastes a piece of text to a place at the beginning]. OK, I already changed it, changed the location of it.
7. Noticing of negative evidence:
8. Hmm... Although I changed the location, it is still regarded as m3. All right, now I know that it is not because of the location. Maybe it’s because of the language...
9. Enhanced understanding:
10. Let’s see some examples. [goes to the help options, Annotated Corpus (AC), and looks at one annotated introduction] [whistles quietly] M3 in this example does not have the previous research review, so I will look at other examples. [opens other annotated texts trying to find review in green] I realized that all the examples, in the m3, there’s no previous research. So, I guess, that’s a problem... I can just keep the previous research in m1 and try to add something in m3, like... let’s see... methods... add some parts like the structure or the summary of the outcomes. So...
11. Output modification:
12. Now I’m focused on the last paragraph of my introduction. I’m trying to add something. [reads the last part of his text. Goes to IADE’s colored text. Goes back to the revision box and adds: “For its methodology, this study used content analysis to find out the historical flow of Chicago downtown. The paper is structured in accordance with the five stages of Florida’s creative economy. The results show that, indeed, the Chicago downtown, although has a relatively short history, matches Florida’s theory.”] Just now I just added some sentences in m3. First of all, I added a summary of my paper. Secondly, I added a brief review of my structure, and then, finally, I added something about the results. Now let’s see how the program will do my next draft. [submits]
13. Focus on discourse form:
14. Now the program told me that it indeed detected some additional m3, and yeah!...
15. I’ll try to revise more, I think. [reads on. highlights a sentence and reads carefully. Highlights it again]
Example of the learning cycle stimulated by IADE's feedback (linearly unfolded excerpts from combined think-aloud protocol and Camtasia transcript, Student 27)

Output modification:
- I'm trying to use the word "assume" as an indicator of m2 hypothesis. [Changes: 57] "So it is possible to set downtown Chicago into Florida's five economic transitions"
- "So this paper assumes that it is possible ..." Let me see if it works and if the program can recognize this change.

Enhanced understanding:
- [opens texts in the AC] This example text does not have this part [meaning step], so I'll look at another one. [clicks on red sentences, most of which are 'problem']
- Since he doesn't find hypothesis, he opens another annotated text. "I'm looking for some indicator words for hypothesis in m2. I already went through two example texts, but both of them m2 just have justification and statement of problems."
- I didn't find indicator words for hypothesis, so I'll just keep looking for it. No, gap, justification, problem, gap ... I went through several other examples, but still I didn't find indicator words for hypothesis in m2. So, I'm wondering why the previous research papers do not have that. Maybe they just don't include hypothesis in my field? [opens the stats in Help Options and confirms his idea] I guess so... because only 0.14% for hypothesis. So, according to my observation, maybe it's true that the other authors in my field don't have hypothesis included in m2. Then I will do more what they do. Let me look at my results.

Focus on discourse form:
- I'm looking at the last sentence ... and to see if I can make some changes and the program can respond to these changes. Maybe I can try to be specific and say "that the results of this paper". Yeah, let's try to do that. [makes the change] "The results show that..." --> "The results of this study show that..." and submits

Noticing of negative evidence:
- And here I find another misunderstanding. [reads the sentence] My intent is to ... I guess it could be m2 or m3, but not m1, so I'm thinking, if it is m2 or m3. [reads the sentence aloud again and takes time to think] It is hypothesis, yeah.

Enhanced understanding:
- Let's see some examples. I'll find out what kind of indicators words the previous papers. The annotated corpus...
- Hmm... no? it is still recognized as m1. Let's see some examples. I'll find out what kind of indicators words the previous papers. The annotated corpus...
- This example text does not have this part [meaning step], so I'll look at another one. [clicks on red sentences, most of which are 'problem']
- Since he doesn't find hypothesis, he opens another annotated text. "I'm looking for some indicator words for hypothesis in m2. I already went through two example texts, but both of them m2 just have justification and statement of problems."
- I didn't find indicator words for hypothesis, so I'll just keep looking for it. No, gap, justification, problem, gap ... I went through several other examples, but still I didn't find indicator words for hypothesis in m2. So, I'm wondering why the previous research papers do not have that. Maybe they just don't include hypothesis in my field? [opens the stats in Help Options and confirms his idea] I guess so... because only 0.14% for hypothesis. So, according to my observation, maybe it's true that the other authors in my field don't have hypothesis included in m2. Then I will do more what they do. Let me look at my results.

Focus on discourse form:
- I'm looking at the last sentence ... and to see if I can make some changes and the program can respond to these changes. Maybe I can try to be specific and say "that the results of this paper". Yeah, let's try to do that. [makes the change] "The results show that..." --> "The results of this study show that..." and submits

Noticing of negative evidence:
- And here I find another misunderstanding. [reads the sentence] My intent is to ... I guess it could be m2 or m3, but not m1, so I'm thinking, if it is m2 or m3. [reads the sentence aloud again and takes time to think] It is hypothesis, yeah.

Enhanced understanding:
- Let's see some examples. I'll find out what kind of indicators words the previous papers. The annotated corpus...
- This example text does not have this part [meaning step], so I'll look at another one. [clicks on red sentences, most of which are 'problem']
- Since he doesn't find hypothesis, he opens another annotated text. "I'm looking for some indicator words for hypothesis in m2. I already went through two example texts, but both of them m2 just have justification and statement of problems."
- I didn't find indicator words for hypothesis, so I'll just keep looking for it. No, gap, justification, problem, gap ... I went through several other examples, but still I didn't find indicator words for hypothesis in m2. So, I'm wondering why the previous research papers do not have that. Maybe they just don't include hypothesis in my field? [opens the stats in Help Options and confirms his idea] I guess so... because only 0.14% for hypothesis. So, according to my observation, maybe it's true that the other authors in my field don't have hypothesis included in m2. Then I will do more what they do. Let me look at my results.

Focus on discourse form:
- I'm looking at the last sentence ... and to see if I can make some changes and the program can respond to these changes. Maybe I can try to be specific and say "that the results of this paper". Yeah, let's try to do that. [makes the change] "The results show that..." --> "The results of this study show that..." and submits
APPENDIX G

Example of sequential and causal relation between the elements of the learning cycle stimulated by IADE’s feedback (interview transcript, Student 43)

Focus on discourse form and noticing of negative evidence:
Researcher (following up on the observation note [doesn’t like the distribution of her moves]): What did you mean when you said that your introduction is “not that centralized”?
Student 43: I mean not coherent, not logical, not as should be.
Researcher: How do you know?
Student 43: Because I saw how colors are all over the place.

Output modification:
Researcher (following up on the observation note [decides to try combining a green and a blue sentence]): Why did you decide to combine those two blue and green sentences at the beginning?
Student 43: Oh, because it wasn’t right. They should both be move 1, but I didn’t write like that. I thought if I combine it helps, but it doesn’t.
Researcher (following up on the answer and on the observation note [combining doesn’t work, so he checks examples in the AC and makes a number of changes at lexical level]): So that’s why you changed that part several times?

Enhanced understanding:
Student 43: Yes. I look at the corpus and I didn’t see “we” in move 1, but I saw “we” in move 3. So I finally understood the problem there.

Researcher (following up on the answer and on the observation note [repeatedly looks at examples of moves/steps in HO and then highlights/reads his own sentences]): Yes, I noticed that you used the Help Options to see examples and then you went back to your own sentences. Why were you doing that?
Student 43: I just compare corpus with my sentence. Because I can see that they are different, and then I think why it’s different and I find out why. I like also definitions because if I go to definitions and examples, then I understand little things … like why I can’t use “we” in move 1. I understand that “we” tells about this research, not research in general, right? I didn’t pay attention before.
Researcher: So that helps?

Output modification:
Student 43: Yeah, sure. Then I know finally what to change, and it works.

Focus on discourse form:
Researcher: How do you know?
Student 43: Because I see the colors and I know what to do next to make it good.
APPENDIX G

Example of sequential and causal relation between the elements of the learning cycle stimulated by IADE's feedback (interview transcript, Student 43)

Focus on discourse form and noticing of negative evidence:
Researcher (following up on the observation note [doesn't like the distribution of her moves]): What did you mean when you said that your introduction is "not that centralized"?
Student 43: I mean not coherent, not logical, not as should be.
Researcher: How do you know?
Student 43: Because I saw how colors are all over the place.

Output modification:
Researcher (following up on the observation note [decides to try combining a green and a blue sentence]): Why did you decide to combine those two blue and green sentences at the beginning?
Student 43: Oh, because it wasn't right. They should both be move 1, but I didn't write like that. I thought if I combine it helps, but it doesn't.
Researcher (following up on the answer and on the observation note [combining doesn't work, so he checks examples in the AC and makes a number of changes at lexical level]): So that's why you changed that part several times?

Enhanced understanding:
Student 43: Yes. I look at the corpus and I didn't see "we" in move 1, but I saw "we" in move 3. So I finally understood the problem there.

Researcher (following up on the answer and on the observation note [repeatedly looks at examples of moves/steps in HO and then highlights/reads his own sentences]): Yes, I noticed that you used the Help Options to see examples and then you went back to your own sentences. Why were you doing that?
Student 43: I just compare corpus with my sentence. Because I can see that they are different, and then I think why it's different and I find out why. I like also definitions because if I go to definitions and examples, then I understand little things ... like why I can't use "we" in move 1. I understand that "we" tells about this research, not research in general, right? I didn't pay attention before.
Researcher: So that helps?

Output modification:
Student 43: Yeah, sure. Then I know finally what to change, and it works.

Focus on discourse form:
Researcher: How do you know?
Student 43: Because I see the colors and I know what to do next to make it good.