Rasch Modeling of the Test of Early Mathematics Ability–Third Edition with a Sample of K1 Children in Singapore

Shih-Ying Yao, David Munez, Rebecca Bull, Kerry Lee, Kiat Hui Khng, and Kenneth Poon

National Institute of Education, Nanyang Technological University

Author Note

Shih-Ying Yao, David Munez, Rebecca Bull, Kerry Lee, Kiat Hui Khng, and Kenneth Poon, Education and Cognitive Development Lab, National Institute of Education, Singapore.

This study was supported by a grant from the Office of Education Research (grant number: OER 09/14 RB). Views expressed in this article do not necessarily reflect those of the National Institute of Education. The authors wish to thank all of the children, teachers, and principals who participated in this research. The authors are also grateful to research staff from the Singapore Kindergarten Impact Project who collected the data for this study.

Correspondence concerning this article should be addressed to Shih-Ying Yao, National Institute of Education, 1 Nanyang Walk, Singapore 637616. Email:shihying.yao@nie.edu.sg.

Rasch Modeling of the Test of Early Mathematics Ability–Third Edition with a Sample of K1 Children in Singapore

Abstract

The Test of Early Mathematics Ability – Third Edition (TEMA-3) is a commonly used measure of early mathematics knowledge for children aged 3 years to 8 years 11 months. In spite of its wide use, research on the psychometric properties of TEMA-3 remains limited. This study applied the Rasch model to investigate the psychometric properties of TEMA-3 from three aspects: technical qualities, internal structure, and convergent evidence. Data were collected from 971 K1 children in Singapore. Item fit statistics suggested a reasonable model-data fit. The TEMA-3 items were found to demonstrate generally good technical qualities, interpretable internal structure, and reasonable convergent evidence. Implications for test development, test use, and future research are further discussed.

Keywords: Test of Early Mathematics Ability, psychometrics, Rasch model, mathematics assessment

Rasch Modeling of the Test of Early Mathematics Ability–Third Edition with a Sample of K1 Children in Singapore

The acquisition of early numerical abilities, such as understanding the rules of counting, quantity comparison, and representing numerical magnitudes in the form of a number line, has been found to provide the foundation for the early learning of formal mathematics in school (e.g., Duncan et al., 2007; Fuchs et al., 2010; Jordan, Kaplan, Ramineni, & Locuniak, 2009; Martin, Cirino, Sharp, & Barnes, 2014; Passolunghi, Lanfranchi, Altoe, & Sollazzo, 2015; Watts, Duncan, Siegler, & Davis-Kean, 2014). Given the importance of early numerical skills to continuing mathematics achievement, it is critical to have good measurement tools to help identify children who are struggling at this early stage. Of the few summative assessments in mathematics appropriate for young children, the Test of Early Mathematics Ability– Third Edition[1] (TEMA-3; Ginsburg & Baroody, 2003) is a widely used measure of mathematics knowledge for children aged 3 years to 8 years 11 months. Researchers and practitioners have commonly used TEMA-3 to understand learning disabilities (e.g., Murphy, Mazzocco, Hanich, & Early, 2007), to document children's progress in mathematics knowledge, and to compare children's mathematics knowledge across countries (e.g., Ryoo et al., 2014).

TEMA-3 is conceptualized to assess two forms of mathematics knowledge: informal (acquired outside the context of schooling) and formal (skills and concepts learned in school). TEMA-3 assesses informal mathematics knowledge through four categories of items: (1) numbering (e.g., verbal counting by ones), (2) number comparisons (e.g., choosing the larger number), (3) calculation (e.g., addition of concrete objects), and (4) concepts (e.g., number constancy). Formal mathematics knowledge is also assessed with four categories: (1) numeral

[1] Please refer to the TEMA-3 manual (Ginsburg & Baroody, 2003) for a detailed account of changes in TEMA across versions.

literacy (e.g., reading or writing numerals), (2) number facts (e.g., subtraction facts), (3) calculation (e.g., written addition accuracy), and (4) concepts (e.g., written representation of sets). Each category contains a number of items assessing corresponding concepts or skills from basic to advanced level. The TEMA-3 yields a standard score (i.e., Math Ability Score), which is converted from the raw score and provides an indication of a child's overall mathematics ability. Note that the authors of TEMA-3 (Ginsburg & Baroody, 2003) do not validate sub-constructs and hence sub-scores are not available.

In spite of its wide use, there exists little research on the properties of TEMA-3. In developing TEMA-3, Ginsburg and Baroody (2003) investigated the validity and reliability of the measure with content analysis and conventional item analysis using a U.S. sample. They reported that TEMA-3 evidenced a high degree of internal consistency, alternate-form, and test-retest reliability. Also, TEMA-3 was found to demonstrate good content, criterion, and construct validity. Ryoo and colleagues (2015) further examined the factor structure of the TEMA-3 using longitudinal data from a U.S. sample. Their finding suggested the potential of identifying subscales within TEMA-3 to understand children's knowledge of specific mathematics concepts. There are several limitations to these existing studies. First, the psychometric properties of TEMA-3 have not been studied with contemporary measurement models. Conventional indices used in previous item analysis (Ginsburg & Baroody, 2003), such as using percentage of children succeeding on an item as an indication of item difficulty, are known to be sample dependent (Daniel, 1999). Second, the administration of TEMA-3 adopts basal-and-ceiling rules. Following the TEMA-3 manual, past research (Ginsburg & Baroody, 2003; Ryoo et al., 2015) scored all items below the basal correct and all items above the ceiling incorrect. Although such scoring practice is conventional, filling missing item responses with the expected item scores poses

threats to the investigation of the item properties (Daniel, 1999). Specifically, if items are not well ordered in difficulty, the expected item scores filled in based on the basal-and-ceiling rule will not be a proper reflection of what a child may have performed, which in turn will mask the true properties of the items. Third, to our knowledge, the properties of TEMA-3 were only investigated with U.S. samples to date. Considering the wide use of TEMA-3 as an indication of children's mathematics ability, more studies are needed to examine its validity and reliability.

This study applied the Rasch model to investigate the psychometric properties of the TEMA-3 items with a sample of children at first year of kindergarten in Singapore. The Rasch model was selected for the following reasons. First, presence of missing data does not impede the concurrent estimation of item and person parameters in the Rasch analysis. Hence, the investigation of item properties can be based on actual item responses (e.g., missing data can be retained for non-administered items resulting from the basal-and-ceiling rule in test administration)[2]. Second, if data fit the Rasch model well, the resulting measurement scale has several desired features. Specifically, the Rasch model provides a full model of response behaviors as separate parameters are estimated for both persons and items. Person and item parameters are calibrated onto a common interval scale through the Rasch analysis, which allows direct and invariant comparisons of person abilities and item difficulties (see Embretson & Reise, 2000 for details). This study investigated the properties of the TEMA-3 items through the following aspects:

[2] Mislevy and Wu (1996) suggested that the missingness caused by the selection of items to administer to an examinee based on the observed responses to previous items in adaptive testing (as the case with TEMA-3 in this study) is ignorable for ability estimation. As to item calibration, the key is the amount of data available per item. See analysis section for a detailed discussion.

1. Technical qualities: Did the TEMA-3 data fit the model acceptably such that the reliability and validity inference drawn from the Rasch scale are supported? How well did each of the items discriminate children with respect to their mathematics ability?
2. Internal structure: To what extent was the empirical pattern of item difficulty consistent with the hypothesis of item design? Specifically, were items designed to assess advanced mathematics knowledge more difficult than items designed to assess basic knowledge? Also, how was item ordering (with respect to difficulty) determined with the Singapore sample related to item ordering in the TEMA-3 test form determined with the U.S. sample?
3. Convergent evidence: What was the relationship between children's performance on TEMA-3 and Number Sets Test (Geary, Bailey, & Hoard, 2009), which is another measure of early numerical ability?

Method

Participants

The data used in this research was from a longitudinal study examining the impact of kindergarten on children's development. The sample had 971 children (494 girls), which comprised 564 Chinese, 191 Indian, 120 Malay, and 42 who identified with other ethnic groups. No valid ethnicity information was available for 54 children. Parental consent and child assent were received for these children prior to data collection. All children were tested between February and September in their first year of kindergarten (average age=56.99 months, S.D. = 3.78 months).

Instruments

The TEMA-3 and Number Sets Test were administered to the participating children at their kindergartens as part of a larger task set by trained research assistants on a one-to-one basis.

TEMA-3. The test was normed in U.S. and has two parallel forms. In this study, only Form A, comprising 72 items, was administered. Each item was scored dichotomously (i.e., 0 or 1). Test administration began with an entry point suitable for a child's age and was terminated when ceiling and basal were established (Ginsburg & Baroody, 2003). Raw data were used in the subsequent Rasch analysis. For example, missing data on the non-administered items were retained and not scored with the basal-and-ceiling rule.

Number Sets Test. This test assesses the speed and accuracy with which children can identify and process quantities represented by Arabic numerals and object sets in a paper-and-pencil format. Pairs or trios of objects (e.g., [●●●|♦♦]), Arabic numerals (e.g., [3|2]), or both (e.g., [♦♦♦|2]) were combined to create domino-like rectangles in the test form. Each combination is considered an item. Children were asked to circle items that matched a target number (5 or 9) as quickly and accurately as possible (Geary et al., 2009).

The Rasch Model

The Rasch model (Rasch, 1960) formulates the probability of person *n* succeeding on item *i* (P_{ni}) as follows:

$$P_{ni} = \frac{\exp(\theta_n - \delta_i)}{1 + \exp(\theta_n - \delta_i)} \quad (1)$$

where θ_n denotes the ability of person *n*, and δ_i denotes the difficulty of item *i*. The weighted mean square statistic (Wright & Masters, 1982), known as the infit statistic, was used to evaluate how well the item data fit the Rasch model. The infit statistic has an expected value of one. Items

with infit statistic outside the range of 0.75 and 1.33 and with the absolute value of the weighted *t* statistic larger than 1.96 required further investigation (Wilson, 2005).

Analysis

As the TEMA-3 targets children up to 8 years old and the current sample comprised K1 children only, test administration terminated before many of the children reached the later items due to the implementation of the basal-and-ceiling rule. As the entry point is item 15 for our sample, items at the beginning of the test were taken by fewer children than items at the middle.

Past research (e.g., Linacre, 1994; Chen et al., 2014) has discussed the impact of small sample sizes on the Rasch analysis, such as larger standard errors, less robust parameter estimates, and less powerful fit analysis. Considering the impact of sample size on item calibration, the analysis procedure was multi-step. In step 1, only items that had at least 250 observations and at least 10 observations per response category were used for the Rasch analysis (i.e., items 8 through 41). Calibrations of these 34 items were expected to be reasonably robust and precise. Past research (e.g., Linacre, 1994; Chen et al., 2014) has suggested that the Rasch analysis can still be performed for items with smaller sample sizes for exploratory purposes, in spite of the aforementioned impact of small sample sizes on calibration. Thus, in step 2, the Rasch model was fit to all TEMA-3 items that did not have null categories[3]. In this calibration, item difficulty parameters obtained from step 1 were imported as anchors. The result of this second calibration showed that items located at the end of the test (i.e., items 53 through 65 and item 72) displayed large infit statistics. These items all had very small sample sizes and fewer than 8 observations per response category. In step 3, the problematic items identified in step 2

[3] 6 items (i.e., items 66 through 71) had null categories (i.e., no observations in either correct or incorrect category). Item parameters could not be identified for items with null categories.

were excluded. A total of 52 items (i.e., items 1 through 52) were retained in the final Rasch analysis, where the item parameters from step 1 were used as anchors.

The Rasch model was estimated with the software ConQuest (Wu, Adams, Wilson, & Haldane, 2007) using the marginal maximum likelihood estimator, where the ability parameter θ_n is assumed to be normally distributed as $N(\mu_\theta, \sigma_\theta^2)$. For model identification, μ_θ was constrained to 0 in step 1. As anchors were used in subsequent steps, no additional model constraints were needed (Briggs & Wilson, 2003).

We also performed the Rasch analysis using data coded with the basal-and-ceiling rule for exploratory purpose. The model could not converge properly. The convergence issue still remained after we re-ran the analysis using the parameter estimates from the previous calibration as initial values and also increasing the nodes used in the numerical integration as suggested by ConQuest. These results demonstrated the problem of using data recoded with the basal-and-ceiling rule in item analysis and supported the use of raw data in the current study.

Results

1. Technical Qualities

Reliability. The expected a posterior (EAP) reliability (Bock & Aitkin, 1981) was found to be 0.93. This high person reliability suggests that the TEMA-3 items differentiate the K1 children with respect to their mathematics ability well. In addition, the item separation reliability (Wright & Stone, 1979) was found to be 0.997. This high item separation reliability suggests a nice spread of items along the continuum of the latent construct.

Item fit. As shown in Table 1, only item 27 had the infit statistic falling outside the range of 0.75 and 1.33, and a *t* statistic larger than 1.96. Figure 1 presents the item characteristics curve of item 27. It shows that the empirical impact of changes in ability level on probabilities of

succeeding on this item was smaller than the model prediction, i.e., this item contributed less to the measurement of the latent mathematics ability compared with other items. Item 27 asked children to compare several number sets, each of which contained 3 numbers (e.g., Here is an 8. Which is closer, 1 or 6?). The 3 numbers were presented as a triangle, with the target number (8) at the top and the two other numbers (1 and 6) at the bottom. The greater variability in children's response to this item might be due to some children misunderstanding "closer" as referring to their physical distance as presented in the stimulus rather than a mental number line as intended in item design. As the infit statistic of item 27 still fell within the range of 0.5 and 1.5 and was considered as productive for measurement following Wright and Linacre (1994), all 52 items were retained.

Item discrimination. Point-measure correlation is a good substitute to evaluate item discrimination, particularly when data are missing (Kelley, Ebel, & Linacre, 2002), which is the Pearson correlation between the ability measures calibrated from the Rasch model and scores on a particular item. Table 1 shows the point-measure correlation coefficients between the EAP ability measures from the Rasch model and item scores. All of the items had positive correlation coefficients as desired. Items with values of discrimination index lower than 0.2 are generally considered weak (Ebel, 1965). Among the investigated items, item 17 had an extremely small value of point-measure correlation (i.e., 0.09), while items 25 and 34 had relatively small values (i.e., 0.20 and 0.21 respectively). A discussion about these three items is presented in the next section, as part of the examination of item pattern.

2. Internal Structure

Figure 2 shows the Wright map, which displays children's ability distribution and item difficulty estimates on the same logit scale. The first column presents the ability distribution

from most able at the top to least able at the bottom of the map. Each "x" represents approximately 1.4 children. The location of the "x"s indicates the level of their ability estimates. The remaining columns present the item difficulty estimates. Items located at the top are more difficult than items at the bottom. If the ability estimate of a child and the difficulty estimate of an item are located at the same position on the map, the probability of this child succeeding on the item is 0.5.

An overview of the child ability and item difficulty estimates suggests that the studied TEMA-3 items are well targeted to our sample. Specifically, the distribution of the ability estimates approximated a bell curve (mean= −0.03 logit, variance=4.04 logit). This evidence suggests that the investigated TEMA-3 items differentiate the sample of K1 children successfully. Furthermore, an examinee's latent ability can be most precisely estimated if items administered are at the examinee's ability level. From the Wright map, it is clear that the range of item difficulty estimates covered the entire span of ability estimates well and there was no gap in the spread of difficulty estimates among these items. This finding suggests that collectively these items assess the mathematics ability of the K1 children well.

The pattern of the difficulty estimates was further examined by item types. While the items assessing informal mathematics spanned across the entire ability distribution, the items assessing the formal mathematics were mostly located at the top of the ability distribution with a few exceptions that will be discussed later. This empirical pattern supports the hypothesis that formal mathematics items assess more advanced concepts and skills compared to informal mathematics items. Also, it is expected that most of the formal mathematics items cluster at the top of the map, as these items target older children (e.g., item 32 is the entry point for 7-year-olds).

Pattern of informal mathematics items. A content analysis of the items suggests that the observed pattern of item difficulty estimates supports the theoretical hierarchy of informal mathematics concepts and skills generally. Items located at the bottom of the map assess pre-counting abilities, such as verbal counting by ones up to 10 (e.g., item 12) and perception of small numbers (e.g., item 1), and are considered basic (Ginsburg & Baroody, 2003). It is notable that the administration of these easy items is accompanied with external representations, such as tokens or pictures. Subsequently, the items located at the middle range assess counting abilities, such as verbal counting by tens up to 90 (e.g., item 33), comparisons of numbers up to 10 (e.g., item 20), nonverbally adding or subtracting two small and previously-viewed collections (e.g., item 8), and understanding of number constancy principle (e.g., item 9). These middle-level items use a mix of external representations (e.g., tokens) and verbal instructions. Last, items located at the top assess advanced counting abilities that involve deeper conceptual understanding, such as verbal counting by tens up to 190 (e.g., item 42), comparing numbers up to 99 (e.g., item 37), solving arithmetic word problems with sums up to 12 by counting or reasoning (e.g., item 32), part-whole concept (e.g., item 17), and equal partitioning (e.g., item 25). These items are not supported with external representations.

Pattern of formal mathematics items. Most of the studied formal mathematics items were clustered at the top of the Wright map. Only items in the numeral literacy category spread out on the map. The empirical hierarchy of the numeral literacy items matches the hypothesized order: the more digits a numeral has, the more difficult it is for children to read or write the numeral accurately. Items in the other categories (i.e., number facts, calculation, and concepts) require children to have a clear understanding of arithmetic rules or retrieve arithmetic facts from

memory that are typically acquired via formal schooling. It is expected that these K1 children who just started kindergarten would find these items difficult.

The TEMA-3 items were arranged from easy to difficult using data from the U.S. sample during the test development (Ginsburg & Baroody, 2003). It was of interest to investigate whether the ordering of the items determined with the Singapore sample was consistent with the ordering determined with the U.S. sample. Kendall's tau-b correlation (Kendall & Gibbons, 1990) was used to compare the association between the original item ordering in the TEMA-3 test form and the difficulty estimates (Table 1) obtained from the Rasch analysis of the Singapore sample for the 52 items under investigation. Kendall's tau-b correlation coefficient is a non-parametric measure of the rank association between two variables. The result showed a positive and strong correlation in the ordering of items between the U.S. and Singapore samples ($\tau b = 0.76$, $p<0.001$).

Ordering of informal mathematics items. As shown in the Wright map, the ordering of the informal mathematics items from the K1 Singapore sample is consistent with the U.S. sample in general, with a few exceptions in the numbering and the concepts category. For the numbering category, item 12 was easier than expected. This item assesses verbal counting skill by ones up to 10. This finding suggests that Singapore children might practice basic counting skills earlier than U.S. children. In the concepts category, items 17 and 25 were more difficult than expected. Item 17 assesses part-whole concept with several story problems (e.g., Rebecca has 7 candies left after eating 2 candies. How many candies did Rebecca have?). This requires children to solve the unknown whole start value (i.e., X–2=7). Whilst approximate answers are acceptable (children can provide any answer greater than 7 to indicate their understanding that the whole must be larger than the individual parts), children in our sample were observed trying to calculate the exact answer. As such, the question is not assessing children's general understanding of part-

whole concept as intended in item design; instead it is being seen by our children as an algebraic problem (e.g., solve X−2=7). Item 25 assesses equal partitioning with story problems about fair-sharing of discrete quantities (e.g., does each share have the same number?). Succeeding on this item requires an understanding of terms like "fairly" and "share." Past research has suggested that difficulty in comprehending particular linguistic terms could limit young children's ability to solve verbally-presented calculation problems (e.g., Gelman & Gallistel, 1986; Levine, Jordan, & Huttenlocher, 1992). It is possible that unfamiliarity with the linguistic terms resulted in the unexpected difficulty level of this item.

Ordering of formal mathematics items. As discussed previously, most of the formal mathematics items were clustered together with a few exceptions. Specifically, all of the items related to reading or writing numerals (items 14, 15, 29, 30, and 35) were much easier than expected, except for the reading or writing of 3-digit numerals (items 44 and 45). Item 18, which assesses written representation up to 5, was also easier than expected. These observations suggest that Singapore children might develop basic numeral reading and writing skills at a younger age than the U.S. children. In the concepts category, item 34 (matching a number sentence to a stated word problem) was more difficult compared to the original U.S. ordering; this item might be difficult for the K1 children as it requires buffering numerical information in memory while identifying both direct and commuted numerical representations. Also, this item contains some word problems where both direct and commuted representations are correct. Children might not have been aware that there was more than one correct answer for these problems.

3. Convergent Evidence

The key outcome variable of the Number Sets Test is the d′ score, which represents a child's sensitivity in the detection of a target quantity (i.e., 5 or 9). d′ was calculated as the difference between the z-scores for hits and false alarms (Geary et al., 2009). The Kendall's tau-b correlation was performed to investigate the relationship between the EAP ability estimates on TEMA-3 and the d′ scores on Number Sets Test. The Kendall's tau-b correlation was selected as it makes no assumption about the distribution of the variables being examined and is robust to extreme observations and to nonlinearity. The EAP ability estimates and the d′ scores were positively and moderately correlated ($\tau b = 0.39$, $p<0.001$).

Discussion

This study investigated the psychometric properties of TEMA-3 with a sample of K1 children in Singapore. To our knowledge, this study is the first attempt to investigate the properties of TEMA-3 using the Rasch model and a non-U.S. sample. Findings from the Rasch analysis indicated that the TEMA-3 items generally demonstrated good technical qualities, interpretable internal structure, and reasonable convergent evidence.

All but one of the 52 items examined fit the Rasch model well. The only item with a relatively large infit statistic was still considered productive for measurement as discussed previously. In addition, all but three items were found to discriminate children with different ability levels well. Collectively, these items demonstrated high internal consistency. Our examination of the four problematic items showed that construct-irrelevant nuisances, such as item representation, might have contributed to the unsatisfying values of fit and discrimination indices observed. Note that despite of the variation in the post-hoc item discrimination measure, these items had reasonable fit statistics, which suggests that the assumptions underlying the

Rasch model (including equal item discrimination) hold generally in the current case and supports the use of the Rasch scaling. Nonetheless, further investigations, such as conducting think-aloud interviews with children, and revisions of the items displaying lower discrimination on the post-hoc measure or larger fit statistics could improve the function of the test.

In addition to the good technical qualities, the pattern of the difficulty estimates of the studied TEMA-3 items generally supports the hypothesis about children's early mathematics development in item design. As expected, items assessing formal mathematics were more difficult than those assessing informal mathematics. Among the informal mathematics items, the observed structure is generally consistent with the progression in the development of mathematics ability. Specifically, items located at the bottom of the Wright map mostly assess pre-counting abilities, items located at the middle assess counting abilities generally, and items located at the top assess advanced counting abilities. Among the formal mathematics items, most of the items appeared to be too difficult for the current sample, which is expected as these items target children aged 7 or older. The empirical hierarchy of the numeral literacy items is consistent with the expectation that the more digits a numeral has, the more difficult it is for children to read or write the numeral.

The pattern of the difficulty estimates also unveils the impact of item characteristics on item difficulty. Specifically, the low-level items are typically accompanied with external representations, such as pictures or tokens, the middle-level items commonly use a mix of external representations and verbal instruction, and the high-level items mainly rely on verbal instruction. This finding is consistent with past research on the role of the cognitive capabilities of young children. It has been suggested that children think in relatively concrete terms, and tend to perform better on mathematics tasks with concrete objects or manipulatives than they do on

mental tasks (e.g., Ginsburg, Lee, & Boyd, 2008). It is also possible that items that just rely on verbal instruction demand fairly sophisticated language skills that children at this age do not yet possess (e.g., Mix, Huttenlocher & Levine, 2002).

The ordering of items determined with the Singapore sample was highly and positively correlated with the original item ordering in the TEMA-3 test form, which was determined with a U.S. sample. As the TEMA-3 manual advices examiners to use the basal-and-ceiling rule to administer and subsequently score TEMA-3 items, item ordering has a crucial impact on the inference made about a child's performance. For example, suppose several difficult items are placed at the beginning of the test due to an underestimation of their difficulty levels. Because of this misplacement, ceiling may be established at an earlier point than it should have been for a child, which in turn leads to an underestimation of the child's total score. Our finding suggests that the item placement in the TEMA-3 test form is generally in line with the order of item difficulties from the Rasch scaling. Note that the percentage of children succeeding on a given item was used to rank items in the U.S. sample, which is different from the use of the Rasch parameter estimates in the current study. Also, the U.S. sample is composed of children aged 3 to 8 years old, while our sample is based on K1 children only. As such, some difference in item ordering between the U.S. and our samples is expected.

The comparison of item ordering also unveils potential differences in children's development of early mathematics between U.S. and Singapore. Specifically, items observed to be easier for the Singapore sample all assess basic counting or numeral reading/writing skills. This finding is consistent with findings from Lee and Bull (2015), where the majority of the Singapore kindergarten children in their sample performed at ceiling on items assessing basic number recognition, writing, and counting on a different standardized measure. These

observations suggest that Singapore children may well practice these basic numeracy skills at a younger age than the U.S. children.

Children's Rasch ability estimates from the TEMA-3 items were positively and moderately related to the d′ score on Number Sets Test. Past research has shown that d′ scores were significantly and positively correlated with mathematics achievement in kindergarten through third grade, and were also related to measures of number, counting, and addition, after controlling for IQ and executive functioning (Geary et al., 2009). The positive relationship between performance on TEMA-3 and Number Sets Test provides support for the validity of using TEMA-3 to draw inference on children's mathematics ability.

The current study clearly has some limitations. This study investigated the properties of TEMA-3 with data from K1 children only. In addition, we did not have much data on items placed towards the beginning and the end of the test, as few children in our sample reached these items. Further studies and more data are needed to evaluate the function of all TEMA-3 items for the targeted age range (3 years to 8 years 11 months). Nonetheless, the findings from this study provide valuable implications for future test development, test use, and research. First, this study demonstrated the potential of using the Rasch model to scale TEMA-3 data. Successful application of the Rasch model has been shown to yield desired features for test development (e.g., Embretson & Reise, 2000; Wilson, 2005). For example, the separability of item and person parameters with the Rasch scaling allows test developers to equate different test forms or to continuously improve the measurement scale of an instrument over time. Second, through the Rasch scaling, person abilities and item difficulties can be compared directly. If a child's ability estimate equals an item's difficulty estimate, the probability of this child passing or failing this item is equal. Hence, in addition to norm-referenced meaning, a child's performance on the

TEMA-3 can also be related to the concepts or skills assessed by the items and have criterion-referenced meaning. Furthermore, more studies are needed to investigate whether TEMA-3 items function invariantly across other countries to establish validity evidence for cross-country comparisons.

References

Bock, R.D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*(4), 443-459.

Briggs, D. C., & Wilson, M. (2003). An introduction to multidimensional measurement using Rasch models. *Journal of Applied Measurement*, *4*(1), 87–100.

Chen, W. -H., Lenderking, W., Jin, Y., Wyrwich, K. W., Gelhorn, H., & Revicki, D. A. (2014). Is Rasch model analysis applicable in small sample size pilot studies for assessing item characteristics? An example using PROMIS pain behavior item bank data. *Quality of Life Research*, *23*(2), 485–493.

Daniel, M. H. (1999). Behind the scenes: Using new measurement methods on the DAS and KAIT. In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement: What every psychologist and educator should know* (pp. 37–63). Mahwah, NJ: Lawrence Erlbaum Associates.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., … Japel, C. (2007). School readiness and later achievement. *Developmental Psychology*, *43*(6), 1428–1446.

Ebel, R. L. (1965). *Measuring educational achievement*. Englewood Cliffs, NJ: Prentice Hall.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Fuchs, L. S., Geary, D. C., Compton, D. L., Fuchs, D., Hamlett, C. L., & Bryant, J. D. (2010). The contributions of numerosity and domain-general abilities to school readiness. *Child Development*, *81*(5), 1520–1533.

Geary, D. C., Bailey, D. H., & Hoard, M. K. (2009). Predicting mathematical achievement and mathematical learning disability with a simple screening tool: The number sets test. *Journal of Psychoeducational Assessment*, *27*(3), 265–279.

Gelman, R., & Gallistel, C. R. (1986). *The child's understanding of number*. Cambridge, MA: Harvard University Press.

Ginsburg, H. P., & Baroody, A. J. (2003). *Test of early mathematics ability (Third Ed.): Examiner's manual*. Austin, TX: PRO-ED.

Ginsburg, H. P., Lee, J. S., & Boyd, J. S. (2008). Mathematics education for young children: What it is and how to promote it. *Social Policy Report*, *22*(1), 3–22.

Jordan, N. C., Kaplan, D., Ramineni, C., & Locuniak, M. N. (2009). Early math matters: Kindergarten number competence and later mathematics outcomes. *Developmental Psychology*, *45*(3), 850–867.

Kelley, T., Ebel, R., & Linacre, J. M. (2002). Item discrimination indices. *Rasch Measurement Transactions*, *16*(3), 883–884.

Kendall, M. G., & Gibbons, J. D. (1990). *Rank correlation methods* (5th ed.). London: Griffin.

Lee, K., & Bull, R. (2015). Developmental changes in working memory, updating, and math achievement. *Journal of Educational Psychology*. Advance online publication. http://dx.doi.org/10.1037/edu0000090

Levine, S. C., Jordan, N. C., & Huttenlocher, J. (1992). Development of calculation abilities in young children. *Journal of Experimental Child Psychology*, *53*(1), 72–103.

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions*, *7*(4), 328.

Martin, R. B., Cirino, P. T., Sharp, C., & Barnes, M. (2014). Number and counting skills in kindergarten as predictors of grade 1 mathematical skills. *Learning and Individual Differences*, *34*, 12–23.

Mislevy, R. J., & Wu, P. (1996). *Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing* (RR 96-30-ONR). Princeton, NJ: Educational Testing Service.

Mix, K.S., Huttenlocher, J., & Levine, S.C. (2002). *Quantitative development in infancy and early childhood*. New York, NY: Oxford University Press.

Murphy, M. M., Mazzocco, M. M., Hanich, L. B., & Early, M. C. (2007). Cognitive characteristics of children with mathematics learning disability (MLD) vary as a function of the cutoff criterion used to define MLD. *Journal of Learning Disabilities*, *40*(5), 458–478.

Passolunghi, M. C., Lanfranchi, S., Altoe, G., & Sollazzo, N. (2015). Early numerical abilities and cognitive skills in kindergarten children. *Journal of Experimental Child Psychology*, *135*, 25–42.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danmarks Paedogogiske Institute. (Reprinted 1980 by University of Chicago Press).

Ryoo, J. H., Molfese, V. J., Brown, E. T., Karp, K. S., Welch, G. W., & Bovaird, J. A. (2015). Examining factor structures on the Test of Early Mathematics Ability—3: A longitudinal approach. *Learning and Individual Differences*, *41*, 21–29.

Ryoo, J. H., Molfese, V. J., Heaton, R., Zhou, X., Brown, E. T., Prokasky, A., & Davis, E. (2014). Early mathematics skills from prekindergarten to first grade score changes and

ability group differences in Kentucky, Nebraska, and Shanghai samples. *Journal of Advanced Academics*, *25*(3), 162–188.

Watts, T. W., Duncan, G. J., Siegler, R. S., & Davis-Kean, P. E. (2014). What's past is prologue: Relations between early mathematics knowledge and high school achievement. *Educational Researcher*, *43*(7), 352–360.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*(3), 370.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. *Rasch measurement*. Chicago: MESA Press.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACER ConQuest version 2.0: Generalised item response modelling software*. Camberwell, Victoria: Australia Council for Educational Research Press.

Table 1

Item Difficulty, Fit, Discrimination, and Percentage of Missingness

Item	Difficulty	S.E.	Infit	*t*	Discrimination	% of missingness
1	-5.69	0.33	1.10	0.50	0.40	90.53
2	-4.57	0.26	0.89	-0.90	0.61	90.01
3	-6.36	0.38	1.02	0.20	0.46	86.10
4	-4.19	0.21	1.06	0.60	0.46	83.93
5	-3.35	0.19	1.07	0.80	0.49	83.21
6	-4.83	0.23	0.90	-0.80	0.52	82.18
7	-4.15	0.20	0.84	-1.60	0.59	82.08
8	-2.53	0.15	1.04	0.70	0.55	74.05
9	-2.53	0.15	0.99	-0.10	0.57	73.64
10	-3.53	0.15	0.90	-1.20	0.60	63.65
11	-4.12	0.16	0.77	-2.50	0.61	40.68
12	-5.19	0.21	1.07	0.50	0.34	28.73
13	-1.70	0.10	0.89	-2.30	0.65	28.53
14	-4.11	0.15	0.89	-1.10	0.53	28.32
15	-1.76	0.09	0.94	-1.10	0.61	0.10
16	0.55	0.08	1.13	3.10	0.56	0.00
17	5.43	0.24	1.21	1.00	0.09	0.10
18	-2.63	0.11	1.02	0.30	0.52	0.21
19	-0.56	0.08	1.04	1.00	0.61	0.21

20	0.48	0.08	0.93	-1.70	0.63	8.03
21	-0.34	0.09	0.88	-3.00	0.65	11.33
22	0.50	0.09	0.88	-3.10	0.65	12.15
23	-1.56	0.10	1.26	4.60	0.36	12.56
24	-1.22	0.10	0.90	-1.90	0.56	18.02
25	4.94	0.20	1.17	1.10	0.20	19.77
26	0.88	0.09	1.05	1.20	0.55	20.49
27	0.12	0.09	1.35	7.70	0.37	21.42
28	0.60	0.09	1.03	0.70	0.55	21.83
29	-0.59	0.10	0.92	-1.70	0.54	28.63
30	0.97	0.09	0.90	-2.50	0.61	32.23
31	1.37	0.09	0.83	-4.30	0.66	28.73
32	3.20	0.12	0.91	-1.20	0.54	33.99
33	1.45	0.10	0.95	-1.10	0.57	40.78
34	5.80	0.29	1.06	0.30	0.21	43.87
35	0.97	0.11	0.92	-1.90	0.54	55.61
36	1.98	0.12	0.89	-2.30	0.59	60.35
37	3.21	0.14	1.05	0.70	0.43	63.65
38	2.18	0.13	1.19	3.50	0.34	64.98
39	2.40	0.13	0.85	-2.80	0.63	68.49
40	2.47	0.13	0.99	-0.10	0.51	68.80
41	4.00	0.18	0.88	-1.10	0.57	72.81
42	3.44	0.16	0.98	-0.30	0.52	74.77

43	4.70	0.22	0.91	-0.60	0.57	75.90
44	4.88	0.24	0.96	-0.20	0.51	80.64
45	4.61	0.24	0.82	-1.30	0.65	83.42
46	5.02	0.29	1.00	0.10	0.54	89.19
47	5.64	0.37	1.16	0.70	0.33	91.14
48	4.51	0.33	0.92	-0.60	0.62	94.44
49	5.62	0.42	1.19	0.70	0.37	95.06
50	5.19	0.40	1.15	0.70	0.37	95.78
51	5.23	0.43	1.15	0.70	0.41	96.40
52	4.91	0.43	0.98	-0.10	0.59	96.91

Note. The discrimination reported is the point-measure correlation.

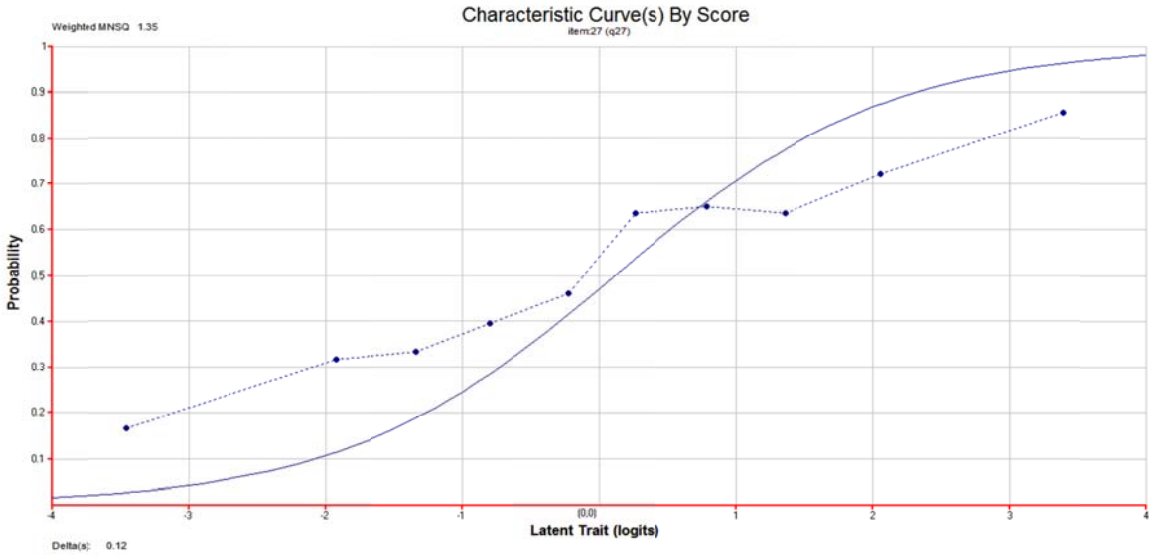

Figure 1. Item characteristics curve of item 27 (dotted line=empirical curve, solid line=modeled curve).

		Item Difficulty								
		Informal				Formal				
logit	Ability Distribution	Numbering	Number Comparisons	Calculation	Concepts	Numeral Literacy	Number Facts	Calculation	Concepts	
7										
6										
	X							49	34	47
	XXX				17					
	X						50	51		
5					25	44	46	52		
	XXX					45	43			
	XX	48								
4	XXXX						41			
	XXXX									
	XXXXX									
	XXXXXXX	42	37	32						
3	XXXXXXXXX									
	XXXXXXXXXXXX									
	XXXXXXXXXXXXXX	39	40							
	XXXXXXXXXXXXXXXXXXXX	38								
2	XXXXXXXXXXXXXXXXXXXXXX	36								
	XXXXXXXXXXXXXXXXXXXX									
	XXXXXXXXXXXXXXXXXXXXXXX	31	33							
	XXXXXXXXXXXXXXXXXXXXXXXX									
1	XXXXXXXXXXXXXXXXXXXXXXXXXXX			26		30	35			
	XXXXXXXXXXXXXXXXXXXXXXXXXXX	28		16						
	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	22	20							
0	XXXXXXXXXXXXXXXXXXXXXXXXXXXXX		27							
	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX									
	XXXXXXXXXXXXXXXXXXXXXXXXX	21	19							
	XXXXXXXXXXXXXXXXXXXXXXXXXXXXX					29				
-1	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX									
	XXXXXXXXXXXXXXXXXXXXXXXXXXX	24								
	XXXXXXXXXXXXXXXXXXXXXXXXXXX	23								
	XXXXXXXXXXXXXXXXXXXXXXXXX	13				15				
-2	XXXXXXXXXXXXX									
	XXXXXXXXXXXXXX									
	XXXXXXXXXXXXXX			8	9				18	
	XXXXXXXXX									
-3	XXXXXXX									
	XXXXXXXX	5								
	XXXXXX	10								
-4	XXXX									
	XXXX	11	4		7	14				
	XXXX	2								
	XXXXX	6								
-5	XX									
	XXX	12								
	XX									
	X	1								
-6	X									
	X	3								
-7										

Figure 2. Wright map.