AWE for Formative Assessment: Investigating Accuracy and Efficiency as Part of Argument-based Validation

Jim Ranalli
Stephanie Link
Evgeny Chukharev-Hudilainen
<table>
<thead>
<tr>
<th>Automated Essay Scoring (AES)</th>
<th>Automated Writing Evaluation (AWE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Testing Context</td>
<td>Classroom Context</td>
</tr>
<tr>
<td>System-centric Research</td>
<td>User-centric Research</td>
</tr>
<tr>
<td>Developers</td>
<td>Practitioners</td>
</tr>
</tbody>
</table>
Interpretation/Use Argument

Development Stage

“... tends to produce evidence that supports proposed interpretations and uses, because any indication of a flaw in the assessment design or a weakness in the IUA triggers an effort to fix the problem”

Appraisal Stage

“... the IUA should be challenged, preferably by a neutral or skeptical evaluator ... [it] would provide a critical review of the assumptions built into the IUA [and] include empirical investigations of the most questionable assumptions.”

Kane, 2013, p. 17
A Validity Argument for AWE as Diagnostic Assessment
Chapelle, Cotos, & Lee, 2012

Revision attempts following *Criterion* feedback

- No change: 160
- Remove: 10
- Add: 20
- Delete: 20
- Change: 60
- Transpose: 10

Frequency
A Validity Argument for AWE as Diagnostic Assessment

Chapelle, Cotos, & Lee, 2012

**Study 1**
Evaluation: *Criterion* feedback provides students with accurate information to target relevant areas for revision/improvement/learning.

**Study 2**
Utilization: Diagnostic results on the quality of academic writing obtained from *Criterion* are useful for students to make decisions on revisions.
Context of the Study

English Placement Test

Graduates

Undergraduates

101D
research writing

101C

Criterion

genre, discourse, rhetorical focus

101B

Criterion

sentence- and paragraph-level focus; more form-oriented
Study 1

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

(Chapelle, Cotos, and Lee, 2012)

Assumption: The feedback is 80% accurate.
Since new e-rater microfeatures must demonstrate an 80% level of precision ... before they are approved for integration into the e-rater scoring engine, we might assume that they are performing well—unless we have evidence to the contrary.

Quinlan, Higgins, & Wolff, 2009, p.18
Our Definition of Accuracy is Based on ...

- Precision/false positives
- Categorization of feature
- Explanation/suggested remedial action, if any
Study 1 Research Question

How accurate is Criterion feedback in terms of the errors most commonly identified in our students’ writing?
<table>
<thead>
<tr>
<th>Error Type</th>
<th>Frequency</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Repetition of Words</td>
<td>3997</td>
<td>Style</td>
</tr>
<tr>
<td>Missing or Extra Article</td>
<td>3456</td>
<td>Usage</td>
</tr>
<tr>
<td>Spelling</td>
<td>2556</td>
<td>Mechanics</td>
</tr>
<tr>
<td>Missing Comma</td>
<td>2031</td>
<td>Mechanics</td>
</tr>
<tr>
<td>Preposition Error</td>
<td>1704</td>
<td>Usage</td>
</tr>
<tr>
<td>Fragments</td>
<td>1567</td>
<td>Grammar</td>
</tr>
<tr>
<td>Subject-Verb Agreement</td>
<td>1478</td>
<td>Grammar</td>
</tr>
<tr>
<td>Extra Comma</td>
<td>1259</td>
<td>Mechanics</td>
</tr>
<tr>
<td>Ill-formed Verbs</td>
<td>1235</td>
<td>Grammar</td>
</tr>
<tr>
<td>Determiner Noun Agreement</td>
<td>1218</td>
<td>Usage</td>
</tr>
<tr>
<td>Run-on Sentences</td>
<td>1151</td>
<td>Grammar</td>
</tr>
<tr>
<td>Passive Voice</td>
<td>1128</td>
<td>Style</td>
</tr>
<tr>
<td>Compound Words</td>
<td>1042</td>
<td>Mechanics</td>
</tr>
<tr>
<td>Confused Words</td>
<td>974</td>
<td>Usage</td>
</tr>
</tbody>
</table>
Compound Words (3)

The population of the world is getting bigger and bigger due to the need of people to have babies, they even worry about whether they will be able to have more. It is interesting that some of them do not like to have children. Although a minority of people want to bring children to a world full of danger and chaos, having children brings meaning to one's life and it is a way to procreate. To begin with, we cannot deny that children bring happiness and meaning to our lives. I think having a baby is similar to having an aim in our lives. Most of us have targets in life.
Methodology

Sampling of data for coding

- 10 error categories
- 2 trained raters
- Calibration set: 360 errors
- Annotation set: 700 errors

Manual coding of feedback

- CyWrite Corpus Annotation Tool
- 24 decision rules
- Agreement on the calibration set: Krippendorff’s $\alpha = .72$
Missing or Extra Article

is important part when choosing career because we are interested in our work and salary is important <<< factor because we can reward more salary when we work hard we can work more happy.

You may need to use an article before this word. Consider using the article a.
## Study 1: Results

<table>
<thead>
<tr>
<th>Error category</th>
<th>n</th>
<th>Completely accurate</th>
<th>Partially accurate</th>
<th>Not accurate</th>
<th>% Completely + partially accurate</th>
<th>% Completely accurate</th>
<th>Published precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Confused words</td>
<td>70</td>
<td>45</td>
<td>6</td>
<td>19</td>
<td>72.86</td>
<td>64.29</td>
<td></td>
</tr>
<tr>
<td>Determiner-noun agreement</td>
<td>70</td>
<td>65</td>
<td>2</td>
<td>3</td>
<td>95.71</td>
<td>92.86</td>
<td></td>
</tr>
<tr>
<td>Extra comma</td>
<td>70</td>
<td>33</td>
<td>6</td>
<td>31</td>
<td>55.71</td>
<td>47.14</td>
<td></td>
</tr>
<tr>
<td>Fragment</td>
<td>70</td>
<td>68</td>
<td>2</td>
<td>2</td>
<td>97.14</td>
<td>97.14</td>
<td></td>
</tr>
<tr>
<td>Ill-formed verbs</td>
<td>70</td>
<td>63</td>
<td>3</td>
<td>4</td>
<td>94.29</td>
<td>90.00</td>
<td></td>
</tr>
<tr>
<td>Missing comma</td>
<td>70</td>
<td>44</td>
<td>11</td>
<td>15</td>
<td>78.57</td>
<td>62.86</td>
<td></td>
</tr>
<tr>
<td>Missing or extra article</td>
<td>70</td>
<td>38</td>
<td>18</td>
<td>16</td>
<td>77.14</td>
<td>54.29</td>
<td></td>
</tr>
<tr>
<td>Preposition error</td>
<td>70</td>
<td>33</td>
<td>23</td>
<td>14</td>
<td>80.00</td>
<td>47.14</td>
<td></td>
</tr>
<tr>
<td>Run-on sentences</td>
<td>70</td>
<td>42</td>
<td>9</td>
<td>19</td>
<td>72.86</td>
<td>60.00</td>
<td></td>
</tr>
<tr>
<td>Subject-verb agreement</td>
<td>70</td>
<td>54</td>
<td>4</td>
<td>12</td>
<td>82.86</td>
<td>77.14</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>80.71</strong></td>
<td><strong>69.29</strong></td>
<td></td>
</tr>
</tbody>
</table>
Study 1 Discussion

• Depending on the criteria one adopts, Criterion is marginally adequate, or inadequate, in terms of accuracy for the intended use as formative assessment.

• Some features are clearly problematic:
  – missing comma errors,
  – missing article errors, and
  – preposition errors
Study 2

Assumption: Students are efficient, in terms of both performance and mental effort, at using the feedback to correct errors at least 60% of the time.

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

(Chapelle, Cotos, and Lee, 2012)
Why include mental effort?

Working memory’s role in coordinating writing processes is well established.

Skilled writing involves heavy demands on cognitive processing.

“When the cost/benefit ratio becomes prohibitive ... people refrain from seeking feedback.”

Olive, 2012

e.g., Hayes, 2006; Torrance & Galbraith, 2008

Hattie & Timperley, 2007, p. 94
Study 2 Research Questions

How efficient are students at different proficiency levels at
• distinguishing between accurate and inaccurate CFB?
  and
• using CFB to correct errors?
Methodology

**Students**

101B \( n = 36 \),  
101C \( n = 46 \)

40-minute web-based task: error correction plus mental effort ratings

Internal consistency of mental effort ratings, Cronbach’s \( \alpha = .95 \)

**Raters**

2 raters

Polytomous scoring:  
0=not correct, 1=partially correct, 2=fully correct

Inter-rater reliability  
Cronbach’s \( \alpha = .93 \)
Error-correction task based on accurate feedback (10 accurate)

Mental effort ratings

Part 1

Determiner-Noun Agreement (1)

In order to get accurate information and sources, I need to interview a lot of people that related to the news. Thus, interacting with a variety of people makes me a more patient people that can better communicate with different kinds of people.

Make your correction here.

In order to get accurate information and sources, I need to interview a lot of people that related to the news. Thus, interacting with a variety of people makes me a more patient people that can better communicate with different kinds of people. Also, I have to deal with the situation when people don’t have willingness to talk with me.

This task required...

very little mental effort  a lot of mental effort

Move the slider by clicking and dragging.
Part 2

Accuracy discrimination and error correction (10 accurate + 10 inaccurate, interspersed randomly)

Mental effort ratings
## Results

Error Correction Performance

<table>
<thead>
<tr>
<th>Level</th>
<th>Part 1</th>
<th>Part 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$M$</td>
<td>$SD$</td>
</tr>
<tr>
<td>101B ($n = 36$)</td>
<td>12.08</td>
<td>2.45</td>
</tr>
<tr>
<td>101C ($n = 46$)</td>
<td>11.33</td>
<td>1.81</td>
</tr>
</tbody>
</table>

Parts 1 and 2 = 20 possible

Wilcoxon signed-rank test

101B: $Z = -2.04$, $p = .04$

101C: $Z = -2.43$, $p = .02$
## Results: Perceived Mental Effort

<table>
<thead>
<tr>
<th>Class</th>
<th>Part 1</th>
<th>Part 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$M$</td>
<td>$SD$</td>
</tr>
<tr>
<td>101B $(n = 36)$</td>
<td>2.09</td>
<td>.65</td>
</tr>
<tr>
<td>101C $(n = 46)$</td>
<td>2.32</td>
<td>1.37</td>
</tr>
</tbody>
</table>

1 = very little mental effort, 7 = a lot of mental effort

Wilcoxon signed-rank test

101B: $Z = -.75$, $p = .46$

101C: $Z = -.31$, $p = .75$
Efficiency = \frac{z\text{MentalEffort} - z\text{Performance}}{\sqrt{2}}

Study 2 Discussion

Our students ... 
• use CFB to make appropriate corrections in about 6 of 10 cases
• do not report high perceived mental effort in 
  • distinguishing between accurate and inaccurate CFB 
  • using CFB to make corrections
• in the lower level course appear to use Criterion FB more efficiently
General discussion

- Limited support for use of Criterion as formative assessment
- Value of argument-based validation of formative assessment
- More accuracy work on recall
- Design changes to enhance AWE tool and thus validation
  - option to turn off specific error types
  - make system data and aggregated student data easy to access
References


Acknowledgments

Our sincere thanks to ... Ahmet Dursun
Kelsey Campbell-Gagen
Kelly Cunningham
Joe Geluso
Thank you!

Questions/Comments?