

- MENU
- PRINTABLE VERSION
- HELP & FAQs

# An Analysis of Testing Methods for English Composition

**Fumie Takakubo**

**Matsuyama School of Nursing**

*Two students were given 30 minutes to write their assignments on single topic on the computer, and the total number of words and numbers of misuse of words or idioms, misspelled and/or mistyped words and grammatical mistakes were counted. Overall evaluations were made according to the criterion-referenced scale-the TOEFL scale for written English, the Test of Written English scoring guide. The topic preference and the topic familiarity were examined by asking each participant to rank*

*the topics. The results suggest the possibility that the number of words and the overall evaluation of assignments depend upon testees' topic preference, and that a great part of topic preference will come from their familiarities with topics. In this sense, direct testing in writing by giving only one topic to testees can cause a method effect in measurement that might limit the generalizability of test scores as indicators for universal scores.*

TOEFLのライティングテスト(TWE)形式で生徒2名が作成した英作文14編について、語数、語彙、スペリング、文法上の誤りを調べ、TWEスコアリングガイドを参考に総合評価を行った。生徒のトピックの好き嫌い、知識があり親しみを持ったトピックであると感じたかは自己評価で順位づけをしてもらい分析した。その結果、生徒が好きなトピックは、知識があり親しみが持てると答えたトピックである場合が多く、好きなトピックであるほど作成された英文は語数が多く総合評価が高い傾向がみられた。以上から、選択肢に複数のトピックではなく1トピックのみ与えるライティングテスト方法では、トピックに対する好き嫌い、知識の多少が評価に影響を与える可能性が示唆された

Information about people's language ability is very useful and necessary on many occasions (Bachman, 1995; Bachman and Palmer, 2000). Within the teaching system, teachers need to obtain information about the achievement of their students in order to assess them and improve their teaching by applying the results to their teaching. To use language tests for these purposes and to make decisions, the quality of the information upon which the decisions are based must be reliable and relevant. Thus, scores from language tests need to be generalizable, and decision makers need to be able to make fair decisions.

The number of applicants for the Test of English as a Foreign Language (TOEFL) in Japan has been more than 100,000 per year, the largest number in Asia (Reedy, 2000). The computer-based TOEFL (CBT) test was introduced in Japan in 2000, and the writing testing method has become a part of the TOEFL test, not a supplement, in the CBT. This means that more weight has been put on the writing section in evaluations of examinees' English proficiency than before. In this study, two students were given writing assignments as a preparation for the Test of Written English (TWE), the TOEFL writing test. The method effect in measurement that may raise questions about the generalizability of results has been focused on and illustrated by presenting their English assignments.

Typical concerns with direct assessments of writing are: (1) the specific effect associated with an interaction between persons and test method facets, for example, individuals differ in their familiarity with certain topics (Bachman, 1995), which may affect the generalizability of results. (2) Assessments of a writing product may not provide information about the writing process, and (3) a brief, timed and single writing sample may not represent an examinee's ability to write in various contexts (Bachman and Palmer, 2000).

Regarding Japanese students, few studies on the specific effect described under (1) in the above paragraph have been reported. Some anxiety in writing, such as 'low self-confidence in writing English' and 'aversiveness of writing in English', may increase when examinees have to write on unfamiliar topics and this might be more obvious in L2 writing than in L1 writing (Cornwell and McKay, 2000). Topic selection would be crucial in obtaining reliable and valid results with Japanese test takers. In this study, effects of Japanese testees' topic familiarity on assessments of writing are focused on, and the author attempts expressing the effects numerically.

## Methods

### Participants

Two Japanese students of a private language school, who have been studying English to prepare for the TWE, participated in this study. Student A (a male) was a middle-school teacher and hoped to study computer technology in the USA. Student B (a female) was fresh from high school and intended to enter a community college in the USA. Their paper based TOEFL scores are in between 440 and 460, and their English levels estimated from the scores would be typical of the first year Japanese university student's level.

### Writing

The writing was conducted based on the TWE, the computer-based TOEFL writing test; that is, the participants were given 30 minutes to write their assignments on single topic on the computer. Topics for writing assignments were selected from the book for TWE (Miyamae and Trevil, 2000) and listed in Table 1 (see Appendix 1). In the text of Miyamae and Trevil (2000), ten topics from various fields compose the basic writing training section for TWE. The targeted writing skill of the section was considered to be suitable for the participants and a topic of the ten was allotted to them as a writing assignment in a lesson. The compositions on seven topics out of ten were completed and submitted by both of the students, which were analyzed in this study.

## Method for analysis

The number of words, the number of misuse of words/idioms, the number of misspelled and/or mistyped words and the number of grammatical mistakes were counted for each assignment. Overall evaluations were made by the author of this report according to the criterion-referenced scale-the TOEFL scale for written English, the TWE scoring guide, which is described in Table 2 (see Appendix 2). The topic preference in writing testing and the topic familiarity were examined by asking each participant to rank the topics.

## Results

### Analysis of the writing assignment

The results are summarized in Table 3 for student A and in Table 4 for student B.

### Word number and topic preference

The average number of words in the assignments was 168 for student A and 218 for student B. The word numbers for topic 5 were much larger than those of other assignments for both students, which greatly contributed to the increase in the average word numbers. One of the reasons for this could be that both students allocated many words to describe 'countryside' and 'big city' in their assignments. Student A told me that he was not in a very good condition when he wrote a composition on topic 2, which might be the reason that this was the shortest among the seven assignments. Table 3 shows that except the composition on topic 4, compositions on topics 5 and 1 which were preferred

over topics 3, 7 and 6, were longer than those on topics 3, 7 and 6. These data suggest that compositions on the topics ranked higher in the students' preference tend to be longer.

### Familiarity of topics and topic preference

The familiarity of topics and the topic preference for writing assignments were closely related. This tendency was demonstrated more clearly by student B's results (Table 4).

### Overall evaluation and familiarity of topics

The overall evaluation tended to coincide with the topic preference or the participants' familiarity with the topics.

## Discussion

Although the number of participants was very small, the results suggest the possibility that the number of words and the overall evaluation of the assignments depend upon students' topic preference. The data also suggest that a great part of topic preference will come from their familiarities with topics. Unlike writers who are evaluated by use of portfolios, TWE test takers are given only one topic to write about in a limited time; therefore when test takers happen to meet a topic which is familiar to them, their scores in writing can be higher than those for unfamiliar topics, and the opposite would also be true. The data in Table 3 and Table 4 are not statistically significant, but they suggest that the numbers of misuses of words/idioms and grammatical mistakes seen in the compositions on the familiar topics tend to be smaller than those for less familiar topics. This may be due to the fact that the participants spent much of their time

**Table 3: Analysis of the writing assignment by student A**

Topic preference	Topic number	Topic	Number of misused words or idioms	Number of misspelled or mistyped words	Number of grammatical mistakes	Number of words	Deviation from the average word number	Overall evaluation	Familiarity with topics
1	No.4	Subjects in high schools	4	6	5	166	+1	3	1
2	No.5	Where to live when you retire	2	0	3	220	+55	3.5	2
3	No.1	Selecting work	7	2	10	172	+17	3	3
4	No.2	Judging a person's character by first impressions	8	2	6	138	-17	2.5	5
5	No.3	How to treat pets	13	3	8	152	-13	2.5	4
6	No.7	Vitamin pills as replacements for vegetables	8	1	9	154	-11	2.5	6
7	No.6	Government intervention in economy	5	0	10	151	-14	2	7

**Table 4: Analysis of the writing assignment by student B**

Topic preference	Topic number	Topic	Number of misused words or idioms	Number of misspelled or mistyped words	Number of grammatical mistakes	Number of words	Deviation from the average word number	Overall evaluation	Familiarity with topics
1	No.2	Judging a person's character by first impressions	8	1	8	224	+6	3	1
2	No.1	Selecting work	5	1	10	213	-5	2.5	4
3	No.4	Subjects in high schools	4	0	8	209	-9	3	2
4	No.3	How to treat pets	11	1	11	211	-7	3	3
5	No.5	Where to live when you retire	10	0	10	257	+39	2.5	5
6	No.6	Vitamin pills as replacements for vegetables	8	0	9	208	-10	2	6
7	No.7	Government intervention in economy	11	1	8	206	-12	2	7



on considering and deciding what to write when they were given unfamiliar topics, which allowed them to have an insufficient amount of time for correcting their mistakes in writing.

The specific effect related to an interaction between Japanese test takers and English writing testing methods cannot be wiped out completely, but it might be avoided by improving the test method, such as giving test takers a right to choose a topic to write about. In this sense, direct testing in writing by giving only one topic to testees can cause a method effect in measurement that might limit the generalizability of test scores as indicators of universal scores.

However, giving a choice of topics to testees might have a risk of generating other problems. For example, some testees may spend much of testing time on choosing a topic to write about, or they may start writing on one topic, abandon it for some reason, try the other topic and complete neither of them. In a study of prompt choice, Polio and Glew (1996) reported that students didn't think they took an inordinate amount of time to choose a topic and they overwhelmingly agreed that there should be a choice of topic on a writing test. These findings are supported by the study of Weigle et al (2000). Weigle et al (2000) also note that topic selection is of less importance if the topics are non-personal, with perceived ability to organize and develop a topic being more important factor. Students' ability in organizing and developing their writing could be increased to some extent by giving them good knowledge and training in writing. What makes testees choose one topic among choices needs to be examined further in subjects at various levels of English writing proficiency.

## Conclusions

The method effect in measurement that may create doubts on the generalizability of results (Spolsky, 1985) can be interpreted to mean that results could lack authenticity, validity and reliability when testers employ inappropriate testing methods. Because of this, testers can fail to demonstrate how test performance reflects language ability.

The analysis of writing assignments presented in this report suggests that there is likely to be a measurement error which varies across individuals who take the test. There are tests which offer testees a choice of topic, such as the test for Cambridge First Certificate in English. TWE was focused on in this study in order to examine effects of Japanese testees' topic familiarity on assessments of writing. The traditional cut scores of TOEFL test are 500-550 for English-medium universities, which is an English level higher than that of the participants. The findings presented in this report may not apply to testees whose scores are closer to this level, and further research will be needed to determine if unfamiliarity of topics leads to weaker writing scores in testees with English writing proficiency at this level. As it is stated in the information bulletin for TOEFL, TWE and Test of Spoken English, the purpose of the TOEFL test is to evaluate the English proficiency of people whose native language is not English. People at various English levels take the TOEFL test aiming at not only fulfilling the language requirement for admissions purposes but also having their English proficiency analyzed objectively. In any case, a measurement error should not exist.

## References

- Bachman, F. (1995). *Fundamental Considerations in Language Testing*. (3rd ed.) Oxford: Oxford University Press.
- Bachman, F. & Palmer, A. (2000). *Language Testing in Practice*. (3rd edn.) Oxford: Oxford University Press.
- Cornwell, S. & McKay, T. (2000). Establishing a valid, reliable measure of writing apprehension. *JALT Journal*, 22: 114-139.
- Miyamae, K. & Trevil, D. (2000). *How to conquer the TOEFL writing test*. Tokyo: Nihonjitugyo Press.
- Polio, C. & Glew, M. (1996). ESL Writing assessment prompts: How students choose. *Journal of Second Language Writing*, 5(1): 35-49.
- Reedy, S. (2000). TOEFL scores in Japan: Much ado about nothing. *The Language Teacher*, 24: 5-11.
- Spolsky, B. (1985). What does it mean to know how to use a language?. *Language Testing* 2, 2: 180-191.
- Weigle, S., Lamison, B. & Peters, K. (2000). Topic selection on a standardized writing assessment. Paper presented at Southeast Regional TESOL, October, Miami, FL

## Appendix 1

**Table 1: Topics for writing assignments**

Topic number	Topic
1	Some people say that if you don't like your work, you must look for something that you like. If you don't like it, you cannot take an interest in it. Others insist that even if you do not like your work, you have to stick to it. Then you will find something interesting in it. Which would you agree with? Use specific reasons and details to support your opinion.
2	Some people believe that they can instinctively judge a person's character. Other people believe that they should take some time in judging it because first impressions are often wrong. Compare these two attitudes. Which attitude do you agree with? Support your choice with specific examples.
3	What do you think of people who treat their dog or cat as a member of their family? Use specific reasons and examples to support your answer.
4	Do you agree or disagree with the following statement? High schools should have required courses for students to balance the subjects taken by their students. Use specific reasons and examples to support your answer.
5	It is better for elderly people to retire in the countryside than in a big city. Do you agree or disagree? Use specific reasons and examples to develop your essay.
6	Some people believe that the government should intervene in the economy of its country. Others believe that the government should not impose any restrictions on the economy. Which one would you support? Use specific reasons and examples to support your answer.
7	Some people believe that taking vitamin pills may replace eating vegetables. Others believe that our body needs real food for its necessary nutrients. Which would you prefer? Use specific reasons and details to support your choice.

(Miyamae and Trevil, 2000)



## Appendix 2

**Table 2: The TOEFL scale for written English**

Scores	Explanations of TWE Scores
6	Clearly demonstrates competence in writing on both the rhetorical and syntactic levels, though the essay may have occasional errors.
5	Demonstrate writing on both the rhetorical and syntactic levels, though the essay will have occasional errors.
4	Demonstrate minimal competence in writing on both the rhetorical and syntactic levels.
3	Demonstrate some developing competence in writing, but the essay remains flawed on either the rhetorical or syntactic level, or both.
2	Suggests incompetence in writing.
1	Demonstrates incompetence in writing.
INR	Did not write an essay.
OFF	Did not write on an assigned topic.

(TWE scoring guide, 1999)