The Impact of Item Formats on Singaporean Students' Performance in the Trends of International Mathematics and Science Study

Kim Koh
Nanyang Technological University

Kadriye Ercikan
University of British Columbia

Berinderjeet Kaur
Nanyang Technological University

In many of the large-scale international assessments (e.g., TIMSS, PISA), both multiple-choice and constructed-response item formats are used to assess student achievement. In terms of measuring learning outcomes, it is widely accepted that multiple-choice items are limited to measuring factual knowledge and simple recall skills. On the contrary, constructed-response items are known as more effective tools for assessing deep understanding of content knowledge and higher-order thinking skills. However, some researchers have shown that multiple-choice and constructed-response items measured the same basic trait or proficiency. Based on the released TIMSS 2003 reports, Singaporean students were among the top performers in both mathematics and science at the 4th and 8th grade levels. But yet little is known about the effects of item formats on the Singaporean students' performance. Are the multiple-choice and constructed-response items measuring the same cognitive and knowledge domains? This study will report the results of the construct comparability of the multiple-choice and constructed-response items in the TIMSS 2003 mathematics achievement test, Singaporean Grade 8 population. The impact of the item formats on student performance will be discussed.

Introduction

Most of the international large-scale achievement assessments (i.e., TIMSS, PISA) have used a mixture of item formats to measure students' achievement in the language arts (reading and writing), mathematics, and sciences. The use of both multiple-choice and constructed-response items in a single test is deemed necessary for capturing important information about examinees. Multiple-choice item is defined as any item in which examinee is required to choose a correct answer from a set of response options (e.g., four or five). On the other hand, constructed-response items refer to any item that requires the examinee to produce a response (e.g., short or extended answer) other than choosing among a list of alternative answers. Constructed-response tasks may require the examinee to give a simple answer, add an arrow to a diagram, write an essay, solve a multistep mathematics problem, draw a graph or diagram, evaluate and critique information or musical performance, or solve a real world problem.

Dating back to the Bloom's knowledge taxonomy, factual and procedural knowledge are typically measured by traditional pencil-and paper test items whereas higher order thinking skills such as analysis, synthesis, and evaluation can only be assessed by open-ended or constructed response assessment tasks. Many have argued that multiple-choice assessments tend to encourage the teaching and learning of discrete facts and decontextualized procedures as well as rote memorization at the expense of deep conceptual understanding and the development of problem-solving skills (e.g., Resnick & Resnick, 1990; Shepard, 1991).

In educational reform, many researchers have advocated for a transformation of the assessment methods to measure academic achievement. Given that constructed response tasks as in most of the authentic assessments require complex cognitive processes and extended problem solving, many believe that the use of such item format can overcome the limitations of multiple-choice format.

Construct validity as evidenced by the comparability of construct between multiple-choice and constructed-response items in a single test or measuring instrument is an important psychometric issue. According to Frederiksen (1984), item format affects the meaning of the test scores by restricting the nature of the content and processes that can be measured. However, research on the construct comparability or equivalence of multiple-choice and constructed-response items has yielded mixed findings. Most of the studies showed that multiple-choice and constructed-response items measured the same latent variable (Bennett, Rock, & Wang, 1991; Bridgeman, 1992; Lukhele, Thissen, & Wainer, 1994; Thissen, Wainer, & Wang, 1994; Perkhounkova, Hoover, & Ankemann, 1997). For example, using the item response models, Lukhele et al. (1994) found that the constructed-response items on the College Board's Advanced Placement exams added little information to those that provided by the multiple-choice items. In other words, when data from constructed-response items are combined with data from multiple-choice items, little new information about the latent variable (e.g., skills or proficiency) being measured is added. In contrast, some researchers found that item format did make a difference when the purpose of the assessment was diagnostic. Constructed-response items were found to provide better information on the students. Using confirmatory factor analysis, the Bennett et al. (1991), Bridgeman and Rock (1993), Ercikan and Schwarz (1995), and Pollack (1997) studies have shown that multiple-choice items loaded on one factor and constructed-response items loaded on a separate factor.

Objectives of Inquiry

Multiple-choice items can measure only static knowledge (Tatsuoko, 1991). For most of the large-scale achievement assessments and international comparative studies of educational system, multiple-choice items are still widely used. This is because constructed-response items are more expensive to score and require a great deal of time from the examinee to answer. In the TIMSS 2003 assessment, constructed-response items made up more than 40 percent of the total assessment time (Martin, 2003). Constructed-response formats are said to be useful for assessing complex and dynamic cognitive processes, identifying students' misconceptions in the process of producing answers, and providing students with the opportunities to apply their problem-solving skills to real-world tasks. If both the multiple-choice and constructed-response formats measure the same construct, then the use of constructed-response items will add little information value to assessment and curriculum. This is contradictory to the belief that constructed-response items can be used to promote students' higher-order thinking and real-world problem solving.

The Singaporean grades 4 and 8 students are the top performers in both the TIMSS 2003 Mathematics and Science assessments. Yet, little is known about the impact of multiple-choice and constructed-response items on Singaporean students' performance. This study aims to answer two important questions: Do the multiple-choice and constructed-response item formats measure different cognitive skills? Did Singaporean students perform better on the constructed-response items?

Methodology

Subjects

Subjects in all 12 booklets of the TIMSS 2003 Mathematics achievement data (Grade 8 population) ranged from 494 to 507.

Items

There are 194 Mathematics items at grade 8. 66 percent of them are multiple-choice items and 34 percent are constructed-response items. The internal consistency of the items is good. The coefficient alpha estimates for the items in each of the booklets ranged from .85 to .94.

Statistical Procedure

Using confirmatory factor analysis, a two-factor model composed of multiple-choice and constructed-response factors was used to examine the relationship of the cognitive skills measured by the two item formats. The two factors were allowed to be intercorrelated, and the items marking a given factor were constrained to load only on that factor. The maximum likelihood estimation method in LISREL was used to estimate the unknown factor loadings from the Pearson covariance matrix. The chi-square/df value and goodness-of-fit indices of the two-factor model were compared to those obtained by fitting a one-factor CFA model (all items loaded on a single factor).

Results and Conclusions

This section will be presented at the conference.