Hitotsubashi Journal of Arts and Sciences 57 (2016), pp.13-19.

STUDENT MISCONCEPTIONS OF ENGLISH PROFICIENCY EXAM ESSAY EXPECTATIONS

Brent Amburgey

Abstract

Data provided by ETS, the company that creates and administers the TOEFL exam, reveals that Japanese students perform poorly relative to students in other Asian countries. The Japanese English education system's relative lack of focus on speaking and listening skills explains the poor performance on those areas of the test. However, Japanese students do spend significant time studying grammar, vocabulary, and writing, which presents an apparent inconsistency with TOEFL writing scores. This paper examines misunderstanding of what constitutes a *good essay* as a partial explanation of this phenomenon.

I. *Introduction*

Despite Japan's focus on English education, and a governmental push towards improving TOEFL scores specifically (Hongo, 2013; Yoshida, 2013), the results have not been strong. According to data provided by ETS, the company that creates and administers the TOEFL exam, the average total score of Japanese students on the TOEFL iBT test in the year 2015 was 71 (out of a possible 120) points (ETS, 2016, p. 14). This score would place them 25th out of 30 countries in the Asia region with large enough sample sizes to calculate reliable data, meaning the 6th lowest average score. While Japan has consistently received criticism for its English education, this has primarily been aimed at speaking and listening skills, which are not a large point of focus in Japanese schools. This is due to most English courses being oriented towards preparing students for university entrance examinations (Ushioda, 2013, p. 5).

Grammar, vocabulary, and writing skills, on the other hand, are a significant focus of English education in Japan. However, this time and effort is not reflected in students' performance on the TOEFL iBT test. Average scores on the reading and writing sections of the TOEFL iBT are only 1 point higher than on the speaking and listening sections; test takers average 18 points for reading and writing compared with 17 for speaking and listening (ETS, 2016, p. 14).

II. *Research Objectives*

Having established that Japanese test takers' performance on the writing section of the TOEFL iBT appears inconsistent with the amount of time devoted to relevant skills by

coursework in mandatory English courses, two research questions were developed to explore misunderstandings of grading criteria as one possible cause of this disparity:

RQ1. Are the participants able to accurately order official sample scored TOEFL essays from best to worst?

RQ2. To the extent that students struggle with ordering the scored examples, do the justifications they provide for their decisions reveal misunderstandings of what is considered a *good essay*.

III. *Methodology*

1. Design

The purpose of this study was to explore misunderstanding of what constitutes a *good essay* as a possible contributor to Japanese students' low writing section scores on English proficiency exams. To do so, groups of students were given five sample scored TOEFL essays (with their official scores redacted) and asked to place them in order from best to worst on a provided worksheet. The essays had been officially rated from "score-5" (best) to "score-1" (worst). After collecting the completed worksheets, students' rankings were compared with the official scoring.

2. Participants

Participants for this study were students in three Intermediate level English communication courses at Hitotsubashi University. All participants were Japanese and therefore, while individual differences could not be completely eliminated, had roughly similar mandatory English education experiences up until the point of the study. The exercise was part of the essay writing curriculum for the semester, and therefore participation was required. However, students were informed that the inclusion of their results in this study was completely optional and that their participation (or lack thereof) in the study would not impact their grade in the course.

A total of 57 students took part. For the purpose of managing a significant amount of reading, as well as to facilitate a discussion of merits and demerits, students were placed into groups of three to four to complete the task. This resulted in a total of 15 groups across the three classes (four groups per class). There were 12 groups of four and three groups of three.

3. Data and Data Collection

Participants were asked to read the five officially scored sample TOEFL essays, discuss their merits and demerits within their group, and then place them in order from best to worst. A worksheet was provided for groups to indicate their chosen order, provide detailed justification for their choices of "best" and "worst" essays, and a short list of merits/demerits for their middle three choices.

Essay ordering was randomized and each essay was labeled at the top with a letter from A

to E. The essay randomization was different for each class of students to prevent sharing of results between classes. As the exercise was also a precursor to the essay writing curriculum of the course, students were able to be afforded as much time as needed to analyze and rank the essays. Groups generally completed the worksheet within 60 minutes.

4. Analysis

Two forms of analysis were used to evaluate the groups' performance on the task. Firstly, the ordering of essays was considered in terms of (1) how close groups came to the correct ordering of essays and (2) which essays were most often chosen as best and worst.

The second form of analysis was evaluation of written justifications for ordering. In this analysis, coding was created for categories such as "grammar mistakes," "spelling mistakes," "organization," "simple/easy," "strong support," and "advanced vocabulary." It was necessary to make some inferences in coding; for example, understanding "good vocab" to mean "advanced vocabulary." Worksheets were then examined to discover how often each of these codes appeared in their justifications for rankings. This was done in hopes of gaining a better understanding of students' internal criteria for a *good essay*.

IV. *Results*

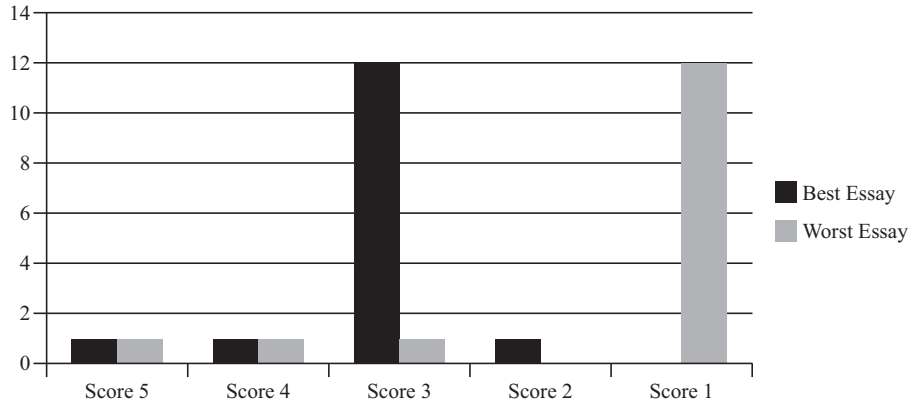
1. Successful Ordering of Officially Rated Essays

Across the three classes and 15 groups, none were able to perfectly order the officially scored essays from best to worst. Two groups had three essays correctly ordered, with one of these coming particularly close, switching only the order of the score-4 and score-3 essays. Of the remaining groups, two groups had two essays correctly ordered, ten groups had one essay in the correct position, and one group had no essays ordered correctly. However, the number of correctly ordered essays does not tell the whole story of students' understanding of what constitutes a *good essay*. As an example of this, within the groups with only one correctly ordered essay, some were able to generally rank the higher scored essays near the top, while others had the order nearly inverted. To better understand general trends among the groups, the popular choices for best and worst essay were also examined.

Examining which essay groups chose as best immediately presented an interesting finding: 12 of the 15 groups chose the same essay: the score-3 essay. Only one group correctly identified the score-5 essay as best. The remaining two groups split their votes between the score-4 and score-2 essays. No groups selected the score-1 essay as their top choice. Figure 1, below, shows the distribution of votes for best essay.

In general, the groups had much more success at identifying the worst essay, with 12 of the 15 groups correctly selecting the score-1 essay as their choice. Interestingly, the remaining three groups were not close to the correct choice, splitting their votes between the score-5, score-4, and score-3 essays (Figure 1).

FIGURE 1. GROUP VOTES FOR BEST AND WORST ESSAY



On the surface, there is a large disparity between student success at identifying the highest and lowest scored essays. Students' own explanations may better explain this disparity, as well as the general thought process with which they evaluated the essays.

2. Justification for Choices

Reading through the groups' comments justifying their ranking decisions, there are two clear criteria that heavily informed their decisions: (1) the frequency of spelling and grammar mistakes and (2) the simplicity of the essay (considering grammar, vocabulary, and organization). These criteria are quite helpful in explaining the rankings discussed in the preceding section.

By an overwhelming margin, the most frequent considerations mentioned by students were the presence or absence of spelling and grammar mistakes. This may help to explain the success groups had with identifying the score-1 essay. Students justified choosing this essay as their "worst" with comments like the following: "There are a lot of mistakes in terms of grammar, spelling, and capital letter." Indeed, the score-1 essay was rife with errors, and the groups generally were quick to make note of this.

The second criteria by which students seemed to heavily judge the essays was where they perceived the essay to fall on a scale of complexity. To simplify this discussion, comments regarding the simplicity and/or complexity of an essay's grammar, vocabulary, and organization can be discussed together. As mentioned previously, only one group was able to correctly identify the score-5 essay as the best. The reason why this eluded most groups is possibly best revealed in their discussion of the complexity of the essay. Groups frequently made comments on the score-5 essay which were similar to the following examples: "this essay is not easy to read in that this contains difficult example," "bad grammar to understand meaning," or "it is too difficult to understand." These groups struggled to understand the essay, likely due to its advanced vocabulary and grammar, and therefore decided it was not a *good essay*. Only one group praised this complexity, saying "nice vocabs."

A more advanced approach to organization also seemed to trouble the groups. The score-3 essay was regularly praised as having "the best organization so it is easy for us to read," or

"good discourse marker." Students appreciated the overt organization provided by its "firstly," "secondly," "thirdly," "in conclusion" structure. In contrast, the score-5 essay, which relied more on logical argument ordering than on obvious transitional words and phrases, was not as appreciated by students. One group said simply "this essay is not organized."

A possible third criteria, which received attention from some groups, was the content and number of reasons provided in the essays. While there was insufficient data from this study to draw a definite conclusion, it is possible students generally preferred essays with many reasons (and from multiple perspectives), rather than those with few, well-developed reasons, which support a single perspective on the issue. Several groups praised the score-2 and score-1 essays for having "two perspectives" on the issue, even though doing so weakened the author's argument. On the other hand the score-5 essay was criticized for "saying the same thing over and over." Several groups also noted that the author of the score-5 essay says there are many reasons to support his/her opinion, "but there is only one reason." Further investigation may reveal that students' value of having many reasons and perspectives does not match test-makers' expectations, who may place greater value on a persuasive essay that has well-developed reasons, which strongly support a single chosen perspective on an issue.

V. *Conclusion*

This exploratory study did find that students struggled with understanding what a *good essay* looks like. The participants often showed a preference for essays that were "easy to read," which manifested in several ways. In terms of organization, this meant choosing formulaic five-paragraph style essays with overt transitional phrases such as "first of all," "secondly," "on the contrary," "in conclusion," etc. The official score-4 and score-5 essays did not follow a strict formula. The writers trusted in logical ordering rather than reliance on discourse markers or transitional phrases and displayed more creativity in topic transitions and separation of paragraphs.

Some participant groups also marked down essays with complex grammar and highly advanced vocabulary, calling them "too difficult to understand." It could be that these groups lost sight of the goal of the activity, which was to find the objective best essay, and instead chose based on which essays they most enjoyed reading.

Another issue complicating students' understanding of what constitutes a *good essay*, is the fact that there are multiple English proficiency exams being promoted in the Japanese marketplace, and more specifically to students at Japanese universities. Those who want to study abroad will be required to take the TOEFL iBT or IELTS. However, within Japanese society, a high score on tests such as the EIKEN or TOEIC may be beneficial or required for certain jobs or for promotion within a company.

All of these tests differ from each other in certain aspects. Perhaps a relevant example to this discussion is a comparison of the EIKEN writing section with the TOEFL iBT writing section. The EIKEN is a Japan based examination that is separated into multiple levels. Even at the highest level, the EIKEN requires only one essay of a suggested length of 200 - 240 words (EIKEN, 2016, p. 14). In contrast, the corresponding TOEFL iBT section requires an essay of "about 300 words" (ETS, 2012, p. 206), as well as an additional section which integrates a reading and listening exercise into another 150 - 225 word essay response.

Logically, it makes sense that the TOEFL iBT, which is designed as a test of a student's ability to use English in a native speaking environment, would be more stringent than a test like the EIKEN, which is used exclusively within Japanese society. It also shouldn't be ignored that Japanese students are more likely to be exposed to the EIKEN first, which will shape their expectations regarding English essays.

The important takeaway from this for English teachers is that one roadblock to student success on the writing section of tests like the TOEFL iBT may be that students don't properly understand the goal of the writing exercise. Rather than assuming students understand the criteria, the combination of a consciousness raising activity, such as the exercise presented in this study, and specific instruction regarding the rating criteria of the test in question may be a crucial starting point.

Appendix A: Essay Rating Worksheet

Rating Sheet

Organize the essays from best to worst:
Best _____ 2nd _____ 3rd _____ 4th _____ Worst _____

*Thinking of your choice for **best** essay, what did you like about it?*

_____
_____
_____
_____

*Thinking of your choice for **worst** essay, what did you dislike about it?*

_____
_____
_____
_____

Briefly discuss the good and bad points of the remaining essays:
2nd *Good points:* _____
Bad points: _____

3rd *Good points:* _____
Bad points: _____

4th *Good points:* _____
Bad points: _____

References

EIKEN (2016). Grade 1. Retrieved from:
http://www.eiken.or.jp/eiken/exam/grade_1/pdf/201601/2016-1-1ji-1kyu.pdf

ETS. (2012). Official guide to the TOEFL test. New York, NY: McGraw-Hill

ETS. (2016). Test and score data summary for TOEFL iBT tests. Retrieved from:
https://www.ets.org/s/toefl/pdf/94227_unlweb.pdf

Hongo, J. (2013, March 25). Abe want TOEFL to be key exam. *The Japan Times.* Retrieved from:
http://www.japantimes.co.jp/news/2013/03/25/national/abe-wants-toefl-to-be-key-exam/#.VpOA_VR97IX

Ushioda, E. (2013). Foreign language motivation research in Japan: An 'insider' perspective from outside
Japan. In M. Apple, D. Da Silva, & T. Fellner (Eds.), Language learning motivation in Japan
(pp.206-224). Bristol, UK: Multilingual Matters.

Yoshida, R. (2013, April 5). To communicate in English, TOEFL is vital: LDP panel. *The Japan Times.* Retrieved from:
http: //www.japantimes.co.jp/news/2013/04/05/national/to-communicate-in-english-toefl-is-vital-ldp-panel/#.V8esO5grLIV