Modern English Education, Vol. 14, No. 3, Fall 2013

Creating Student Profiles Based on the Distractor Analysis of a Vocabulary Test*

Sookyung Cho**
Hankuk University of Foreign Studies
Chanho Park
Korea Institute for Curriculum and Evaluation

Cho, Sookyung & Park, Chanho. (2013). Creating student profiles based on the distractor analysis of a vocabulary test. *Modern English Education, 14*(3), 51-65.

Despite some criticisms, multiple-choice (MC) tests are widely used in vocabulary tests for English language learners. It was pointed out that distractors affect examinees taking an MC-format vocabulary test. Thus, in this study the distractors of a vocabulary test were analyzed to create student profiles on the distractor categories. A 50-item MC test was administered to 2, 206 secondary school students. Then the distractors of the test were classified into three categories: word meaning, graphemic variation/affix, and contextual. Using these categories, the data sets were constructed to analyze the examinee's tendency of choosing distractors when they could not find the correct answer. The results reveal that the patterns to choose three distractor types closely relate to examinees' proficiency: high proficient students are likely to choose more meaning-based distractors but less graphemic variation/affix than low proficient students. By showing the patterns of choosing distractors, a distractor analysis of a vocabulary test will inform the examinees of what to correct and improve in the future as well as what they have mastered, and thus it will contribute to creating a detailed profile for the examinees.

[distractor analysis/vocabulary test/multiple-choice item/
선택지 분석/어휘 시험/선다형 문항**]**

---

* This work was supported by Hankuk University of Foreign Studies Research Fund of 2012.
** First author: Sookyung Cho, Corresponding author: Chanho Park

I. INTRODUCTION

Multiple-choice (MC) tests are still widely used in educational measurement (Haladyna, Downing & Rodriguez, 2002), including the assessment of English language learners' vocabulary knowledge. In spite of its popularity, however, MC tests have also been extensively criticized for their limitations (Read, 2000). One of the criticisms raised by Wesche and Paribakht (1996) is that test items may test learners' knowledge of distractors rather than their knowledge of the target words. By investigating eight different types of distractors in a vocabulary test, Goodrich (1977) pointed out the influence of distractors on the assessment of vocabulary knowledge; in particular, test takers are more likely to be distracted by the options that fit in the given context and those semantically related to the correct answer than the options of similar forms and spellings.

Given the fact that distractors affect test takers' performance in an MC vocabulary test, it is necessary for distractors to be properly examined. If we understand what types of distractors an examinee tends to choose from these distractor categories, we can obtain in-depth and fine-tuned information about the knowledge state of the test taker. In particular, we may find some patterns of choosing distractors by the levels of vocabulary knowledge. Not only the correct answers test takers choose but also their preferred types of distractors will broaden the information about them that can be obtained from the test results.

Methods of distractor analysis have been suggested as item response theory (IRT) models (Thissen & Steinberg, 1984; Wang, 1998). These methods focused more on the functionality of the options and were mostly useful when reviewing the performance of the options. In this study, we develop a method of creating profiles by coding the distractors of an MC vocabulary test, and, through an analysis of distractors, we seek to address the following research questions:

1) Are language learners more likely to choose certain types of distractors than others depending on their proficiency levels?
2) What patterns are discernible across the groups of different vocabulary levels when choosing distractors?
3) What information can be provided to examinees through an analysis of distractors?

II. LITERATURE REVIEW

1. Multiple-Choice Vocabulary Test

Because of its convenience to administer and easiness to analyze the score, the MC format has been widely used when testing language learners' knowledge of vocabulary as well as native speaker's. In spite of its popularity, MC items have not been considered straightforward to interpret as it is claimed to be, for test takers' performance on the test involves factors other than just their knowledge of the target words tested. For instance, by examining the vocabulary section of the Michigan Test of English Language, Perkins and Linnville (1987) found that frequency, number of syllables, and abstractness of the target words well predicted test takers' proficiency levels. That is, frequent, short and concrete words were easier to answer than less frequent, long and abstract words. McQueen (1996) also explored what factors would contribute to the difficulty of MC items in the reading section of an elementary test of Chinese as a foreign language, which was administered in Australia and New Zealand. She found a close relationship between key words in the reading passage and examinee's reading comprehension. In particular, the difficulty of MC items was found to depend on how difficult the word pronunciation was and whether the word characters were presented in Chinese or Romanized letters.

In addition to these factors, distractors in MC items seem to affect test taker's performance. While analyzing several problems with MC vocabulary tests through a meta-analysis of present vocabulary tests, Wesche and Paribakht (1996) pointed out the problems, such as "items may test students' knowledge of distractors rather than their ability to identify an exact meaning of the target word" and "the learner may miss an item either for lack of knowledge of words or lack of understanding of syntax in the distractors" (p. 19). Goodrich (1977) also investigated the effects of distractors on test performance. He examined the effects of eight different types of distractors—false cognate, cloze, antonym, false synonym, graphemic variation, arbitrary, affix, and contextual—and discovered that contextual distractors that relate to the given context as well as semantically-related distractors, such as false synonyms and antonyms, tend to attract test takers from the correct answers in comparison with the other types of distractors.

These effects of distractors as found in the previous studies, however, may not be always detrimental to the test validity but may rather provide an insight on examinee's vocabulary knowledge, if the distractors can be systematically analyzed and controlled. Furthermore, which distractor examinees prefer to the other distractors may reflect their language proficiency levels. According to Talamas, Kroll and Dufour (1999), less fluent learners are more likely to be interfered by form-related words while more fluent learners are prone to be interfered by semantically-related words. Seeing that learners' proficiency levels closely relate to the extent to which they are distracted by meaning- or form-based word interference in the acquisition of second language vocabulary, an analysis of distractors as well as the correct answers will expand the information that can be obtained from the test about each test taker's vocabulary knowledge.

2. Distractor Analysis

Distractors of MC tests are often analyzed only as part of item analysis using frequencies and correlation with the total score (Crocker & Algina, 1986). When checking the frequencies of distractors, the rule of thumb is that all distractors should seem plausible to the examinees who cannot find the correct answer choice and should not seem reasonable to the examinees who are knowledgeable enough to find the correct answer choice (Haladyna et al., 2002). In other words, the frequencies of incorrect answer choices should be neither zero nor exceed that of the correct option. A distractor is not functioning properly if no or very few examinees choose the distractor. If the proportion of the examinees who choose a distractor is larger than that of the examinees who choose the correct option, the item developer needs to check whether the answer key is correctly specified.

For distractor analysis, Thissen and Steinberg (1984) proposed the multiple choice model as an item response theory (IRT) model. The model is similar to the nominal response model suggested by Bock (1972), in which the probability of endorsing each answer choice is expressed as a function of latent traits. By using this model, it can be shown how the endorsement rate of the options of MC test items change as the ability level changes. Ideally, as test taker's ability level increases the probability of endorsing the answer key should increase while the endorsement rate of wrong options should decrease. This model provides an easy way to check how each option of an MC test performs.

Wang (1998) also proposed a model for analyzing distractors of an MC test. This model is a special case of the multiple choice model proposed by Thissen and Steinberg (1984), and the basic difference between the models is that Wang's model is based on the Rasch model. That is, Wang's model assumes that all test items discriminate the latent ability equally while Thissen and Steinberg's model assumes that discrimination is the same across all items.

The models of Thissen and Steinberg (1984) and of Wang (1998) are useful when analyzing how the distractors of an MC test perform, especially when one is interested in how each category is attracting the examinees of different proficiency levels. However, these two models should be used cautiously. These models can be expressed as simple forms using item parameters and are thus called parametric models. These parametric models are relatively easy to apply but require a strong assumption that the model fits the data. In other words, the results produced by these models can be misleading if a wrong model (i.e., a model that does not fit the data) is used.

Similar IRT analysis of distractors can be conducted nonparametrically using TESTGRAF (Ramsay, 2000). TESTGRAF can also provide the results of how the options of an MC item function as the ability level changes. Figure 1 graphically shows the probability of endorsing the options of an MC test depending on the ability level. Although similar plots can be obtained by applying the models by Thissen and Steinberg (1984) and Wang (1998), TESTGRAF is more flexible than parametric models since it does not assume that the model fits the data.

Figure 1 shows that option 0 is dominant at the lowest level while the highest level examinees choose option 2 most, and option 1 is popular among mid-level examinees. It seems that option 2 is the answer key because the probability of endorsement is increasing as the ability level increases. If option 2 is not the answer key, this question is most probably miskeyed and should be revised. Also, although the probability of endorsing option 3 is not larger than that of option 2, option 3 needs to be revised because its probability is also increasing as the ability level increases. By using this graphical information, item developers may understand how each option is functioning across different proficiency levels.

Distractors of MC items have mostly been useful when item developers want to revise the item or want to write a similarly functioning item. Although researchers begin to utilize the information from the distractors of MC items (e.g., Bolt, Wollack & Suh, 2012), no studies attempted to classify examinees based on what distractor types they are likely to choose when they cannot find the answer key. By examining the patterns of choosing distractors, however, distractor analysis is expected to provide additional information with regard to examinees' knowledge levels and help to create their detailed profiles.

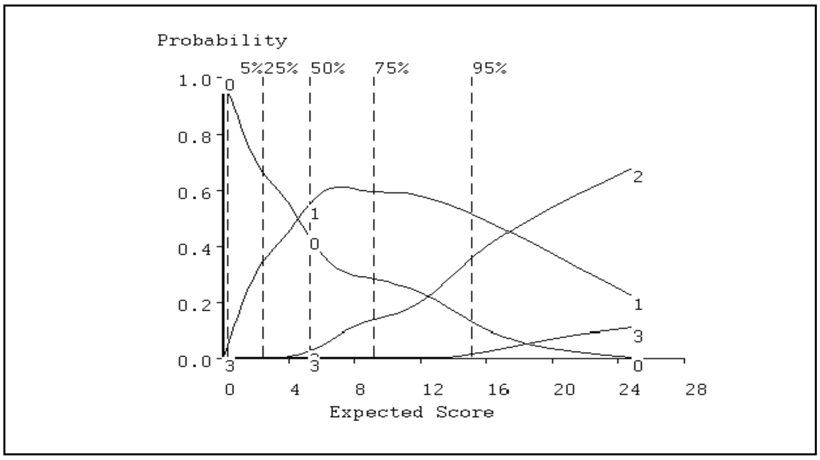

FIGURE 1 TESTGRAF Option Characteristic Curve (Ramsay, 2000)

III. METHOD

1. Participants

2,206 secondary school students participated in this study. The students who gathered in a university in Seoul for another research study took the test. Although the participants came from all over the country, they may not be a nationally representative sample, and the results should be interpreted accordingly. Since it took only 15 minutes to take the test, every participant agreed to participate, and it was made explicit that the test would only be given to volunteers and the students who did not want to participate could leave any time before or during the test. Of the 2,206 subjects, 975 participants (44.20%) were middle school students while 1,231 participants (55.80%) were high school students. Also, 1,315 students (59.61%) were males, and the remaining 891 students (40.39%) were females. No other pieces of background information of the students than their grades and gender were collected from the participants.

2. Materials

1) Instrument

A vocabulary test of 50 MC items was administered to the participants. The test format was adopted from the vocabulary section of the Test of English Proficiency developed by Seoul National University (TEPS); that is, the test had 50 MC items with four options. Of the 50 items the first 25 items were dialogue types, and the last 25 items were monologue types. The participants were asked to find the word of phrase that best fills the gap in one exchange of dialogue or in a sentence. The test items were written by the authors, and sample items are shown in Appendix.

The participants were given 15 minutes to finish the test. Before conducting the distractor analysis, the reliability of the vocabulary test was examined. Cronbach's alpha (Cronbach, 1951), which is an internal consistency measure and a lower bound to reliability, was computed to be .83, which was good and acceptable for a low-stakes examination (Kline, 1999).

Then the distractors were coded according to the categories classified by Goodrich (1977). Although Goodrich suggested eight categories aforementioned, some categories such as false cognate and cloze were hard to apply to the test of English vocabulary for Korean learners of English. Also, since these categories are used to create profiles, the fewest categories possible would maximize the interpretability of the profiles. Some of the remaining categories were combined based on a preliminary analysis. Thus, antonym and false synonym collapsed into one category, and graphemic variation and affix collapsed into one as well. Also, the "arbitrary" category was not used for further analyses because of the difficulty of interpretation. That is, a high or low profile on the inclination to choose arbitrary options may not help make a study plan for the examinee. Thus, the category of "arbitrary" was not considered.

The resulting categories were "word meaning", "graphemic variation/affix", and "contextual." The profiles based on these categories can be interpreted as follows. Examinees would choose "contextual" answer choices if they understand the context but do not exactly know the correct answer. Examinees who do not understand the context but know the meaning of the words would choose "word meaning" choices such as false synonyms and antonyms. Finally, when the examinees who neither know the context nor word meanings may have seen or heard about the vocabulary, they may choose similar-looking answer choices ("graphemic variation/affix"). We hypothesized that advanced learners would be more susceptible to meaning-based options while form-based options would attract low-level examinees more.

Based on the three categories, the answer choices of the items were coded, and the results of coding for the first five items are shown in Table 1. Table 1 shows that among the four options of Item 1, Option 1 is the key, and the three remaining options are all distractors of "word meaning." That is, Options 2 to 4 of Item 1 are false synonyms or antonyms of Option 1. Unlike Items 1 and 2, in which all distractors are of the same type, the distractors of Items 3 through 5 have different categories. For example, Option 3 of Item 3 is the key, and Option 2 is a "word meaning" type. It is noteworthy that some options can be classified into two or more categories. Option 4 of Item 5 is categorized as "word meaning" and "contextual" types simultaneously.

Option categorization was conducted by the two authors separately and independently. After initial categorization, the authors convened and compared the coding. Out of the coding results of the 200 options (i.e., 50 items by 4 options), there were eight options whose categorization on which the authors disagreed with each other. We further discussed the differences and finally agreed on the final coding results.

Three data sets were constructed using the three categories as follows. For the "word meaning" data set, an item was coded 1 when the examinee chose a "word meaning" category for the item; otherwise, it was 0. For example, for an examinee who selected Option 3 for Item 1, Item 1 was coded 0 for the original data set, but was coded 1 for the "word meaning" data set. The other two data sets—"graphemic variation/affix" and "contextual"—were constructed in the same way. As a result, besides the original data set, three additional data sets were created for the three distractor categories. One thing to note is that Item 1 was entirely missing for the "graphemic variation/affix" and "contextual" data sets because no options of Item 1 belonged to those categories.

Table 2 shows the number of items of the original data set (named Vocabulary hereafter) and the three additionally created data sets for distractor analyses. The "word meaning" data had the largest number of items of the three data sets created while the "graphemic variation/affix" data set had only 19 items. The created data sets had lower Cronbach's alpha values ranging from .55 to .70. They were a little low but acceptable considering that they were reconstructed from the data.

TABLE 1
Coding of the Options (First Five Items)

Item	Option	Correct	Word Meaning	Graphemic/Affix	Contextual
1	1	1	0	0	0
	2	0	1	0	0
	3	0	1	0	0
	4	0	1	0	0
2	1	1	0	0	0
	2	0	1	0	0
	3	0	1	0	0
	4	0	1	0	0
3	1	0	0	0	0
	2	0	1	0	0
	3	1	0	0	0
	4	0	0	0	0
4	1	0	0	1	0
	2	1	0	0	0
	3	0	0	1	0
	4	0	0	1	0
5	1	0	1	0	0
	2	0	0	0	0
	3	1	0	0	0
	4	0	1	0	1

TABLE 2
Number of Items and Cronbach's Alpha of the Data Sets

	Vocabulary	Word Meaning	Graphemic/Affix	Contextual
Number of Items	50	44	19	40
Cronbach's Alpha	.83	.65	.70	.55

3. Procedure

First, the Rasch model (Rasch, 1960) was applied to the four data sets of dichotomous items as summarized in Table 2. The Rasch model was used because the latent trait scores of the Rasch model are independent of the items; thus, the effects of unequal numbers of items are minimal with the Rasch model. Examinee profiles can be attempted using the proportions of the three categories for the wrong items. In that case, however, the category of "graphemic variation/affix" will have low profiles due to the small number of items. Also, the sum of the proportions of the three categories may exceed 1.0 because multiple categorization was allowed, and the proportions cannot be interpreted as probabilities.

Since the Rasch model is an IRT model, it requires a constraint for model identification (de Ayala, 2009). In this application, the model was identified so that the latent trait scores had a mean of zero. Once the analyses were completed, all examinees except for one student who answered all 50 items correctly had four latent trait scores for Vocabulary and the three distractor categories. Since each set of scores has a mean of zero across the examinees, the scores can be used as student profiles. Also, the examinees were grouped into three using tertiles of vocabulary score. The three groups were thus named high, intermediate, and low proficiency groups. The three group's mean scores for the three distractor categories were compared.

For the Rasch model analysis, Winsteps software program was used (Linacre, 2009). All other statistical analyses were conducted using R, a free statistical computing system (R Development Core Team, 2010).

VI. RESULTS

First, the latent trait scores on Vocabulary and the three distractor categories obtained from the Rasch model scaling were compared for correlations. Table 3 shows Pearson product-moment correlation coefficients of the four latent trait scores. Since the three distractor categories can have 1's only for the items of incorrect responses, their correlation coefficients with the vocabulary knowledge are close to -1.0 as expected. In other words, as the vocabulary knowledge increased, the tendency to choose each of the three distractor categories—"word meaning," "graphemic variation/affix," and "contextual"—decreased and vice versa. The three distractor categories are highly correlated with each other with the correlation coefficients of about .75.

TABLE 3

Correlation Coefficients among Vocabulary and Three Distractor Categories

	Vocabulary	Word Meaning	Graphemic / Affix	Contextual
Vocabulary	1.00			
Word Meaning	-.90	1.00		
Graphemic / Affix	-.84	.75	1.00	
Contextual	-.83	.77	.74	1.00

Each examinee (except of course for the one examinee with the perfect Vocabulary score) now has one latent trait score for Vocabulary and three scores for the three distractor categories. Table 4 shows a sample of 10 students with their number correct scores on the vocabulary test and latent trait scores under the Rasch model on Vocabulary and the three distractor categories. The student with high number correct scores such as ID # 2 has high scores on Vocabulary and low scores on the other distractor categories. Students with intermediate or low number correct scores have quite different profiles on the three distractor categories.

TABLE 4

Student Scores on Vocabulary and Three Distractor Categories (Sample)

ID	Number Correct	Vocabulary	Word Meaning	Graphemic / Affix	Contextual
1	27	0.69	-0.14	-0.74	-0.69
2	45	3.12	-2.80	-2.99	-3.01
3	10	-1.30	0.60	1.82	0.06
4	41	2.35	-1.90	-2.11	-3.01
5	33	1.33	-1.13	-1.50	-0.69
6	14	-0.73	0.23	0.57	0.74
7	36	1.67	-1.30	-1.79	-1.22
8	39	2.06	-1.68	-1.79	-1.89
9	33	1.33	-1.48	-1.50	-1.22
10	31	1.11	-0.54	-0.99	-0.38

Although Table 4 provides student profiles that can be useful for their learning purposes, the scores are rather hard to interpret due to its arbitrariness of scales. Thus, it was further examined how the scores can turn into meaningful categories. For illustration purposes, the latent trait scores were grouped into three using two tertiles as cut-off scores, which are shown in Table 5. Using the tertiles, the scores as shown in Table 4 can turn into student profiles. Table 6 shows a sample of student profiles on the distractor categories. Although all examinees in this sample are low on Vocabulary, they have different skill profiles on distractors.

TABLE 5

Tertiles on Vocabulary and Three Distractor Categories

	Vocabulary	Word Meaning	Graphemic/Affix	Contextual
1/3	-.37	-.14	-.26	-.09
2/3	.18	.35	.57	.33

TABLE 6

Student Profiles on Vocabulary and Three Distractor Categories (Sample)

Vocabulary	Word Meaning	Graphemic/Affix	Contextual
Low	Intermediate	High	Low
Low	High	High	Intermediate
Low	High	Intermediate	High
Low	Low	Intermediate	High
Low	Low	High	High

It was further analyzed how the three vocabulary-level groups differ on the scores on the distractor categories, and the results are shown in Table 7. The high-level group had the highest mean score on "contextual" and the lowest mean score on "graphemic variation/affix." This means that students with high vocabulary knowledge did not have difficulty with "graphemic variation/affix" distractors while "contextual" distractors were the most difficult distractor for them. On the contrary, students with low vocabulary knowledge were finding the "graphemic variation/affix" category most difficult while "contextual" distractors were relatively easier for them. Although the pattern was not as distinct, intermediate-level groups were similar to low-level groups.

TABLE 7
Mean Scores of Vocabulary Groups on Three Distractor Categories

Vocabulary Group	Word Meaning	Graphemic/Affix	Contextual
High	-.59	-.91	-.49
Intermediate	.18	.21	.12
Low	.53	.89	.45

V. DISCUSSION AND CONCLUSION

In spite of decreasing popularity, MC tests are still widely used in vocabulary tests for English language learners. This study attempted to create student profiles based on the distractor analysis results of an MC vocabulary test. Although the latent trait scores could be used as profiles, they were categorized into three groups using tertiles for easier interpretation of the profile. Only from the sample outputs, it was obvious the examinees had different profiles even with the same total vocabulary score. Also, the scores of the examinees on the three distractor types were compared for the three overall vocabulary level groups.

The distractor analysis using the Rasch model revealed that examinees preferred certain types of distractors to the others depending on their proficiency levels. For example, in comparison with low proficiency students, high proficiency students were more likely to be distracted by the “contextual” type and less distracted by form-based distractors such as the “graphemic variation/affix” type. On the other hand, low proficiency students showed quite the opposite pattern: they chose more form-based distractors but less meaning-based ones than the high proficiency students.

The different patterns of choosing distractors, depending on the examinees’ proficiency levels, confirm the findings of Talamas et al. (1999) that advanced language learners are more distracted at word-meaning levels while low-level learners are distracted at phonemic levels. By showing each examinee’s profile of distractor patterns, the distractor analysis will show the examinees the areas of their weakness and provides more chances of avoiding the same kind of mistakes choosing meaning-based, context-based, or form-based distractors instead of the correct answers.

The systematic analysis of distractors in an MC vocabulary test will increase the information that can be obtained from a test and contribute to creating a new student profile, drawing on their test scores. In this way, distractors do not harm the test validity any more, but they can be utilized positively as a means of showing what types of wrong options examinees choose and what they should be cautious of next time.

A traditional profile of a vocabulary test based on the number of items with correct responses usually shows each examinee what they have mastered and what they have not. On the other hand, the distractor analysis of an MC test will show what they have to pay more attention to—form, meaning, or context—when they learn a new word. In this way, the distractor analysis will shed light on understanding test takers' knowledge.

The results of the distractor analysis can also be useful to item developers. When item writers are working on an MC item, they may choose to use meaning-based distractors or form-based distractors depending on the purpose of the test. It can be inferred that "contextual" distractors would increase the difficulty level of an item while "graphemic variation/affix" distractors may be suitable for less difficult MC items.

One of the limitations of this study is the arbitrariness in grouping the students based on tertiles. In order to create exact and meaningful groups, a standard-setting procedure will be required as used for criterion-referenced evaluation. Because the four data sets were all scaled using the Rasch model, the Bookmark method can easily be implemented as well as the Angoff method (Cizek, Bunch & Koons, 2004). A standard-setting procedure was not adopted in this study because most methods require many panelists to spend quite a few hours (Cizek, 2012; Cizek & Bunch, 2007); however, to maximize the interpretability of the profiles, a standard-setting procedure needs to be implemented. Also, it will be interesting to examine if the student profiles obtained from the distractor analysis are useful in tailoring study plans for the students.

REFERENCES

Bock, R. D. (1972). Estimating item parameters and latent ability when the responses are scored in two or more nominal categories. *Psychometrika, 37*, 29-51.

Bolt, D. M., Wollack, J. A., & Suh, Y. (2012). Application of a multidimensional nested logit model to multiple-choice test items. *Psychometrika, 77*, 339-357.

Cizek, G. J. (Ed.) (2012). *Setting performance standards: Foundations, methods, and innovations*. New York: Routledge.

Cizek, G. J., & Bunch, M. (2007). *Standard setting: A practitioner's guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.

Cizek, G. J., Bunch, M. B., & Koons, H. (2004). Setting performance standards: Contemporary methods. *Educational Measurement: Issues and Practice, 23*(4), 31-50.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Fort Worth, TX: Holt, Rinehart and Winston, Inc.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford.

Goodrich, H. C. (1977). Distractor efficiency in foreign language testing. *TESOL Quarterly, 11*(1), 69-78.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309-334.

Kline, P. (1999). *The handbook of psychological testing* (2nd ed.). London, UK: Routledge.

Linacre, J. M. (2009). *Winsteps* (Version 3.68) [Computer software]. Beaverton, OR: Winsteps.com.

McQueen, J. (1996). Rasch scaling: how valid is it as the basis for content-referenced descriptors of test performance? In G. Wigglesworth & C. Elder (Eds.), *The language testing cycle: From inception to washback* (pp.137-187). *Australian Review of Applied Linguistics*, Series S No. 13. Canberra: ANU.

Perkins, K., & Linnville, S. E. (1987). A construct definition study of a standardized ESL vocabulary test. *Language Testing, 4*(2), 125-141.

Ramsay, J. O. (2000). TestGraf: A program for the graphical analysis of multiple choice test and questionnaire data. Retrieved June 1, 2013, from the World Wide Web: http://www.psych.mcgill.ca/faculty/ramsay/ramsay.html.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Denmark: Danish Institute for Educational Research.

R Development Core Team. (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Read, J. (2000). *Assessing vocabulary.* Cambridge, UK: Cambridge University Press.

Talamas, A., Kroll, J., & Dufour, R. (1999). From form to meaning: Stages in the acquisition of second language vocabulary. *Bilingualism: Language and Cognition, 2*(1), 45-58.

Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika, 49*, 501-519.

Wang, W. C. (1998). Rasch analysis of distractors in multiple-choice items. *Journal of Outcome Measurement, 2*(1), 43-65.

Wesche, M., & Paribakht, T. S. (1996). Assessing second language vocabulary knowledge: Depth & breadth. *Canadian Modern Language Review, 53*(1), 13-40.

APPENDIX

Sample MC Items

1. Dialogue type

A: It's awkward for me to talk about money with my boyfriend.

B: It's just because love and money don't __________ well.

a) make b) spend

c) mix d) speak

KEY: c

2. Monologue type

A tomboy and ________ arts fighter has proved she has beauty as well as brawn.

a) fine b) martial

c) performing d) musical

KEY: b

Examples in: English
Applicable languages: English
Applicable levels: Secondary/Tertiary

Sookyung Cho
Department of English Linguistics
Hankuk University of Foreign Studies
107 Imun-Ro, Dongdaemun-Gu, Seoul 130-791, Korea
Tel: (02) 2173-3194
Email: sookyungcho@hufs.ac.kr

Chanho Park
Division of National English Assessment
Korea Institute for Curriculum and Evaluation
Jeongdong Bldg. 21-15 Jeongdong-gil, Jung-gu, Seoul, 100-784, Korea
Tel: (02) 3704-3613
Email: cpark.irt@gmail.com

Received 15 June 2013
Revised 3 August 2013
Accepted 23 August 2013