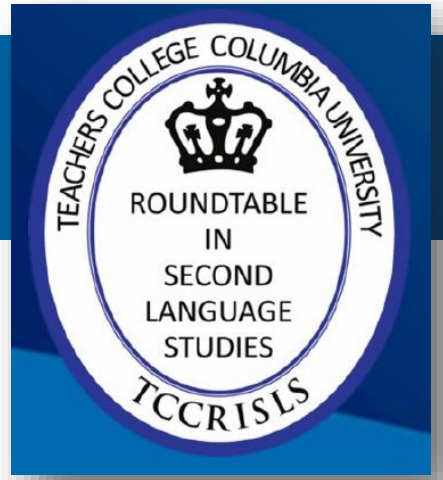
The 3rd Teachers College, Columbia University Roundtable in Second Language Studies



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

Jim Ranalli

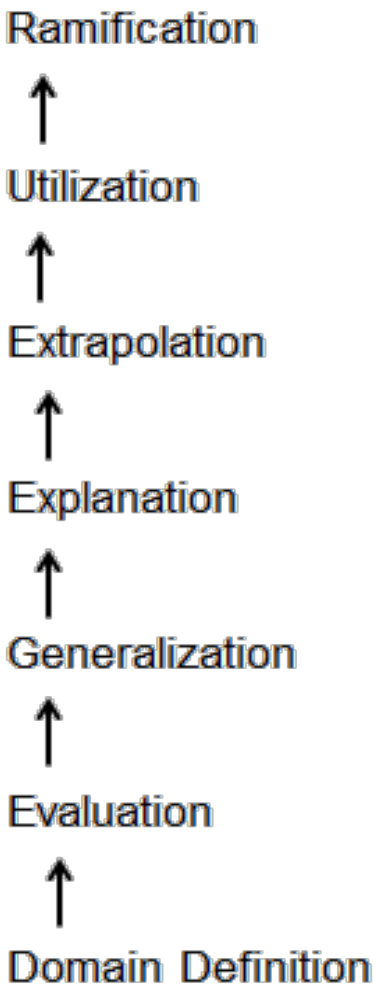
Stephanie Link

Evgeny Chukharev-Hudilainen

IOWA STATE UNIVERSITY

Automated Essay Scoring (AES)	Automated Writing Evaluation (AWE)
Testing Context	Classroom Context
System-centric Research	User-centric Research
Developers	Practitioners

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

Ramification

↑

Utilization

↑

Extrapolation

↑

Explanation

↑

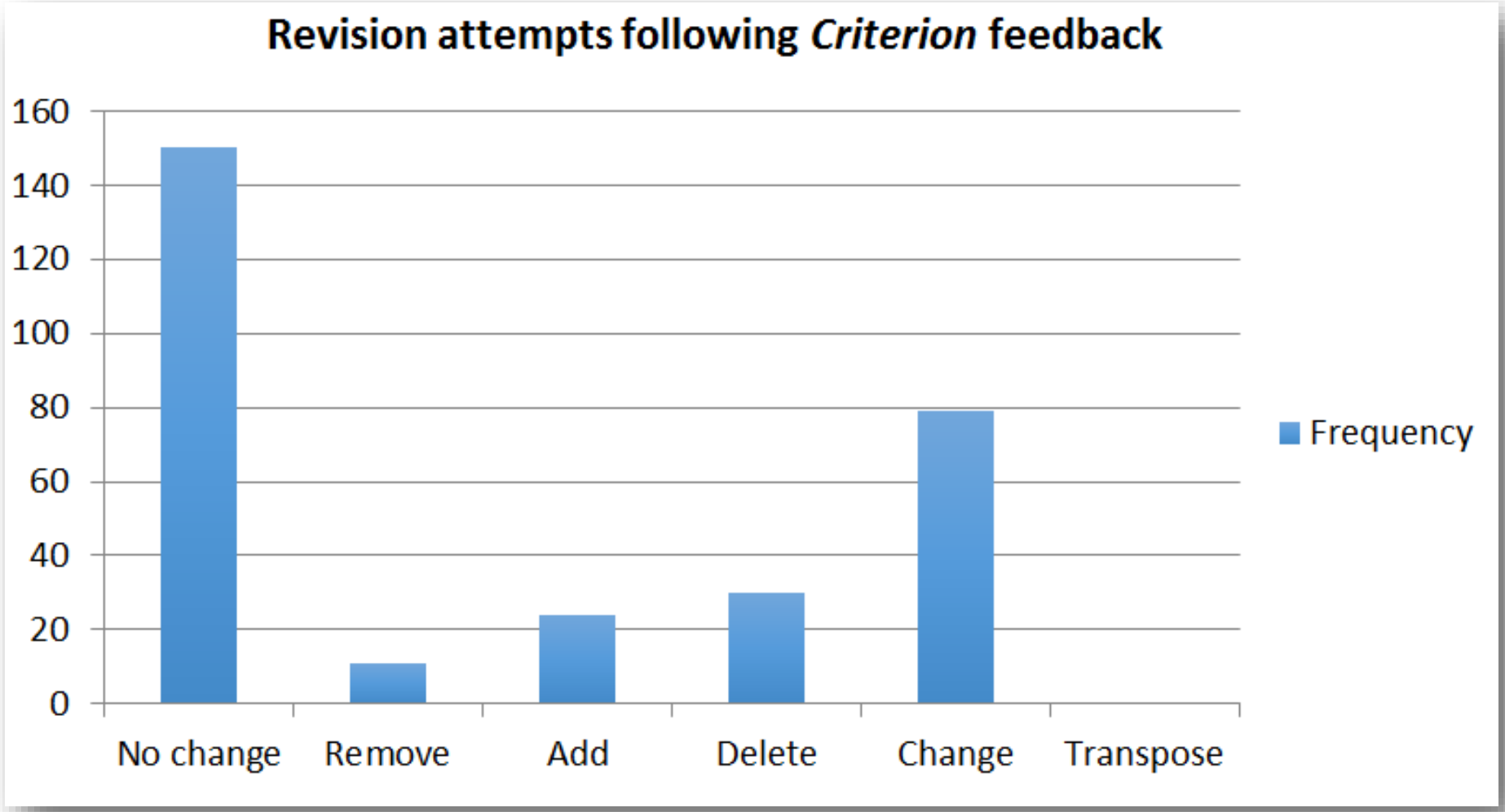
Generalization

↑

Evaluation

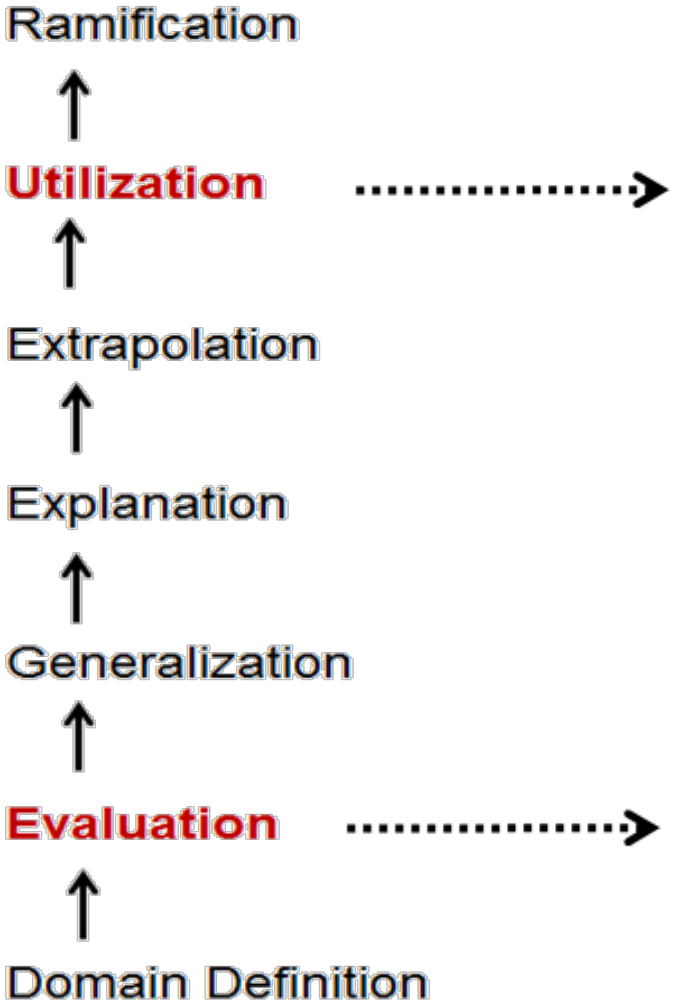
↑

Domain Definition



A Validity Argument for AWE as Diagnostic Assessment

Chapelle, Cotos, & Lee, 2012



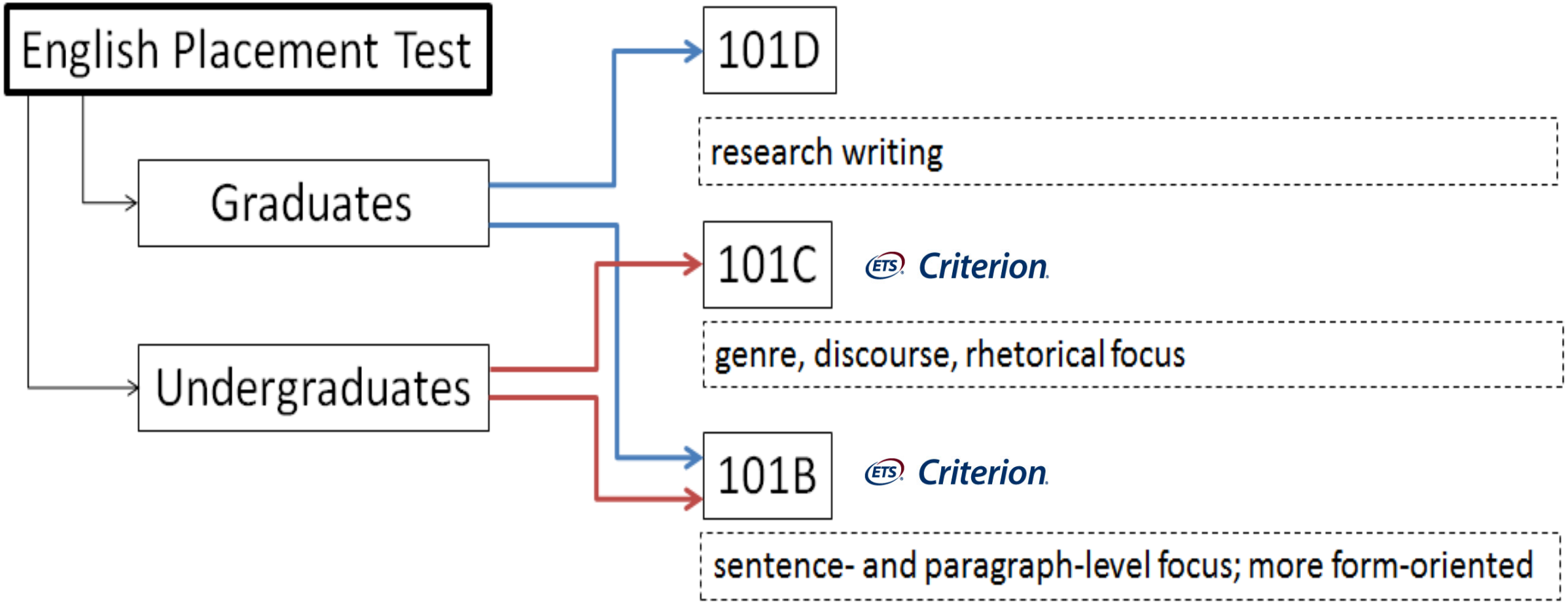
Study 2

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

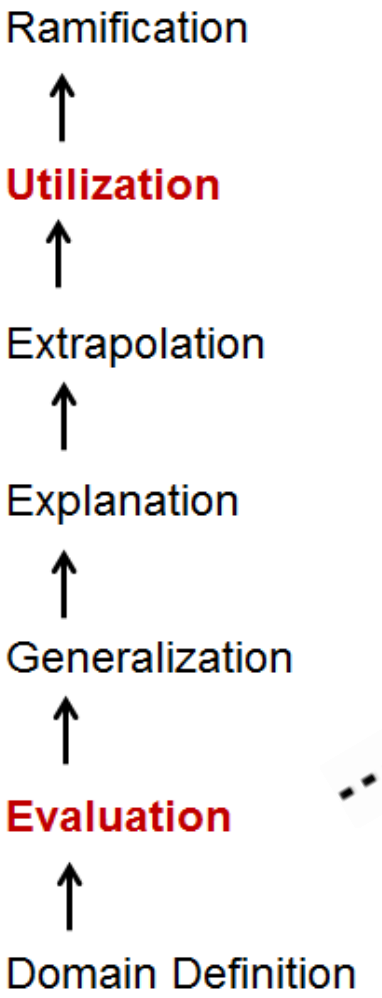
Study 1

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

Context of the Study



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.

2 Dimensions of Accuracy in AWE

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.

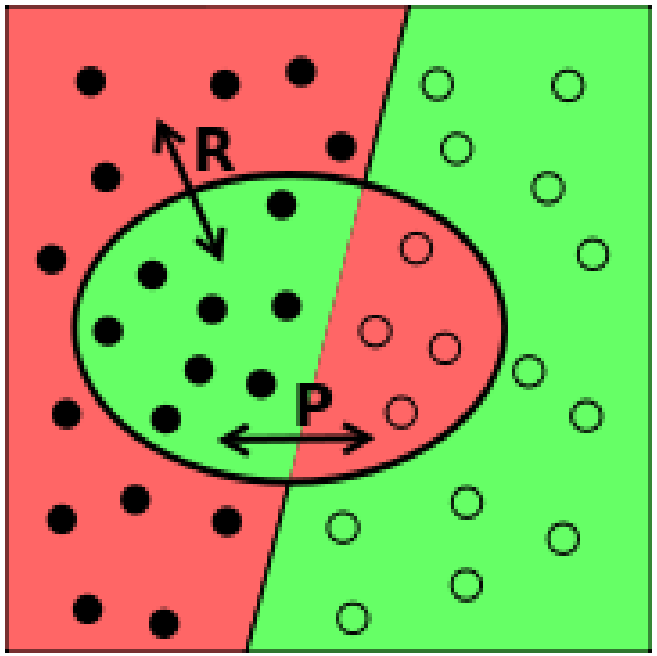
Quinlan, Higgins, & Wolff, 2009, p.18

Precision

false positives

Recall

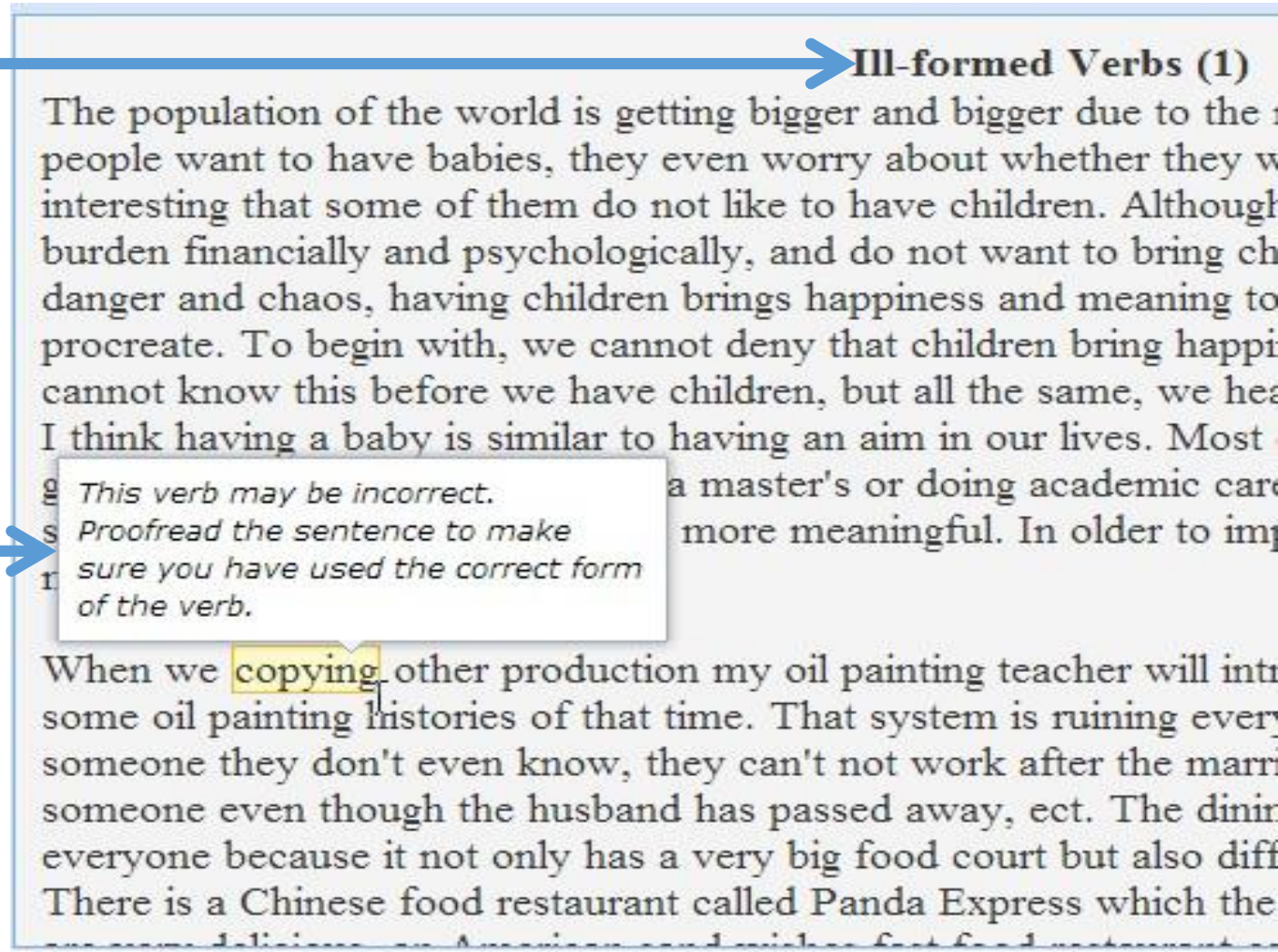
false negatives



"Recall-precision" by Nichtich - Own work. Licensed under Public domain via Wikimedia Commons http://commons.wikimedia.org/wiki/File:Recall-precision.svg#mediaviewer/File:Recall-precision.svg

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?

Error Type	Frequency	Category
Repetition of Words	3997	Style
Missing or Extra Article	3456	Usage
Spelling	2556	Mechanics
Missing Comma	2031	Mechanics
Preposition Error	1704	Usage
Fragments	1567	Grammar
Subject-Verb Agreement	1478	Grammar
Extra Comma	1259	Mechanics
Ill-formed Verbs	1235	Grammar
Determiner Noun Agreement	1218	Usage
Run-on Sentences	1151	Grammar
Passive Voice	1128	Style
Compound Words	1042	Mechanics
Confused Words	974	Usage

Compound Words (3)

The population of the world is getting bigger and bigger due to the need of people to h
people want to have babies, they even worry about whether they will be able to have a
interesting that some of them are unable to have children. Although a minority of peop
burden financially a *These two words should be written as one compound word.* ant to bring children to a world fu
danger and chaos, h nd meaning to one's life and it is
procreate. To begin with, we can not deny that children bring happiness and meaning t
can not know this before we have children, but all the same, we hear about that happi
I think having a baby is similar to having an aim in our lives. Most of us have targets i

Methodology

Sampling of data for coding

Manual coding of feedback

10 error categories

2 trained raters

Calibration set: 360 errors

Annotation set: 700 errors

CyWrite Corpus Annotation Tool

24 decision rules

Agreement on the calibration set:
Krippendorff's α = .72

Welcome | **CyWrite** | Project Management

Annotation | Administration | Monitoring | Disagreements

CyWrite Corpus Annotation Tool

147 sentences to go

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.

- (1) Not accurate
- (2) Partially accurate
- (3) Completely accurate

Submit (Press Enter) or skip this sentence

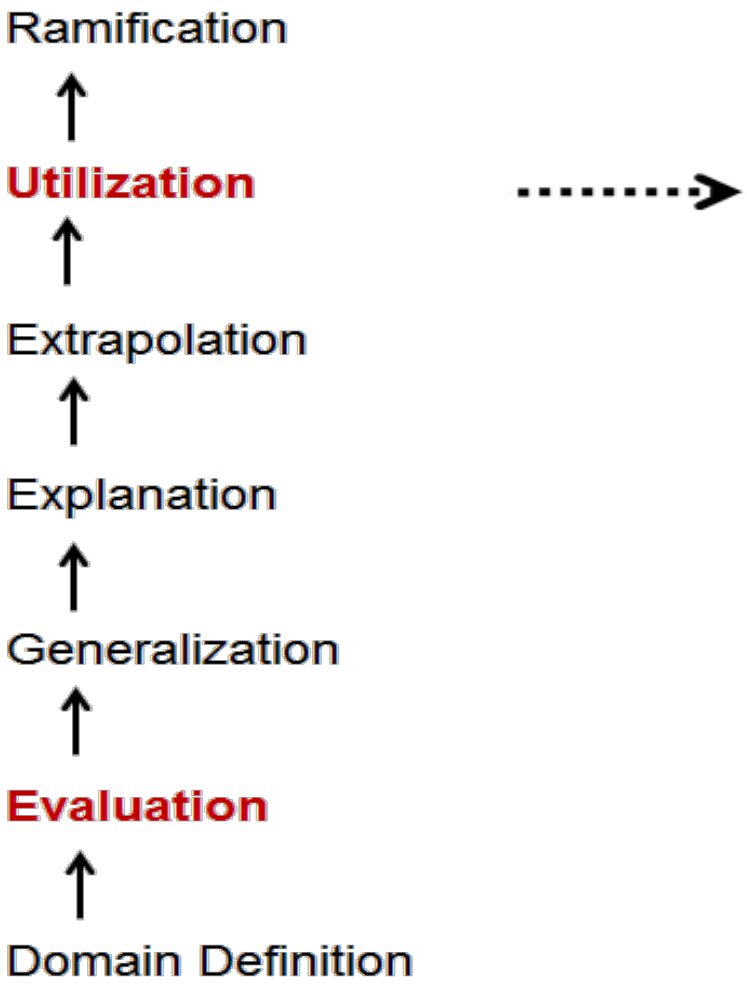
Study 1: Results

Error category	*n*	Completely accurate	Partially accurate	Not accurate	% Completely + partially accurate	% Completely accurate	Published precision
Confused words	70	45	6	19	72.86	64.29	
Determiner-noun agreement	70	65	2	3	95.71	92.86	
Extra comma	70	33	6	31	55.71	47.14	
Fragment	70	68		2	97.14	97.14	
Ill-formed verbs	70	63	3	4	94.29	90.00	
Missing comma	70	44	11	15	78.57	62.86	
Missing or extra article	70	38	18	16	77.14	54.29	90
Preposition error	70	33	23	14	80.00	47.14	80
Run-on sentences	70	42	9	19	72.86	60.00	
Subject-verb agreement	70	54	4	12	82.86	77.14	92
					80.71	**69.29**	

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



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)

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.

Why include mental effort?

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

Olive, 2012

Skilled writing involves heavy demands on cognitive processing.

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

"When the cost/benefit ratio becomes prohibitive … people refrain from seeking feedback."

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 α = .95

Raters

2 raters

Polytomous scoring:

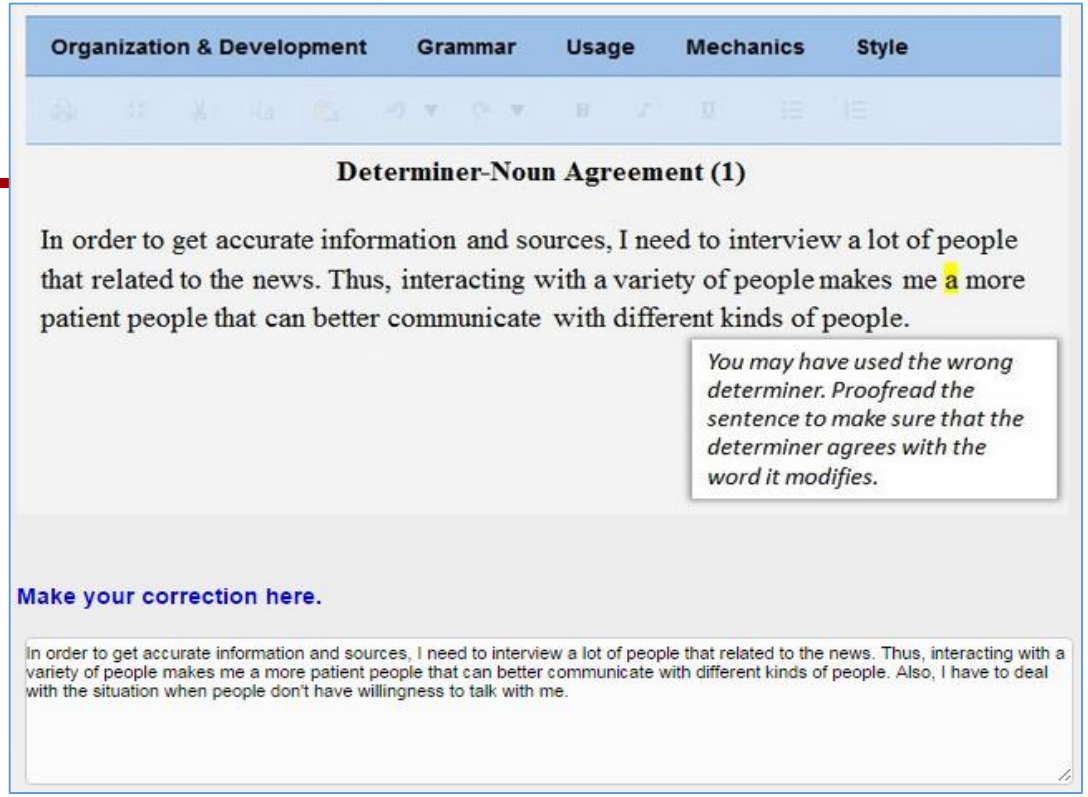
0=not correct, 1=partially correct, 2=fully correct

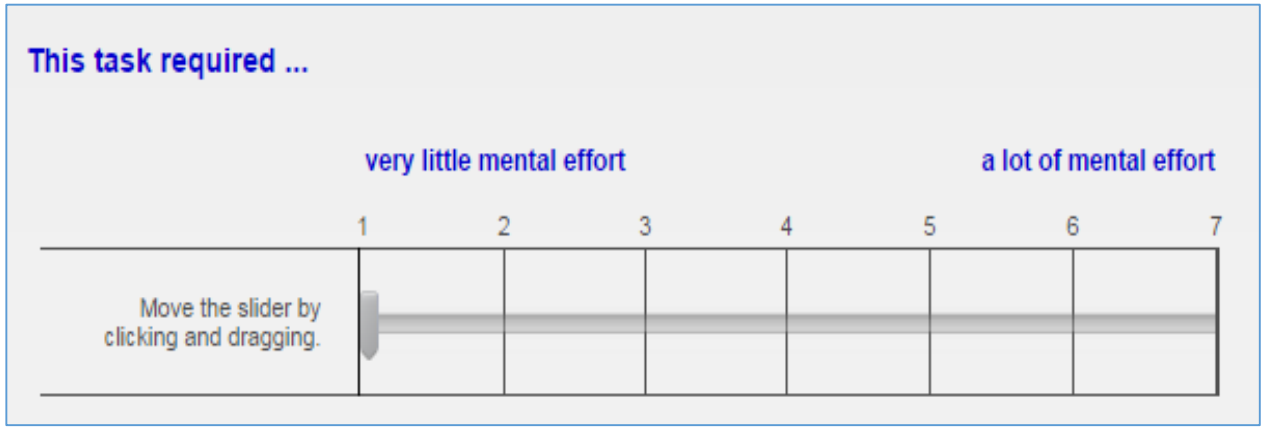
Inter-rater reliability Cronbach's α = .93

Part 1

Error-correction task based on accurate feedback
(10 accurate)

Mental effort ratings

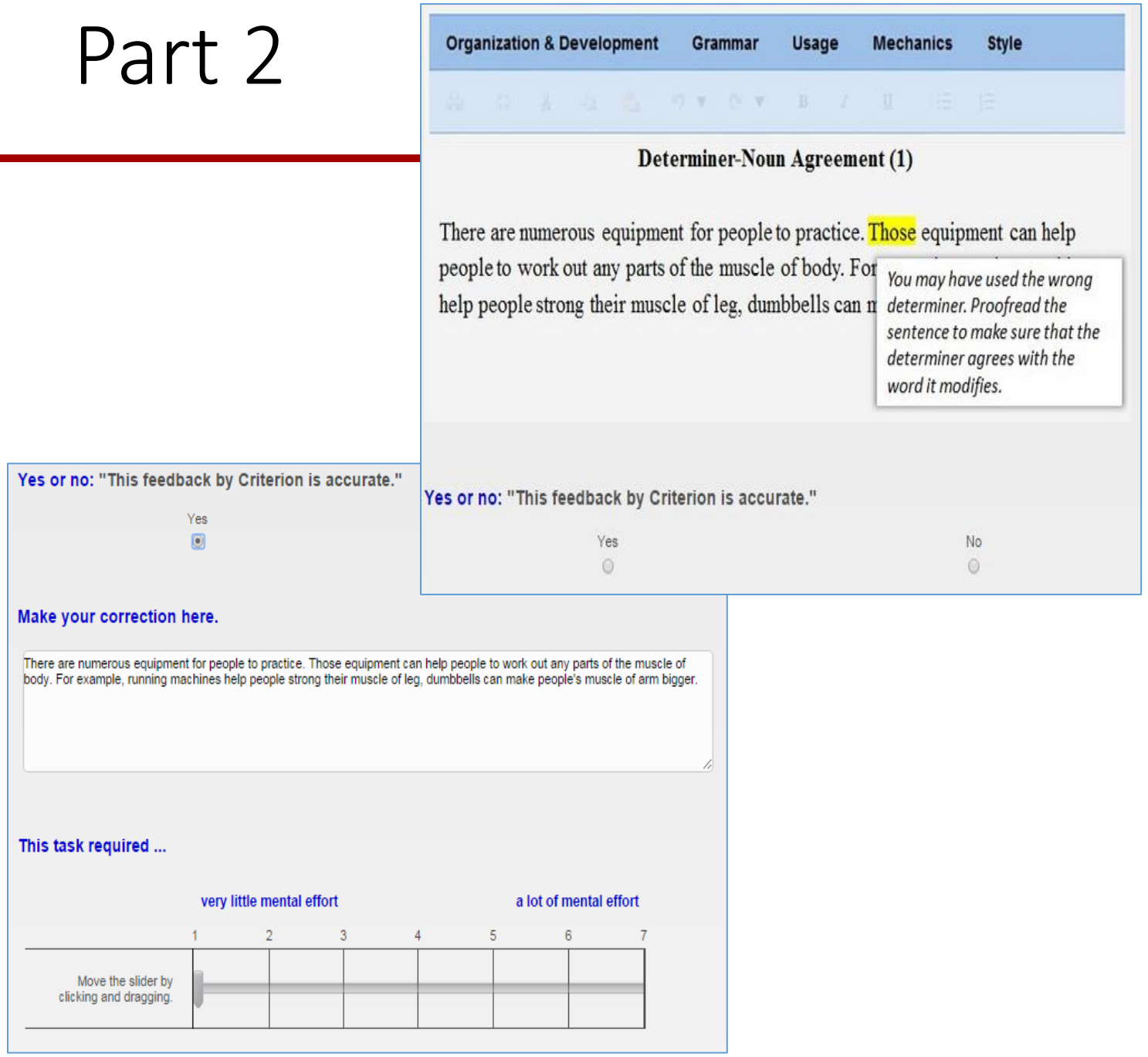




Part 2

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

Mental effort ratings



Results Error Correction Performance

	Part 1 (Accurate only)			Part 2 (Distinguish between accurate and inaccurate)		
Level	*M*	SD	%	*M*	SD	%
101B (*n* =36)	12.08	2.45	60.40	13.22	3.26	66.10
101C (*n*=46)	11.33	1.81	56.60	12.50	4.06	62.50

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

	Part 1		Part 2	
Class	*M*	SD	*M*	SD
101B (*n* =36)	2.09	.65	2.04	.93
101C (*n*=46)	2.32	1.37	2.63	1.34

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

Wilcoxon signed-rank test

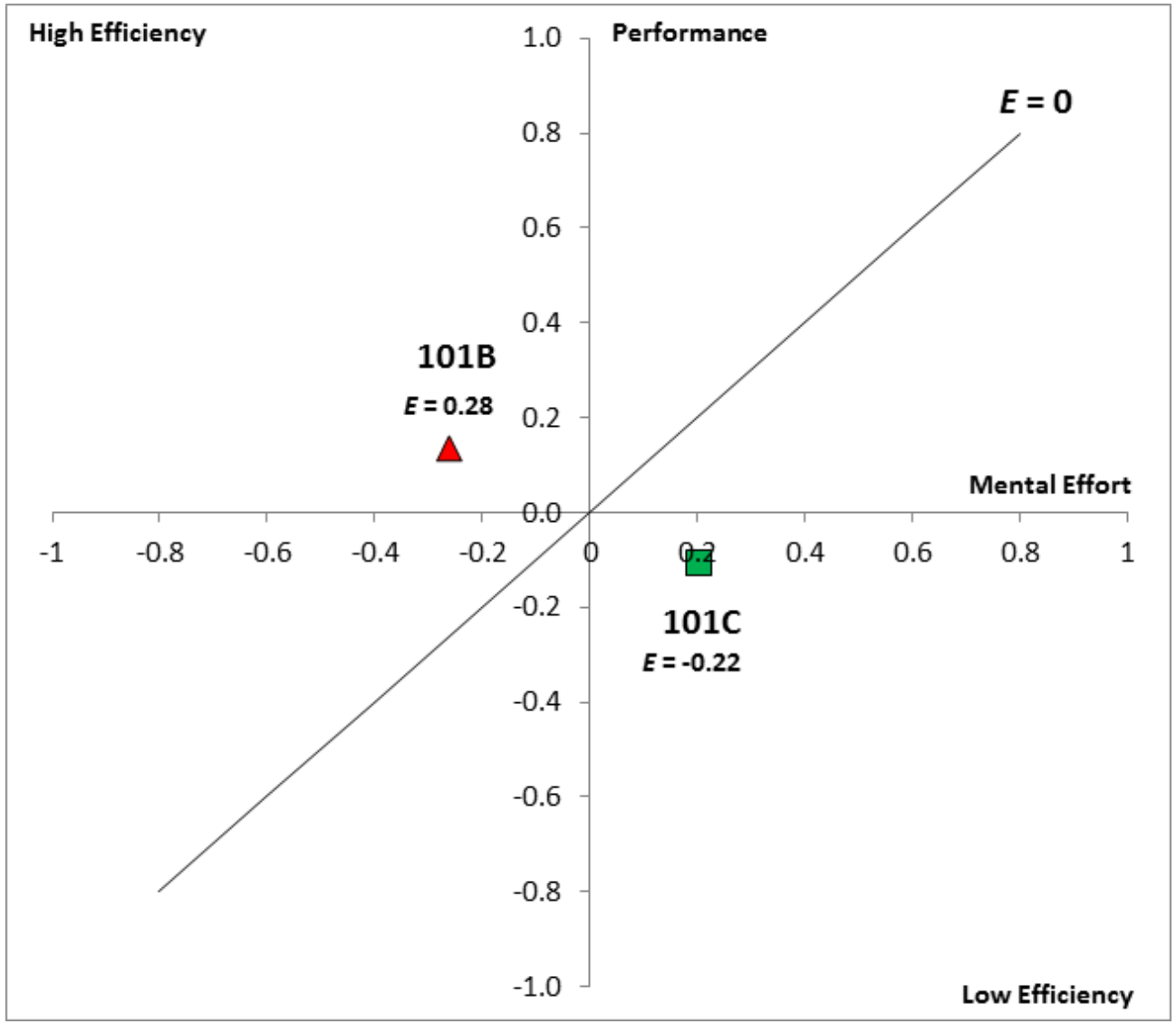
101B: $Z = -.75$, $p = .46$

101C: $Z = -.31$, $p = .75$

Results Efficiency

$$Efficiency = \frac{zMentalEffort - zPerformance}{\sqrt{2}}$$

Paas & Van Merriënboer, 1993; Van Gog & Paas, 2008



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

Chapelle, C. A., Cotos, E., & Lee, J. Y. (2013). *Diagnostic assessment with automated writing evaluation: A look at validity arguments for new classroom assessments. Paper presented at the LTRC 2013, Seoul, Korea.*

Hattie, J., & Timperley, H. S. (2007). The power of feedback. *Review of Educational Research, 77(1), 81-112. doi: 10.3102/003465430298487*

Hayes, J. R. (2006). New directions in writing theory. In C. A. MacArthur, S. Graham & J. Fitzgerald (Eds.), *Handbook of writing research (pp. 28-40). New York: Guilford Press.*

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50(1), 1-73. doi: 10.1111/jedm.12000*

Olive, T. (2012). Working memory in writing. In V. W. Berninger (Ed.), *Past, present, and future contributions of cognitive writing research to cognitive psychology (pp. 485-503). New York:: Psychology Press.*

Paas, F. G. W. C., & Van Merriënboer, J. J. G. (1993). The efficiency of instructional conditions: An approach to combine mental effort and performance measures. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 35(4), 737-743. doi: 10.1177/001872089303500412*

Quinlan, T., Higgins, D., & Wolff, S. (2009). Evaluating the construct coverage of the e-rater scoring engine. Princeton, NJ: Educational Testing Service.

Torrance, M., & Galbraith, D. (2006). The processing demands of writing. In C. A. MacArthur, S. Graham & J. Fitzgerald (Eds.), *Handbook of writing research (pp. 67-80). New York: Guilford Publications.*

van Gog, T., & Paas, F. (2008). Instructional efficiency: Revisiting the original construct in educational research. *Educational Psychologist, 43(1), 16-26. doi: 10.1080/00461520701756248*

Acknowledgments

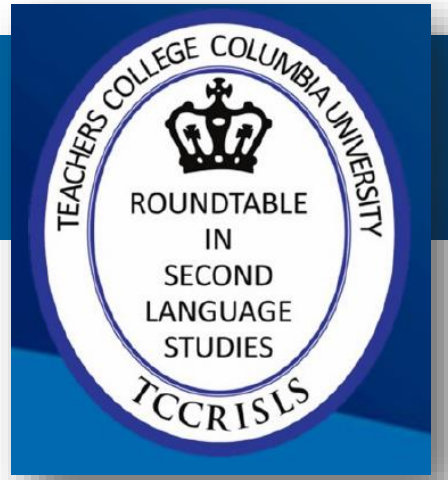
Our sincere thanks to …

Ahmet Dursun

Kelsey Campbell-Gagen

Kelly Cunningham

Joe Geluso



Thank you!

Questions/Comments?