ANALYSIS OF THE E-TEN CALIBRATION DATA

Ying Liu
University of Southern California
liu385@usc.edu

Arthur J. Baroody
University of Illinois at Urbana-Champaign
baroody@illinois.edu

Elizabeth M. McCarthy
WestEd
bmccart@wested.org

Michael D. Eiland
University of Illinois at Urbana-Champaign
meiland@illinois.com

The present report summarizes the analyses of the calibration data for the electronic Test of Early Numeracy (e-TEN), an adaptive, iPad-based test of early numeracy achievement. A total of 794 children age 3 years to 8 years-11 months from two states were tested. Items were designed to map onto seven theoretical domains: verbal counting, numbering, numerical relations, numeral literacy, single-digit calculation, multi-digit calculation and base-ten place-value. Analysis indicated a strong single factor—a unidimensional structure across all domains. A Wright map of calibrated item difficulty indicated that items were fairly evenly distributed along the entire continuum of the early numeracy.

Keywords: Assessment and Evaluation, Early Childhood Education, Number and Operations

Purpose and Background

The primary aim of the calibration analysis is to develop the adaptive and scoring algorithm that is psychometrically sound yet feasible to implement through standard programming (e.g., for the iOS platform). Secondary aims include (a) examination of the dimensionality, (b) refinement of the item or sublevel scoring rule, (c) item calibration, and (d) identification of "problematic" items. This report addresses these four secondary aims, which are prerequisites for achieving the primary aim.

Methods

Sample. For each 6-month interval between 3.0 to 8.99 years (12 age groups), data were collected with a total sample size of 794 (55 young 3s, 77 old 3s, 66 young 4s, 80 old 4s, 88 young 5s, 72 old 5s, 56 young 6s, 56 old 6s, 57 young 7s, 61 old 7s, 60 young 8s, and 66 old 8s). This sample size was chosen to allow appropriate use of the item response theory (IRT) model, our primary analytic framework.

Procedure. Participants were administered all items deemed appropriate for their 6-month age group regardless of their individual performance to determine the success rate for each item for each age group. The number of items scored per age group ranges from 17 (young 3s) to 28 (older 6s).

Results

Dimensionality

Dimensionality & issue of age. Among the seven proposed dimensions, all three "calculation-2" items loaded on the calculation-1 dimension. Collapsing calculation-1 & -2 resulted in a 6-dimensional model. All items loaded on a single dimension, except two (writes 2- and 3-digit numbers), which loaded on both the "grouping-and-place-value" and the "numeral literacy" dimensions.

Multi- versus unidimensionality. Table 1 compares the goodness of fit information

across models. By comparing model 1 and 3 (unidimensional vs. 6-dimensional Rasch), the 6-dimensional structure is slightly more favorable (smaller AIC and BIC). All dimensions are highly correlated, with pairwise correlation between grouping and place value and the other five dimensions being the lowest (ranging from 0.78 to 0.90). The remaining pair-wise correlations arranged from 0.90 to 0.99. Exploratory factor analysis on the estimated correlation matrix suggests a strong single factor (eigenvalue = 5.48, with the remaining eigenvalues all less than 1), which is consistent with what the correlation matrix indicates. Therefore, **a unidimensional structure across all domains is most appropriate** (see Purpura, 2010; Ryoo et al., 2015)**.**

Table 1: Summary of Model Fit Statistics.

Model #	Item Score	Dimen-sion	Age regres-sor	Deviance	AIC	AICc	BIC	# Para-meters
1	Full credit or not	1	No	16778.8	16886.8	16880.1	17139.4	54
2	Full credit or not	1	Yes	18042.3	18152.3	18145.3	18409.5	55
3	Full credit or not	6[a]	No	16160.9	16308.9	16296.5	16655.0	74

[a] Calculation-1 and Calculation-2 were collapsed into one dimension, because Calculation-2 was not well identified. A Monte Carlo method was used to estimate the 6-dimensional model.

Refinement of Partial Credit

We started with the finest partial credit scoring based on content area expertise, and refined the scoring rules using item analysis. In particular, we relied on mean-square fit statistics (MNSQ), an item misfit index frequently used in Rasch analysis that signals the amount of distortion in the measurement system due to mis-specification (i.e., misfit between data and model). Values greater than 1.0 suggest unmodeled noise in the data (i.e., something in the data is not yet explained by the model), whereas values less than 1.0 imply that the item is not very informative (i.e., the item may not provide much unique information above and beyond other items). Items with greater MNSQ values are more of a concern in our case, and we revised such items' scoring to improve the fit. In particular, we collapsed sublevels that were not well differentiated with adjacent sublevels. Two examples: (a) Item "composing & decomposing ten & base-10 equivalents of ten" contained two trials: "10 ones=ten" and "ten=10 ones." As the vast majority of students were incorrect or correct on both, we removed the original partial credit of being correct on one trial, so a student either receives a full credit if he is able to do both, or no credit. (b) The original scoring for subtraction items (e.g., "subtract to 18") entailed correctness and fluency (i.e., completion within 3 seconds). However, even the oldest group (old 8) in our calibration sample showed little evidence of fluency. In the revision, we removed the reference to fluency, and considered correctness solely for scoring.

Item Calibration

Figure 1 summarizes the calibrated overall difficulty levels of the items. As expected, the sample is fairly evenly distributed along the entire continuum. Items are located across the entire continuum, implying good coverage of the full spectrum of the developmental levels. Item "blocks" by domain are located, in general, consistent with the developmental orders predicted by theory. For instance, "numbering" appears in the low end of the scale, signaling that it is one of the first few things that a young child learns about numeracy, whereas "grouping and place-value knowledge" and "calculation-2" appear in the high end, implying that they are most

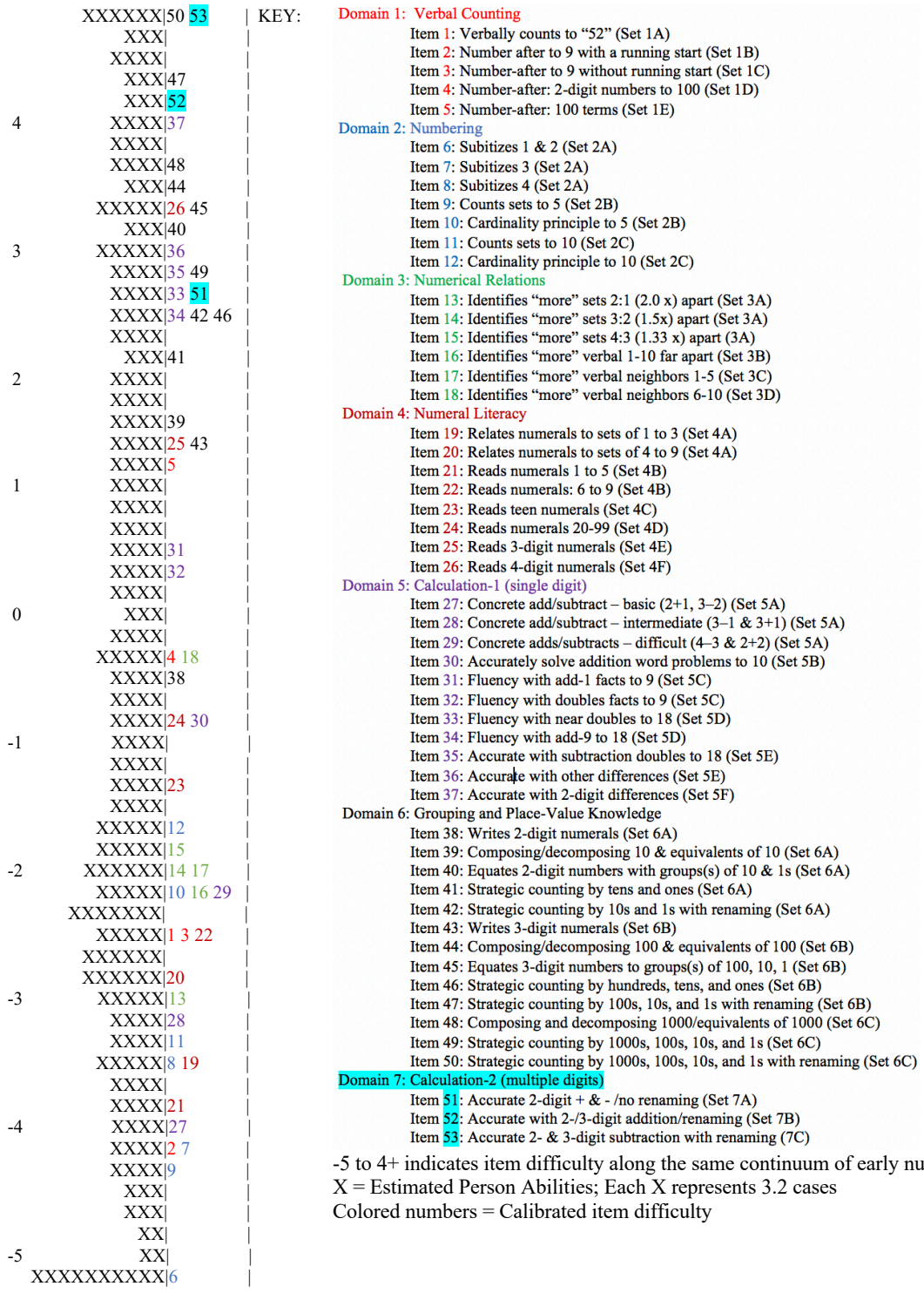

Figure 1: Wright Map of Calibrated Item Difficulty

challenging and most challenging and typically achieved by older and more capable children. Within each domain block, items spread across a range, suggesting varying degrees of abilities that such items try to capture. This also gives us leverage to further develop the adaptive algorithm, so that given a child's performance on earlier items, we will be able to choose the next best item that is closest to the child's true ability along the continuum.

Conclusions

All six dimensions were highly inter-correlated, and a unidimensional structure was strongly supported by the data. These results are not surprising given the interdependency of numeracy knowledge.

Item scoring originally provided partial credit for reasonable answers (e.g., incorrect answers that at least honored the meaning of an operation and were in the right direction), indicated a child basically understood how to determine an answer but slipped up and was off by 1 or 2, or could at least make a reasonable estimate of an answer (i.e., off or within 25% of a correct answer). However, such reflections of "operation sense" occurred so infrequently that they had to be eliminated as a candidate for partial credit.

The item calibration provides reassurance that the e-TEN assesses the whole range of numeracy ability. Some similar items, such as Items 25 (reads 3-digit numerals) and 43 (writes 3-digit numerals) and Item 33 (fluency with large near doubles such as 8 + 7) and Item 34 (fluency with combinations involving 9) had the highly similar or even the same level of difficulty. In such cases, one item may be eliminated to shorten testing times.

The item calibration largely corroborated the developmental trajectories on which items were ordered. For example, consistent with previous research (Huttenlocher, Jordan, & Levine, 1994; Levine, Jordan, & Huttenlocher, 1992), success on a nonverbal addition/subtraction task emerges in a step-like manner (with collections of 1 and 2 first, then those involving 3, and finally with 4 items; Items 27, 28, and 29, respectively) and does so before children are successful with verbally presented word problems (Item 30). The results confirmed that the small doubles such as 3 + 3 and 4 + 4 (Item 32) are among the easiest of combinations and, in something of a surprise, were even slightly easier than the highly salient add-1 combinations (Item 31; cf. Baroody, Purpura, Eiland, & Reid, 2015). Moreover, the large doubles such as 8+8 (Item 33) were as difficult as combinations involving adding 9 (Item 34). Another unexpected result was that Item 42 (strategic counting by 10s and 1s with renaming) proved as easy as Item 41 (strategic counting by 10s and 1s). Theoretically, flexibly counting representations of tens and ones by tens and then by ones should be easier than doing so when there are more than 10 representations of ones, which requires regrouping them into a ten (Chan, Au, & Tang, 2014). Moreover, in something of a surprise, Item 42 (strategic counting with 1s and 10s with renaming) was easier than Item 40 (equating 2-digit numbers with groups of 10s & 1s). (It is less surprising that Item 41 (strategic counting 1s and 10s without renaming) was easier than Item 40 or that Item 46 (strategic counting with 1s, 10s and 100s without renaming) was somewhat easier than Item 45 (equating 3-digit numbers with groups of 100s, 10s & 1s), because flexible enumeration may or may not involve understanding grouping and place-value ideas). Additional research is needed to see if these curious result holds and, if so, why.

For a full report, see: https://www.researchgate.net/project/NSF-funded-Development-of-the-Electronic-Test-of-Early-Numeracy-and-IES-funded-Evaluating-the-Efficacy-of-Learning-Trajectories-in-Early-Mathematics

Acknowledgments

This material is based upon work supported by National Science Foundation under Grant #1621470 ("Development of the Electronic Test of Early Numeracy"). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Baroody, A. J., Purpura, D. J., Eiland, M. D., & Reid, E. E. (2015). The impact of highly and minimally guided discovery instruction on promoting the learning of reasoning strategies for basic add-1 and doubles combinations. *Early Childhood Research Quarterly, 30*, 93–105. doi: 10.1016/j.ecresq.2014.09.003

Chan, W. W. L., Au, T. K., & Tang, J. (2014). Strategic counting: A novel assessment of place-value understanding. *Learning and Instruction, 29*, 78–94.

Huttenlocher, J., Jordan, N. C., & Levine, S. C. (1994). A mental model for early arithmetic. *Journal of Experimental Psychology: General, 123,* 284–296. doi:10.1037/0096-3445.123.3.284

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

Purpura, D. J. (2010). *Informal number-related mathematics skills: An examination of the structure of and relations between these skills in preschool*. (Doctoral dissertation). Retried from ProQuest Dissertations and Theses. (Accession Order No. AAT 3462344)

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.