A practical examination of the use of free-response questions in computerized adaptive testing

G. Gage Kingsbury
Ronald L. Houser

Portland Public Schools

April 15, 1993

Paper presented to the annual meeting of the American Educational Research Association: Atlanta, Georgia.

A practical examination of the use of free-response questions in computerized adaptive testing

Computerized adaptive testing has been shown to be a valuable alternative to traditional paper-and-pencil testing (Weiss, 1983). Through its use of item response theory (Lord and Novick, 1968; Lord, 1980), adaptive testing adjusts test difficulty to the trait level estimate of the test taker, resulting in reliable trait level estimates with shorter test lengths than paper-and-pencil testing.

To date, almost all adaptive testing has been done using multiple choice questions. While multiple choice questions provide an excellent estimate of a test taker's ability to recognize a correct answer to a question, it may be that a test taker's ability to generate a correct answer to a question represents a different and equally important trait to measure. To measure this trait, it would be necessary to use questions administered in a free-response mode.

Several questions arise in any situation in which one is contemplating a change from one response mode to another, or the use of a combination of response modes. Primary among these are questions concerning the dimensionality of item responses and the fit of an item response model to the item responses.

In addition to these theoretical questions, there are many practical considerations concerning the use of free-response items in an adaptive test. Among these are 1) whether a free-response adaptive testing system could select, present, and score items relatively quickly, 2) whether the test takers could learn to enter there responses from the keyboard accurately, and 3) whether the test takers can respond to free-response questions quickly enough to justify their use.

The present study implements a simplified version of a procedure for free-response adaptive testing proposed by Kingsbury and Houser (1990), and investigates the impact of the use of free-response items along with multiple-choice items in a hybrid (FRMC) adaptive test. The three areas of investigation are 1) whether the use of free-response items in

adaptive testing is practical with the procedures we are using, 2) whether the dimensionality of the student responses changes between free-response and multiple-choice questions, and 3) whether the fit of the item response model differs between free-response and multiple-choice administration.

Free-Response Adaptive Testing

The procedure for free-response adaptive testing suggested by Kingsbury and Houser (1990) consisted of twelve steps which included procedures for the growth and refinement of a free-response testing system. While these procedures are quite desirable in an ongoing testing program, it was decided to simplify the process used in this research study. The modified procedure for constructed-response adaptive testing used here consists of six steps, as follows:

1) Select a set of items that have been previously calibrated in multiple-choice mode that may be used as free-response items by removing the response alternatives. These items should require only short answers of one or two numbers or words.

2) Have content area experts examine the items to identify potential answers that will go into a correct list and an incorrect list. The correct list would consist of all possible correct answers that the experts can generate, and the incorrect list would consist of all of the probable but incorrect answers that the experts can generate. For certain items like computation items, the incorrect list may contain a code indicating that all answers not in the correct list are incorrect.

3) Run each list through a dictionary program (several dictionaries are available in the public domain) to identify synonyms. Create lists of the unique synonyms that were not included in the original lists, and add these synonyms to the correct or incorrect list, as appropriate.

4) Feed the items and the expanded scoring lists into the testing system, and prepare to test.

5) As a student takes the test, s/he enters an answer. The answer is scored using the correct and incorrect lists and a soundex routine to identify misspellings. If the answer appears on either list, the question is scored appropriately and the test continues as any other adaptive test.

6) If the answer does not appear on either list, it is added to a list of potential answers to be categorized at the end of some regular time period. The item is not scored, and the student then receives another question of the same difficulty, and continues the test. (Once the unknown responses are reviewed, the correct and incorrect response lists would be updated. This procedure should result in fewer and fewer unknown answers as testing progresses, and improve the efficiency of the testing procedure as a consequence.)

Method

Item Pools

Content area experts identified 60 mathematics items from a larger pool of multiple-choice items which had been previously calibrated to the one-parameter, logistic item response model (1PL; Rasch, 1960). Items were chosen which covered the entire range of difficulty in the large item pool, and which could be used as constructed-response items by simply removing the response alternatives.

These 60 items were used to create two item pools. In the first item pool, the first 30 items were kept as multiple-choice items, while the second 30 items were converted to constructed-response items. In the second item pool, the first 30 items were converted to constructed response items, while the next 30 items were left as multiple-choice items. This created two 60-item pools which had exactly the same items in the two different response modes. While a 60-item pool is a very small item pool from which to draw an efficient adaptive test, it is quite appropriate for the purposes of this study, which are to study the comparability of measures, not the measurement efficiency of adaptive testing.

Since the difficulty estimates for the items in free response mode were not known, an arbitrary constant of .8 theta units was added to the previously calibrated, multiple-choice difficulty of the items. This was designed to allow the items to appear in approximately the correct position in the adaptive test, to enhance the similarity of our test to a normal adaptive test.

Adaptive Tests

Each student was administered a 30-item adaptive test in mathematics. The entry point for the student was based on the student's grade level, or on a regression estimate based on prior test performance, when available. Items were chosen using Owen's Bayesian item selection procedure (1975). Student's final achievement level estimates were computed

using maximum-likelihood scoring. During the test, administration alternated between free-response items and multiple choice items, so that each test consisted of 15 items in each response mode.

Test Takers

Participants in this study were 384 students enrolled in grades four through twelve in a large, metropolitan school district. These students were enrolled in nine schools in the district which volunteered to participate in this study, and were tested as part of an ongoing adaptive testing program. This program is currently in use in all of the schools in the district for testing purposes determined by each school. Any student who was scheduled to receive a mathematics test in one of these schools received the FRMC test instead of the normal adaptive test.

Analyses

Practical considerations. To identify whether the FRMC was a practical alternative to traditional adaptive testing, two indirect indicators of system performance were collected along with a non-systematic sampling of user reactions. The indirect measures were the time that the student took to respond to each item and the number of free-response items for which the answer had to be added to the undecided list. The non-systematic reaction sampling process consisted of asking the test proctors whether the students had any difficulties taking the test, and whether these difficulties were related to computer problems (ie, slow item presentation) or response factors (ie, difficulty in trying to type in a response).

Dimensionality. To investigate the differences in dimensionality between the two response formats, two approaches were used. First, confirmatory factor analysis using LISREL (Joreskog & Sorbom, 1984) was applied to the sparse data matrix for each item pool to compare the fit of a one-factor model to the fit of a model with one general factor and two factors specific to the response mode of the item. The three factor solution was compared to a model with one general factor and two factors specific to a random half of the items.

Second, a simple scatterplot of two trait level estimates calculated from the items administered in different response modes to the same person (following recalibration) would be created to indicate the strength of the relationship of the trait(s) measured by the two response modalities. The sample size of 384 allows the recalibration of some of the items in the pool in both response modes. Given the sample size, the average number of responses to any one item in any one response mode was approximately 100. To allow calibration

with the 1PL model, only those items with more than 100 responses in each response mode were selected.

Calibration was accomplished using a fixed-parameter calibration procedure which used the final achievement level estimate from only the multiple-choice items as the fixed-parameter. Since we are only attempting to identify differences in this study, this scale will serve as well as any other and the use of the fixed-parameter calibration approach would assure that the final calibrations would be on the same scale, to the extent possible.

If there were no substantial indication of multidimensionality in the data sets, then the last analysis would try to determine whether the 1PL model would be adequate to describe the performance of students on the two different types of items. The mean square fit of the recalibrated item parameters would be calculated for each item in its two response modalities. To the extent that the mean square fit statistic is consistently higher in one of the response modalities, it would indicate that the response model is inadequate to describe performance in both response modes.

Results

Practical considerations

Table 1 shows the mean number of seconds that students took to respond to each question, for both test forms and response modes. From this table, it can be seen that students take slightly longer to respond to free-response questions than to multiple-choice questions (4% longer for Form A, and 17% longer for Form B). To make the results of this analysis more practical, it indicates that we would expect a 50-item examination to be about 5 to 6 minutes longer with free-response than with multiple-choice, if it were similar to Form B.

This relatively small difference in response times between the two response modes may surprise some readers. It should be kept in mind that the type of free-responses considered here are very short (only one or two numbers or words). It should be expected that longer free-responses will increase response time proportionately.

Table 1

Mean, Standard Deviation (SD), and number of

observations for student response times to each item on each test form, for free-response (FR) and multiple-choice (MC) items

Test Form	*Mean*	*SD*	*N*
Form A			
FR	*44.42*	*40.08*	*2787*
MC	*42.70*	*39.63*	*2776*
Form B			
FR	*44.94*	*43.62*	*3079*
MC	*38.10*	*34.45*	*3164*

Our ongoing conversations with test proctors confirmed the results from Table 1. They suggest that, for the most part, students take slightly longer to answer the free-response items. They also suggest that a student would occasionally read a question quickly and skip it. The data that we analyzed indicated that the skip rate was approximately 1 percent of the total sample of item responses.

The number of answers that had to be added to the unknown list was also quite small in this study. Less than .3% of the answers had to be added to the list. While we would like to attribute this to the brilliance of our content area experts, it is more likely that we should attribute it to the fact that over half of the questions in the item pools had numeric answers. For these items, we were able to delineate all of the correct responses, and consider all other answers incorrect. Visual inspection of the responses given to these items revealed no errors caused by this procedure.

Dimensionality

Correlations. In order to perform the confirmatory factor analysis (CFA), we first created an inter-item correlation matrix. Before discussing the CFA results, it is worthwhile to mention the characteristics of the correlation matrix. Table 2 shows the mean correlation in each of three portions of the correlation matrix for each test form. The three portions were correlations between multiple-choice items (MCMC), correlations between free-response items (FRFR), and correlations between multiple-choice and free-response items (FRMC), respectively. It should be noted that this and all subsequent dimensionality analyses used only the 26 items which were identified as having at least 100 valid responses in each response mode.

Table 2

Mean, Standard Deviation (SD), and number of correlations (N) in three portions of the correlation matrix for each test form.

Matrix Portion	*Mean*	*SD*	*N*
Form A			
MCMC	*.118*	*.085*	*91*
FRFR	*.091*	*.119*	*66*
FRMC	*.112*	*.108*	*168*
Form B			
MCMC	*.112*	*.119*	*66*
FRFR	*.117*	*.094*	*91*
FRMC	*.102*	*.105*	*168*

Several trends are evident from Table 2. First, all of the correlations were fairly low. This is normal for inter-item correlations, and is even more likely in an adaptive test setting. Second, differences among the mean correlation in the different portions of the correlation matrix were fairly small and inconsistent. If this were a multi-trait-multi-method matrix, we would say that there was some evidence of convergent validity between the response methods used here.

Confirmatory Factor Analysis. In the CFA of Form A, the Chi-squared value for the fit of the model with one general factor loading on all items was 580.1, with 299 degrees of freedom. The addition of two, response-model-specific factors to create a three-factor model resulted in a Chi-squared value of 471.9, with 273 degrees of freedom. This reduction in the Chi-squared value was significant, but may not be meaningful. When two randomly-designed factors are added to the one-factor model, the resultant model had a Chi-squared value of 466.37, with 273 degrees of freedom.

Very similar results occurred in the CFA of Form B. The one-factor model resulted in a Chi-squared value of 592.9. This was reduced significantly to 512.3 with the addition of the two response-mode specific factors, but was reduced to 513.0 by the addition of two random factors, instead of the response-mode factors.

Item Recalibration. The recalibration of the items within the alternative response modes was accomplished with

the 26 items which had over 100 responses in each response mode. The mean difficulty estimate from the free-response administration was 2.00 on the theta scale, while the mean from the multiple-choice administration was 1.42. The correlation between the calibrations in multiple-choice and free-response modes was .81. The correlation between calibrations is relatively low for the one-parameter, logistic model, but this could be due to the restriction of range in the calibration sample due to the adaptive testing paradigm, or due to item-specific variation in the impact of the choices in multiple-choice mode.

Figure 1 shows the relationship between the free-response and multiple-choice calibrations. It is evident from this figure that there was a strong relationship between the calibrations from the two response modes. This strong relationship does not support the theory that a multidimensional relationship exists between items in the two response modes.

Figure 1

Free-Response and Multiple-Choice Calibrations

```
+-----------------------------------------------------+
          ¦
¦         ¦                                   1          111
¦
      245 ¦                                                  1
+
          ¦                                                 1
¦
          ¦
¦
          ¦                                                11
¦
          ¦                                                 1
¦
    227.5 ¦                                              1
+
 F                                                         ¦
¦
 r            ¦                                        1   1
¦
 e            ¦                                        1   1
¦
 e           ¦                                1                1
¦
      210 ¦                                                  1
+
 R                                                         ¦
¦
 e             ¦                                          211
¦
 s            ¦                                   2  1         1
¦
 p                                                         ¦
¦
 o      192.5 ¦                                             1
+
 n                                                         ¦
¦
 s                                                         ¦
¦
 e                                                         ¦
¦
          ¦
¦
```

```
       175¦
+
          ¦
¦
          ¦
¦
          ¦
¦
          ¦
¦
     157.5¦
+
          ¦
¦
          ¦
¦

+-------------------------------------------------+
         150       170       190       210       230       250
```

Multiple Choice

Model fit

The mean-squared fit indices for the items in the two response modes were quite similar. The mean-square differed by an average of less than 15% between the two response modes. For 15 of the 26 items, the model fit statistic was better for the free-response version of the item. There was no indication that the multiple-choice response mode allowed better model fit than the free-response mode. If anything, there was a tendency for the free-response mode to allow slightly better model fit to the data.

Discussion and Conclusions

It appears to be practical to administer adaptive tests to students in FRMC mode, to the extent that this study was able to answer the question. The cost in testing time of the addition of free-response items was no more than 17% in our samples. If this cost is considered acceptable for the additional flexibility in item styles, it should expand the power of our current adaptive tests.

The two response modes in these tests did not seem to measure different traits involved in student learning, but in another content area or testing context, they may. In that instance, questions concerning the meaning of the different

traits arise, and research will be necessary to determine under which circumstances we wish to measure the individual different traits, or whether a single multidimensional measurement is more desirable.

The items in our tests seem to measure the same achievement trait, regardless of response mode. This is in keeping with some past research (Bennett, Rock, & Wang, 1990). If this turns out to be generally true, future adaptive tests could leave the choice of the response mode to the test giver or even the test taker, if desired, or could consist of any desired blend of questions.

It is possible that a response model more complex than the simple 1PL model may be provide a better fit for the different types of items, even if they both measure the same trait. Past research (Vale & Weiss, 1977) has indicated that free-response and multiple-choice items differ in their information characteristics, even if they measure the same trait. Our study has indicated a tendency for items in free-response mode to have slightly better model fit than the same items in multiple-choice mode. While it wasn't investigated here, it might be possible to add a fixed lower asymptote to the items in multiple-choice mode to improve the model fit, while still not estimating any more parameters.

Finally, the scoring procedure used in this study is a simple correct-incorrect-unknown procedure, but there is nothing preventing the use of a much more sophisticated scoring algorithm, incorporating degrees of partial credit, for differing answers. Beyond this, it should be possible to develop an expert system for scoring more complex responses, eventually freeing us from list keeping and reducing the number of unknown responses to a minimum.

As a result of the addition of free-response questions to adaptive testing technology, we should be able to make the use of adaptive tests even more desirable to test developers, and we may be able to improve the validity and acceptability of the scores obtained from these tests.

References

Bennett, R. E., Rock, D. A., & Wang, M. (1990). Free-response and multiple-choice items: Measures of the same ability? (RR-90-8). Princeton, NJ: ETS.

Joreskog, K. G. & Sorbom, D. (1984). LISREL VI Analysis of structural relationships by the method of maximum likelihood. Mooresville, Indiana: Scientific Software, Inc.

Kingsbury, G. G. (1985). A comparison of item response theory procedures for assessing response dimensionality. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Kingsbury, G. G. & Houser, R. L. (April, 1990). Assessing the utility of item response models: Computerized Adaptive Testing. A paper presented to the annual meeting of the National Council of Measurement in Education, Boston, MA.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Lord, F. M. & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. Journal of the American Statistical Association, 70, 351-356.

Vale, C. D. & Weiss, D. J. (April, 1977). A comparison of information functions of multiple-choice and free-response vocabulary items (RR 77-2). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Weiss, D. J. (Ed.) (1983). New horizons in testing: Latent trait test theory and computerized adaptive testing. New York: Academic Press.