The Dimensionality Analysis of the Specialized Mathematics Open-Ended Questions using The Multidimensionality Model for DIF Framework

Leong Chee Kin,
ckleong@recsam.edu.my

Warabhorn Preechaporn, Teoh Boon Tat
w_preechaporn@recsam.edu.my, boontat_teoh@recsam.edu.my

SEAMEO Regional Centre for Education in Science and Mathematics
Penang, Malaysia.

Abstract: *In this paper, we report the dimensionality investigation of the specialized mathematics open-ended questions using conventional approaches and the multidimensional model for Differential Items Functioning (DIF) framework. The dimensionality of these questions was initially analysed exploratory utilizing the Principal Components Factor Analysis and model fitting in the Confirmatory Factor Analysis. The open-ended questions were verified to be multidimensional using the exploratory analysis method. We then advocated the use of the Poly-SIBTEST's Differential Test Functioning (DTF) analysis, which was based on the multidimensional model for DIF framework. This was to confirm that the open-ended questions were multidimensional and to identify a possible secondary dimension. The possible secondary dimension was hypothesized to assess a distinct cognitive dimension to supplement the dominant ability being measured, namely the formal reasoning ability beside the dominant construct, the general mathematics ability. The DTF analyses revealed that the open-ended questions were multidimensional and the multidimensionality was due to the inclusion of the formal reasoning ability. The exploratory and model fitting analyses also revealed that there could be more than one secondary dimension involved. Further studies should investigate the other possible secondary dimensions. Knowledge of the secondary dimensions assessed by the open-ended questions will have an impact on the way the teachers generally teach. Teachers will need to give attention to the other secondary abilities. Another significant implication is that the single score being normally reported for multidimensional questions test influences the validity of inferences and decisions being made about the examines.*

1. Introduction

Mathematics has served nearly all the branches of the sciences and plays a vital role in the current information technology era. Mathematics is the language of all these technologies. The mastery of mathematics needs to be increased so as to prepare a mathematically literate workforce to deal with all these technologies processes that are align to the development and needs of a nation. However, mastering mathematics is not easy and it requires various abilities. The National Council of Teachers of Mathematics (NCTM) specified mathematical literacy as students having the abilities to solve mathematical problems and to reason mathematically. Mathematical literacy equips students with logical reasoning skill, problem-solving skill, and the ability to think in abstracts ways [15]. Various studies have identified that mathematics tests require students to apply various skills ([1]; [22], [23]; [24]). The various skills required, defined from the multidimensional perspective, means different groups of examinees may have different multidimensional ability distributions due to the various abilities involved. In addition, if test questions are capable of measuring these multiple dimensions, then using any unidimensional scaling procedure may produce item bias ([2]). From this perspective, a test item functioning differentially between two groups is an item measuring a secondary dimension that favours one of the groups after controlling the main dimension that the test is intended to measured ([7]).

Secondary dimensions are additional abilities being assessed to supplement the dominant ability being measured.

Hence, there are two major purposes in this study. Firstly, a specialized mathematics open-ended questions test dimensionality was investigated exploratory using the Factor Analysis (FA) and model fitting analysis. This specialized mathematics is known as Additional Mathematics which is an elective subject offer to students of Form 4 and Form 5 in Malaysia. Secondly, these analyses revealed that the test has a multidimensional structure, then the dimensionality was further investigated using a Differential Item Functioning (DIF) approach, which utilized the framework that multidimensionality causes DIF. This study not only confirmed the multidimensionality structure of the test but also investigated a possible secondary dimension involved in the test using this framework.

2. Problem Statement

Researchers such as [1] and [25] and many others have argued that tests are multidimensional and should be treated as multidimensional. Researches on open-ended format have shown that such format required various abilities (see for e.g. [17]; [22]; [23]; [24]). However, there has been no published report of studies investigating the dimensionality of the newly reviewed Malaysian Certificate of Education Additional Mathematics open-ended questions which is the focus of this study. However, based on the reporting of a single total score or a single grade in reporting students' achievement, it is always assumed that the score reflects only the general mathematics ability. Hence, are these open-ended questions test unidimensional or multidimensional? Do these open-ended questions measure a single ability or multiple abilities? What if the open-ended questions test is indeed multidimensional? Assuming a multidimensional structure is verified, what then are the additional abilities being measured? There are various reasons that cause multidimensionality in a test. Some studies on mathematical questions have suggested that various abilities and cognitive processes are involved. Hence, what would be the possible secondary dimensions?

According to [12] and [9], formal reasoning skill is related to mathematics and necessary for mastering mathematics. [6] and [16] demonstrated that students who are non-formal reasoners experienced difficulties in mathematics. Hence is lacking in the formal reasoning ability a disadvantage in answering these specialized mathematics open-ended questions? If it is, then these open-ended questions will be functioning differentially against this group of examinees who are weak in formal reasoning ability.

Although there could be other abilities required in the open-ended specialized mathematics questions, it is hypothesized in this study that the formal reasoning would play a much prominent role in answering these open-ended questions. As such, if the open-ended questions test has a multidimensional structure, this study will determine whether the secondary dimension involved is the formal reasoning.

3. Research Objectives

The first objective of this study is to investigate the dimensionality of the open-ended questions test. If a multidimensional structure is verified by both the exploratory and confirmatory approaches analyses, the questions' multidimensionality will be reconfirmed with the Poly-SIBTEST's DTF analyses.

Secondly, this study will investigate the possibility of formal reasoning ability as being a possible secondary dimension assessed by the open-ended questions besides the targeted primary dimension, the general mathematics ability.

This study has avoided grave consequences due to mismatch of using a unidimensional approach in analyzing multidimensional structure data. The analysis on the open-ended questions test that has been verified to be multidimensional was conducted using an appropriate multidimensional-based framework. This study used the unique and novel application of DIF procedures which was based on [20] multidimensional model for DIF (MMD) framework as well as [19] multidimensional-based analysis paradigm.

4. Research Questions

This study was guided by the research objectives outlined and which endeavoured to develop a framework that could provide answers to the following questions:

1. Are the Specialised Mathematics open-ended questions multidimensional?
2. Do the Specialised Mathematics open-ended questions function differentially for formal reasoners and non-formal reasoners?

5. Theoretical Framework

The theoretical base for this study comes from [20] MMD and [19] multidimensional-based DIF analysis paradigm. The MMD is a theoretical account of how DIF occurs and it is based on the premise that DIF is produced by multidimensionality. The main construct the test intended to measure is the primary dimension. DIF questions are believed to elicit at least one dimension in addition to the primary dimension ([1]; [11]; [13]; [19]; [20]; [23]). The dimensions that produce DIF are referred to as secondary dimensions. When the primary and secondary dimensions characterize item responses, the data is deemed to be multidimensional. Secondary dimensions that are related to the construct on the test are considered auxiliary and the DIF caused by such secondary dimensions is benign ([19]).

The MMD is based on two assumptions. The first assumption is that the DIF questions elicit at least one secondary dimension, η, in addition to the primary dimension, θ. The second assumption is that a difference exists between the two groups of interest in their conditional distributions on the secondary dimension, η, given a fixed value on the primary dimension, θ. (i.e. $\eta \mid \theta$).

The [19] multidimensional-based DIF analysis paradigm is a two-stage procedure built on the foundation provided by the MMD. The first stage is a substantive analysis where the dimensional structure of the test is evaluated. Based on this structure, the DIF hypotheses are generated. A DIF hypothesis specifies whether a single question or bundle of questions that are designed to measure the primary dimension also measures a secondary dimension, thereby producing group differences. In deciding whether the data contain distinct dimensions, organizing principles are used to identify questions that share certain characteristics.

The second stage is statistically testing the dimensionality-based DIF hypotheses generated in the first stage. Statistical analyses are performed to see whether the organizing principles reveal distinct primary and secondary dimensions across the group under study. The Poly-SIBTEST (for polytomous scored items) is used to test the DIF hypotheses and quantify the size of DIF.

6. Methodology

This study was quantitative in nature and it involved a total of 1917 students in Form Four taking the elective specialized mathematics subject. The data were collected from a total of 29 schools. All the participants involved in this study followed the same Form 4 specialized mathematics curriculum and syllabus. Topics covered and taught were the same for all the participants. The average age of the participants was 16 years and they were of various abilities and socioeconomic status. The participants were from both rural and urban schools of different types of secondary schools.

There were three sets of instruments being used in this study, namely (a) Studied Subtest, (b) Matching Subtest, and (c) Test of Logical Thinking (TOLT). All the instruments were in Bahasa Malaysia language. The instruments, Studied Subtest and Matching Subtest, were used to assess the examinees' specialized mathematics performance. The TOLT was used to assess the examinees' formal reasoning ability.

The Studied Subtest consisted of six open-ended questions that were hypothesized to be multidimensional. The Matching Subtest consisted of 16 multiple choice questions that were hypothesized to be unidimensional, measuring only the general mathematics ability. TOLT was designed by [21] to measure five modes of formal reasoning namely, controlling variables, proportional reasoning, combinatorial reasoning, probabilistic reasoning and correlation reasoning. To ensure that TOLT was suitable for use locally, the Bahasa Malaysia version of TOLT was developed using Brislin's back-translation method ([5]).

The convergent validity of both the Matching Subtest and the Studied Subtest were determined using the Pearson Correlation coefficient. The reliability of the Matching Subtest, Studied Subtest, and TOLT were estimated using the internal consistency method where the Cronbach Alpha was determined for each test.

The data from the Studied Subtest instrument was of the polytomous type and the data from the other instruments was dichotomous. The polytomous data from the Studied Subtest was scored using a marking scheme. The reliability among the raters was determined to ensure that the scoring was consistent, accurate and reliable.

FA was conducted using the SPSS's principal component analysis approach on the Matching Subtest and Studied Subtest to determine their dimensionality. Confirmatory Factor Analysis (CFA) was also conducted using both PRELIS and LISREL on the hypothesized multidimensional Studied Subtest. PRELIS was used to overcome problems caused by the use of categorical variables and ordinal variables. CFA was conducted to examine how well the hypothesized models fit the data.

The analysis using Poly-SIBTEST was performed to test the hypothesis that the open-ended questions function differentially against non-formal reasoners. The analysis required the examinees to be grouped into the Reference and Focal groups. The examinees' performance on the TOLT was used to determine the grouping of the examinees based on whether they were formal reasoners or otherwise. The Reference Group was those examinees who scored a total of seven, eight, nine or ten on the TOLT. The examinees in the Reference Group were considered as formal reasoners. The Focal Group was those examinees who scored a total of zero, one, two or three. The examinees in the Focal Group were considered the non-formal reasoners. There were 710 non-formal reasoners and 767 formal reasoners.

7. Results

Both the Matching Subtest and Studied Subtest obtained a convergent validity coefficient of .91 and .88 respectively. The consistency reliability, Cronbach Alpha of the Matching Subtest, Studied Subtest, and TOLT were .81, .83, and .84 respectively. An average interraters coefficient of .99 was obtained for the Studied Subtest.

FA results showed that Matching Subtest had a dominant one-factor solution. However, the results of FA on the Studied Subtest revealed that there were two components having eigenvalues greater than one. The first component had 55.41% of total explained variance and the first two components accounted for a cumulative total explained variance of 73.69%. Although the third component had an eigenvalue of 0.71, which was less than one, its percentage of total explained variance was 11.88%, which was quite substantive. Similarly the examination of the scree plot suggested that Matching Subtest had a single dominant factor and Studied Subtest had three factors.

Table 7.1 shows that the χ^2 goodness-of-fit statistics is 1248.56 with 209 degrees of freedom for the one-dimensional model. However, this statistics dropped to 916.28 with 202 degrees of freedom for the two-dimensional model. As the two-dimensional model was nested within the one-dimensional model, these fit statistics and their corresponding degree of freedom could be subtracted to compare model fit using the χ^2 difference test [10]. Hence, the χ^2_{diff} was 332.28 with 7 degrees of freedom. This was a highly significant improvement in model fit because the critical value for χ^2 with 7 degrees of freedom at the significance level of 0.001 was 24.32.

Table 7.1 also presents the other goodness-of-fit measures. For the one-dimensional model, the Comparative Fit Index (CFI), Goodness-of-Fit Index (GFI), and Adjusted Goodness-of-Fit Index (AGFI) (.94, .98, .97 respectively) showed good fit to the data. However, the chi-square to degrees of freedom ratio, which was greater than 2, did not support the fit. The larger Akaike Information Criterion (AIC) (1336.56 > 506.00) yielded in the one-dimensional model relative to the saturated model suggested the possibility of improvement. The Root Mean Square Error of Approximation (RMSEA) indicated a reasonable fit but not a good fit.

For the two-dimensional model, the CFI, GFI, AGFI (.96, .98, .98 respectively) and RMSEA (.04) indicated a very good fit to the data. However, the chi-square to degrees of freedom ratio was also greater than 2 but smaller than the obtained value for the one-dimensional model. Similarly, the AIC was still large but smaller when compared to the AIC value in the one-dimensional model (506.00< 1018.28 < 1336.56). This result suggests that the two-dimensional model is a more appropriate representation of the structure of the Studied Subtest but with the possibility of improvement

Table 7.1
Comparison of Model Fit

Model	Measure of Fit							
	χ2	df	CFI	GFI	AGFI	RMSEA	AIC-model[a]	AIC-saturated[b]
One-dimensional	1248.56	209	.94	.98	.97	.05	1336.56	506.00
Two-dimensional l	916.28	202	.96	.98	.98	.04	1018.28	506.00

Note. model[a]- AIC value for model as tested
model[b]- AIC value for saturated model

The Poly-SIBTEST analysis indicated that the Studied Subtest functioned differentially against the focal group, which are the non-formal reasoners. The β statistics was 1.73 which

indicated an extreme degree of DTF. The converted SIB index was 2.81, which was also beyond the critical value of 1.96. This indicated that the Studied Subtest also favored the formal reasoners group.

8. Discussion and Conclusion

This study found that the open-ended questions of the Studied Subtest are multidimensional. FA has shown that the questions in the Studied Subtest have more than one-factor solution. There were two components having eigenvalue greater than one and three components having more than 10% total explained variance. Scree plot also revealed that there were three underlying factors. Also, the CFA has shown that these questions best modeled after the two-dimensional model structure (χ^2_{diff} = 332.28 with 7 degrees of freedom). The two-dimensional model suggested improvement in the CFI, AGFI, RMSEA and AIC indexes when compare to the one-dimensional model. Particularly, the RMSEA of the two-dimensional model indicated a very good fit as compared to the one-dimensional model where its RMSEA only indicated a reasonable fit. Also, the AIC, indicated the two-dimensional model was a more appropriate model.

The Poly-SIBTEST DTF analyses also verified the Studied Subtest questions were multidimensional. The result indicated that the open-ended questions did assess secondary dimension, and in this case the formal reasoning ability. The Poly-SIBTEST analysis revealed that the $|\beta|$ statistics was 1.73 indicated extreme large degree of DTF. This means the Studied Subtest questions do assess the formal reasoning skill as the secondary dimension and hence its multidimensionality.

The finding that the Studied Subtest test is multidimensional aligns with many researchers' argument on test's dimensionality. [2], [18], and [25] have contended that test is almost always multidimensional. In fact, many researchers agree that educational test data do not always satisfy the unidimensional assumption. On the mathematics test, [17] who have carried out the dimensionality confirmatory test, have also concluded that the mathematics test is not unidimensional in general. Studies by [23] and [24], on Washington Assessment of Student Learning's mathematics suggested that the mathematics test should be treated as multidimensional.

In Malaysia, the revised specialized mathematics' curriculum and syllabus, has provided a fairly multidimensional view of what is to be taught and learned. The assessment, especially in developing questions, must adhere to the curriculum objectives that were designed to meet these objectives. Hence, specifically, the specialized mathematics open-ended questions were intentionally multidimensional to meet this curriculum's objectives and standards.

It has also been shown that different formats in assessment induce different approaches to problem solving. The open-ended format induced more cognitive strategy usage. [8], who elaborated on the features of open-ended questions format found that these questions type required students to apply their reasoning skill. This implied that such a question format is capable of assessing more than one skill. The results of this study concurred with this line of thinking.

The CFA results indicated that the two-dimensional model was more appropriate to describe the data. However, the large chi-square to degree of freedom ratio (> 2) and the large AIC relative to the saturated model obtained by the two-dimensional model (1018.28 > 506.00) suggested that there is possibility to improve further the two-dimensional model fitting. This could be explained as the FA analysis and scree plot showed a significant involvement of the three factors.

The multidimensionality finding does support earlier findings by researchers like [12] and [14] that one variable associated with mathematics is the reasoning ability. [4] and [12] identified that formal reasoning as the essential ability for success in school mathematics. [3] in fact

discovered in their study that mathematics scores were strongly associated with students' level of reasoning. All these substantiates that the Studied Subtest questions favoured the formal reasoners as the test involved the use of formal reasoning for solving the mathematical problems.

9. Summary and implications for educational practice

This study found that the specialized mathematics open-ended questions were multidimensional and benefited those students who were proficient in formal reasoning. The FA and CFA analyses also revealed that there is more than one secondary dimension involved.

Thus, teachers of the specialized mathematics as well as students should be cognizant of the importance of the secondary dimension to enhance the students' performance and understanding in the specialized mathematics. They need to realize that the additional abilities are different yet relevant to answer these questions. Teaching this specialized mathematics should also emphasize the additional secondary skill. Students should be guided on reasoning mathematically.

The attempt to conclude the multidimensional test through one single score is misleading. It fails to depict the examinees proficiency on the two discerning cognitive activities. The mismatch between the construct's dimensions and its scoring can affect the validity of inferences made based on such test scores. To reflect the examinees' proficiency more precisely, one needs to take the dimensionality into account. One solution is to reevaluate the way these questions are scored.

This study shows the application of an appropriate multidimensional DTF analysis approach on the multidimensional data. Ironically, however this is still lacking in local research. Normally, a unidimensional approach is used and certain assumptions are fulfilled to enable analysis to be applied. Hence, the validity of such study is questionable. This DTF approach could easily be applied to other situations such as researching problem solving in mathematics.

Future research should investigate the other possible secondary dimensions involved in these open-ended questions as the FA and CFA results point to more than formal reasoning ability being involved besides the general mathematics ability.

References

[1] Abedi, J. (1997). *Dimensionality of NAEP Subscale Scores*. University of California, Los Angeles.

[2] Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensionality perspective. *Journal of Educational Measurement, 29,* 67-91.

[3] Berenson, S. B., Carter, G., & Norwood, K. S. (1992). The at-risk student in college developmental algebra. *School Science and Mathematics, 92*, 55-58.

[4] Bitner, B. L. (1991). Formal operational reasoning modes: Predictors of critical thinking abilities and grades assigned by teachers in science and matheatics for students in grades nine through twelve. *Journal of Research in Science Teaching, 28*, 275-285.

[5] Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology, 1*(3), 185-216.

[6] Bunce, D. M., & Hutchinson, K. D. (1993). The use of the GALT (Group Assessment of Logical Thinking) as a predictor of academic success in college chemistry. *Journal of Chemical Education, 70*(3), 183-187.

[7] Camilli, G., & Shepard, L. A. (1994). *Methods for identfying biased test items*. Thousand Oaks, CA: Sage Publications
[8] Foong, P. Y. (2002). *Using short open-ended mathematics questions to promote thinking and understanding.* Paper presented at the Conferences of 21st Century Project, Palermo, Italy.
[9] Genovese, J. E. C. (2003). Piaget, pedagogy, and evolutionary psychology. *Evolutionary Psychology, 1*, 27-137.
[10] Kline, R. B. (1998). *Principles and practice of structural equation modeling*. New York, NY: The Guilford Press.
[11] Kok, F. (1988). Item bias and test multidimensionality. In R. Langeheine & J. Rost (Eds.), *Latent Trait and Latent Class Models* (p. 263-275). New York: Plenum.
[12] Lawson, A. E. (1982). Formal reasoning, achievement and intelligence: An issue of importance. *Science Education, 66*(1), 77-83.
[13] Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum.
[14] Mills, C. J., Ablard, K. E., & Stumpf, H. (1993). Gender differences in academically talented young students' mathematical reasoning: Patterns across age and sub-skills. *Journal of Educational Psychology, 85*(2), 340-346.
[15] National Council of Teachers of Mathematics. (1996). *Curriculum and Evaluation Standards for School Mathematics.* NCTM, Inc., USA.
[16] Niaz, M. (1989). Translation of algebraic equations and its relation to formal operational reasoning. *Journal of Research in Science Teaching, 26*, 785-793.
[17] Perkhounkova, Y., & Dunbar, S. B. (1999). *Influence of item content and format on the dimensionality of tests combining multiple-choice and open-response item: An application on the Poly-DIMTEST procedure*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.
[18] Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1988). Building a unidimensional test using multidimensional items. *Journal of Educational Measurement, 25,* 193-203.
[19] Roussos, L. A., & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20,* 355 – 371.
[20] Shealy, R. & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58,* 159-194.
[21] Tobin, K. G., *& Capie, W. (198*1). The Test of Logical Thinking. *Journal of Science and Mathematics Education in S. E. Asia, VII*(1), 5-9.
[22] Walk*er, C. M., & Beretvas, S. N. (2000). Using multidimensional ability estimates to determine student proficiency in mathematics.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.
[23] Walker, C. M., & Beretvas, S. N. (2001). An empirical investigation demonstrating the multidimensional DIF paradigm: A cognitive Explanaton for DIF. *Journal of Educational Measurement, 38,* (2), 147-163.
[24] Wu, B. C. (2003, October). *Reliability Estimation under a Multidimensional Structure.* Paper presented at the 29th IAEA Conference, Manchester.
[25] Yen, W. M. (1985). Increasing item complexity: A possible cause of scale shinkage for unidimensional item response theory. *Psychometrika, 50,* 399–410.