EasyChair Preprint

№ 3780

Using Examples to Support Arguments in an English Language Assessment

Yi Song, Patrick Houghton, Szu-Fu Chao and
Beata Beigman Klebanov

July 7, 2020

Using Examples to Support Arguments in an English Language Assessment

Yi Song

Patrick Houghton

Szu-Fu Chao

Beata Beigman Klebanov

Educational Testing Service

Author Note

The authors declare that there no conflicts of interest with respect to this preprint.

Correspondence should be addressed to Yi Song. Email: ysong@ets.org

Abstract

In this project, we evaluated the quality of written arguments through analyzing examples in test takers' essays in an English language assessment. Altogether we identified 168 examples used to support arguments in 99 essays. Raters were able to recognize various characteristics of the examples in a relatively consistent manner. The results indicated that the number of examples and clarity were significant predictors of the essay quality. Here, we present our analysis approach, results, and implications.

Keywords: argument, English language assessment, writing, evaluating examples

Using Examples to Support Arguments in an English Language Assessment

The project goal is to evaluate the quality of written arguments through analyzing examples in test takers' essays in a worldwide English language assessment. This writing task asks test takers to express their opinions about a given topic. The test takers need to provide reasons and examples in their responses. Argument from example has been identified as a commonly used argumentation scheme (Walton, 1996). Appropriate examples can provide specifics and details in support of a claim and capture the reader's attention. To evaluate the quality of an example and how well it supports a claim, three critical questions (Walton, 1996) that can be asked: (1) Is the example true? (Accuracy); (2) Does the example support the claim? (Relevancy); and (3) Is the example typical? (Typicality) Critical questions help differentiate fallacious arguments from reasonable ones by appraising unwarranted assumptions that may be present in the reasoning underpinning an argument.

Walton's theory has been influential among philosophers, and has been applied to support automated detection of arguments (e.g., Mochales & Ieven, 2009), to develop computational representation of arguments (e.g., Atkinson, Bench-Capon, & McBurney, 2006), and to teach argumentation skills (e.g., Song & Ferretti, 2013; Nussbaum & Edwards, 2011). However, this approach has not been used widely in assessing or scoring written arguments. Here, we apply it to the analysis of written responses in an English language assessment. Our research questions are:

1. Can we reliably annotate characteristics of examples in the test takers' essays?
2. What are some common characteristics of examples in the test takers' essays?
3. Is there any relationship between characteristics of examples and essay scores?

Method

The data was sampled from essays written by 118 non-native English speakers in an international test of academic English proficiency (https://catalog.ldc.upenn.edu/LDC2014T06). In this writing task, test takers were asked to write an essay in response to whether they agree with the statement. "It is better to have broad knowledge of many academic subjects than to specialize in one specific subject." The task required the test takers to use specific reasons and examples to support their answer. Essay scores ranged from 1 to 5 points and reflected their overall quality (M = 3.29; SD = .92).

We identified 169 examples that were given to support relevant arguments in 99 essays. A large majority of these essays provided one or two examples (one example: 41.4%; two examples: 47.5%; three examples: 9.1%; four examples: 2%). Two raters annotated the examples independently with respect to the following characteristics, some of which were grounded in Walton's theory. They identified which side each example supports (broad knowledge, specialized knowledge, or an integrated position) and the number of academic subjects if an example was used to support broad knowledge, annotated the example type (historical events/figures, group, or personal experience), judged the accuracy of historical examples, and evaluated whether the examples are relevant to academic subjects. The raters also scored the clarity and convincingness aspects on a 0-2 scale.

Results

RQ1. Can we reliably annotate characteristics of examples in the test takers' essays?

Most annotation categories had good exact interrater agreements: 95.3% for side, 88.9% for the number of subject areas, 93.7% for type, 98.3% for accuracy, and 87.1% for relevancy.

Two categories, clarity and convincingness, had acceptable exact agreements: 76.6% and 71.40%, respectively. Clarity and convincingness categories were somewhat challenging because raters' understanding of these characteristics were tied to their background knowledge and their perspectives related to the issue of broad knowledge versus specialized knowledge.

RQ2. **What are some common characteristics of examples in the test takers' essays?**

Altogether we identified 169 examples in 99 essays. More examples were used to illustrate the position that broad knowledge is better than specialized knowledge (56.8% vs. 40.2%), and only five examples (3%) addressed both sides. Among 96 examples that were used to support the position that broad knowledge is preferred, 61.4% mentioned two or more subject areas. Seventy-one percent of the examples referred to a particular group (e.g., doctors, scientists, engineers, and college students), 23.7% of the examples used personal experiences (e.g., "although I'm not good in English, I master in math"), and 5.3% cited historical events or figures (e.g., "Einstein was a mathematician"). Additionally, 74.6% of the examples were clearly written, 24.3% of the examples had some clarity issues, and only 1.2% were completely unclear. The majority of the examples (83.4%) were relevant to academic subjects, as required by the writing prompt. Roughly 30% of the examples were convincing, 53.3% of the examples were partially convincing, and 16% were not convincing at all.

RQ3. **Is there any relationship between characteristics of examples and essay scores?**

We ran a multiple regression analysis to examine the relationship between the essay scores (i.e., the dependent variable) and various example characteristics (i.e., independent variables). For clarity, we combined the completely unclear cases with cases that had some clarity issues because there were only two examples that were completely unclear. The model indicated that 21% of the variance in the essay scores could be explained by the example

characteristics; $F(11, 87) = 3.37$, $p < .001$. The regression coefficients for the number of examples and clarity were significant (number of examples: B = .54, $p < .001$; clarity: B = .48; $p < .01$). Therefore, the number of examples provided in an essay and how clearly the ideas were communicated appeared to contribute to the overall quality of the essays.

Discussion

In this project, we examined the characteristics of examples in support of the test takers' argumentative goal and explored the relationship between the example characteristics and essay quality. Our results indicate that we were able to recognize the characteristics of the examples in a relatively consistent manner. The majority of test takers supported one position in their response, and very few presented an integrated position that acknowledged arguments from both sides in different situations, which is a more advanced argumentation skill. In their examples, test takers tended to use a generalizable example involving a particular group of people. They also developed arguments from personal stories. Many test takers were able to communicate their ideas clearly in writing, but we have to note that we ignored grammatical, spelling, and other minor errors that did not interfere with our understanding of the text. Fewer than one-third of the examples were truly convincing in support of the arguments.

Furthermore, the results indicated that the characteristics of examples accounted for some variance in the essay scores. The number of examples and clarity were significant predictors of the essay quality. It is not surprising because people who generate more ideas and have good general writing skills often receive high essay scores because English proficiency tests typically focus on fluency in English written communication rather than the convincingness of the actual arguments. However, it is essential to recognize that the quality of an argument goes well beyond

length, mechanics, grammar, style, vocabulary, and even structure, because arguments in a well-formulated essay may be invalid if supported by poor examples.

References

Atkinson, K., Bench-Capon, T., & McBurney, P. (2006). Computational representation of practical argument. *Synthese, 152*, 157–206.

Mochales, R., & Ieven, A. (2009). Creating an argumentation corpus: Do theories apply to real arguments? A case study on the legal argumentation of the ECHR. Paper presented at the 12th international conference on Artificial Intelligence and Law, Barcelona, Spain.

Nussbaum, E. M., & Edwards, O. V. (2011). Critical questions and argument stratagems: A framework for enhancing and analyzing students' reasoning practices. *Journal of the Learning Sciences, 20*, 443−488.

Song, Y. & Ferretti, R. P. (2013). Teaching critical questions about argumentation through the revising process: effects of strategy instruction on college students' argumentative essays. *Special Issue: Reading and Writing: An Interdisciplinary Journal, 26*(1), 67–90.

Walton, D. (1996). *Argumentation schemes for presumptive reasoning.* Mahwah, NJ: Lawrence Erlbaum.