Learner Fit in Scaling Up Automated Writing Evaluation

Elena Cotos, Iowa State University, Ames, IA, USA
Sarah Huffman, Iowa State University, Ames, IA, USA

ABSTRACT
Valid evaluations of automated writing evaluation (AWE) design, development, and implementation should integrate the learners’ perspective in order to ensure the attainment of desired outcomes. This paper explores the learner fit quality of the Research Writing Tutor (RWT), an emerging AWE tool tested with L2 writers at an early stage of its development. Employing a mixed-methods approach, the authors sought to answer questions regarding the nature of learners’ interactional modifications with RWT and their perceptions of appropriateness of its feedback about the communicative effectiveness of research article Introductions discourse. The findings reveal that RWT’s move, step, and sentence-level feedback provides various opportunities for learners to engage with the revision task at a useful level of difficulty and to stimulate interaction appropriate to their individual characteristics. The authors also discuss insights about usefulness, user-friendliness, and trust as important concepts inherent to appropriateness.

Keywords: Appropriateness, Automated Writing Evaluation, Discourse-Level Feedback, Effectiveness, Interactional Modification, Learner Fit

INTRODUCTION
Since its beginnings over fifty years ago, Automated Writing Evaluation (AWE) has gained increased popularity as well-considerable technological advancement. Early AWE software, developed to reduce teachers’ workload by automating the scoring of student essays, analyzed the quality of texts by examining language at the surface level (Page, 2003). Modern day AWE software, such as Criterion (Educational Testing Service), MY Access! (Vantage Learning), and Intelligent Essay Assessor (Pearson Knowledge Technologies), employ natural language processing techniques to enable more complex analyses of writing for performance-specific feedback. These products’ scoring and feedback affordances are promoted as being capable of meeting the needs of L2 learners, writing teachers, and institutional administrators.

However, despite the promising potential of AWE, its effectiveness has been the subject of a strenuous debate. On the one hand, AWE programs are deemed to support process writing approaches valued for multiple drafting and scaffolding feedback (Hyland, 2003; Hyland & Hyland, 2006). On the other hand, AWE effects
on the development of writing skills are doubted and even considered harmful (Cheville, 2004). However, the debate largely feeds on empirically unsupported arguments about whether or not AWE should be used rather than how it should be used to better serve the end users (Chen & Cheng, 2008; Grimes & Warschauer, 2010). It has also been pointed out that the debate over AWE effectiveness overlooks design issues such as the lack of relevant theoretical grounding, heavily form-focused feedback, and unspecified learner needs, which are bound to affect AWE impact on L2 writing if not accounted for as the programs are designed (Cotos, 2012). AWE developers have relied on psychometric evidence of accuracy and reliability, but disregarded the possible consequences of re-purposing automated scoring technology from intended summative to formative assessment, ignoring the need to re-conceptualize AWE design. Along these lines, we believe that AWE technologies should be evaluated from the earliest stages of their development, and that the learners’ perspective on the use of AWE for a given task, in particular, should be a fundamentally significant viewpoint in conceptualizing the design, development, and implementation of such tools in order to enhance their effectiveness.

These issues have been considered in the design of the Research Writing Tutor (RWT), an innovative, genre-specific, web-based tool that analyzes the research article Introduction, Methods, Results, and Discussion/Conclusion sections in terms of discourse units that build the communicative effectiveness of each of these sections. RWT represents a scale-up from an earlier prototype - IADE, a program informed by Interactionist SLA, skill acquisition theory, systemic functional linguistics, and genre analysis (Cotos, 2009). IADE analyzes research article Introductions by classifying texts into rhetorical moves (Swales, 1981, 2004) and generates color-coded feedback on the discourse structure of student texts. It also compares student texts with a corpus of Introductions published in fifty academic domains and provides numeric feedback on how well students’ writing approximates the writing in their field. The approach to IADE’s design and empirical evaluation (Cotos, 2010) have motivated scaling up to a more fine-grained operational design of RWT, which not only includes improved functionality of features, but also draws from systematic analyses of formative data obtained from test implementations aimed at validating design decisions and informing continuous development of this emerging tool.

This mixed-methods study is one such test/evaluation implementation, which was aimed at examining the potential effectiveness of RWT considering its learner fit quality. Learner fit is defined by Chapelle (2001) as the capacity of computer-assisted language learning (CALL) materials to engage learners with the language at an appropriate level of difficulty given their individual characteristics. Effectiveness is, thus, regarded here as a user-centric factor rather than a system-centric psychometric measure. Qualitative and quantitative data were gathered when the tool was piloted for the first time with nine learners of English, who were enrolled in a graduate-level academic writing course. Multiple data sources were triangulated to understand how learners perceived, responded to, and engaged with RWT feedback. Our findings regarding learner fit include appropriateness constructs in terms of the level of linguistic difficulty, which is reflected in the degree of interaction, and in terms of usefulness, which is viewed as a combination of helpfulness, user-friendliness, and trust.

**LITERATURE REVIEW**

**Evaluation of Effectiveness**

Integral to the discussion of effectiveness of AWE tools is the question of what “effectiveness” constitutes. The effectiveness of automated scoring of writing has long been viewed through the lens of educational measurement where it is considered effective if it is reliably comparable to human evaluations (for details see Dikli, 2006). With a shift in focus towards AWE use for formative assessment purposes,
Related Content

TeLCU: A Model for Technology-Conditioned Language and Literacy Change
[www.irma-international.org/chapter/telcu-model-technology-conditioned-language/36004/](www.irma-international.org/chapter/telcu-model-technology-conditioned-language/36004/)

Skype-Based Tandem Language Learning and Web 2.0
[www.irma-international.org/chapter/skype-based-tandem-language-learning/21940/](www.irma-international.org/chapter/skype-based-tandem-language-learning/21940/)

Feasibility and Acceptability of In-Depth Annotated Parallel Corpus-Aided Translation Teaching

Translator Intercultural Competence: A Model, Learning Objectives, and Level Indicators
[www.irma-international.org/chapter/translator-intercultural-competence/120789/](www.irma-international.org/chapter/translator-intercultural-competence/120789/)

Developing Language Awareness in Primary School Children with Multilingual Virtual Talking Books: First Results of the Pilot Study