Journal of Education and Development; Vol. 2, No. 1; April, 2018
ISSN 2529-7996 E-ISSN 2591-7250
Published by July Press

Examining Validity and Reliability of a Mathematics Assessment Tool for K-2 Students

Carl F. Siebert[1] & Jonathan L. Brendefur[1]

[1] College of Education, Boise State University, Boise, USA

Correspondence: Carl Siebert, College of Education, Boise State University, 1910 University Drive, Boise, 83725-1745, USA.

Received: January 13, 2018 Accepted: February 22, 2018 Online Published: March 6, 2018

doi:10.20849/jed.v2i1.329 URL: https://doi.org/10.20849/jed.v2i1.329

Research reported in this article was supported in part by the National Institutes of Health under award number R42HD075523-01A1.

Abstract

The Primary Math Assessment (PMA) tool is increasingly being used in multiple districts in a northwestern state. The PMA provides both screening and diagnostic information in six domains to assess mathematical proficiency in young students in their early educational years. A previous study using multidimensional Rasch analyses found support for the PMA's six-dimensional theoretical framework, and that the PMA is a reliable mathematics assessment for early grades. This study extended the examination of a Rasch model, implementing exploratory and confirmatory factor analysis, Item Response Theory, and Differential Item Functioning analyses. In doing so, this study found an IRT 2-PL model to fit best with these data and provided ways to improve the accuracy of measuring mathematical proficiency in early grades.

Keywords: exploratory factor analysis, confirmatory factor analysis, Item Response Theory, Differential Item Functioning, math proficiency assessment

1. Introduction

1.1 Need for Early Math Assessment

National and international mathematics assessments of fourth grade (9-year-old) students point to the need for better mathematics preparation for our youngest students (Clements & Sarama, 2007; Gersten, Beckman, Clarkem Foegen, Marsh, Star, & Witzel, 2009; Gersten, Clarke, Dimino, & Rolfus, 2011; NRC, 2009; Reese, Miller, Mazzeo, & Dossey, 1997). Rising first-graders who have not learned basic math concepts and skills will experience problems in elementary school (ages 6 through 14) that can carry through to high school (ages 14 through 18) (Duncan, Dowsett, Claessens, Magnuson, Huston, Klebanov, & Brooks-Gunn, 2007; Jordan, Kaplan, Ramineni, & Locuniak, 2009; Morgan, Farkas, & Wu, 2009). Nevertheless, enduring improvement can be achieved through early intervention (Jordan & Levine, 2009; LeFevre, Fast, Skwarchuk, Smith-Chant, Bisanz, Kamawar, & Penner-Wilger, 2010). Of great importance, then, are valid, reliable, and efficient assessment instruments (i.e., screeners) that can identify these young students who are experiencing problems with mathematics and mathematical thinking, followed by assessments that can diagnose in which subdomains they experience their problems.

The Common Core Standards in Mathematics in the United States are descriptions of the mathematical skills that students should have at each year of education—a general framework of expectations that instructors use when drafting their teaching plans. These include standards in numbers and operations, along with algebraic thinking, measurement, data analysis, and geometry. However, most current educational screening tools assess only number sense, to the exclusion of other critical mathematical areas, for students in their early years of education. Teachers rely on available screening instruments, despite that the content of the screener is misaligned with the Common Core Standards. Further, dated definitions of mathematics disability have caused disability assessment tools to focus on numerical quantity and number sense, which is again misaligned with current research suggesting that mathematics disability also includes symbolic processing and visual-spatial impairment (Rousselle & Noel, 2007), a deficit in working memory (Geary, 2004; Swanson, Howard & Saez, 2006), and a

hybrid of other deficiencies (Ashkenazi, Black, Abrams, Hoeft, & Menon, 2013). For these reasons, it is crucial that assessment and diagnostic instruments encompass the variety of mathematical problem subtypes, along with the content of the Common Core Standards for U.S. students.

1.2 Primary Mathematics Assessment Tool

To address the limitations of many of the current instruments, the Primary Mathematics Assessment (PMA) (Brendefur & Strother, 2010) has been developed to identify students in U.S. grades kindergarten through second grade (student ages five through eight) who are at risk for poor outcomes in six mathematics dimensions. Initially, students are assessed with the PMA-Screener (PMA-S), and students who identify as at risk are then assessed with the PMA-Diagnostic (PMA-D). These sequential, "multi-gate" assessments can first quickly assess many students, and then confirm initial screening findings for the small group of at-risk students, providing a comprehensive evaluation of the six dimensions of mathematical proficiency/deficiency. These results can be translated into targeted interventions, resulting in a more efficient use of time and the likelihood of improved proficiency for the at-risk students.

The PMA was developed to measure six predictive dimensions of future math achievement: number sequencing, operations (number facts), contextual problems, relational thinking, measurement, and spatial reasoning. A more thorough review of the development and success of the PMA is presented elsewhere (see Brendefur et al., 2015; Brendefur, Strother, & Thiede, 2012).

2. Purpose

Like many assessment instruments, the PMA was subjected to an initial validation that utilized a Rasch model, which can examine items' difficulty. However, the PMA is now in the stage of development that requires a more in-depth validation with a large population of students. To accomplish this, we conducted a multiple-stage validation, beginning with an examination of the 1) conceptual framework, 2) structural validity using factor analysis, and 3) model fit by student grade using confirmatory factor analysis. We followed this with a 4) comprehensive evaluation of each item using Item Response Theory (IRT) and a 5) determination of the appropriate IRT nested model. Finally, we calculated Cronbach's alpha to 6) determine the instrument's reliability, examination of R^2 and utilized Differential Item Functioning (DIF), to 7) determine whether any items were biased against different groups and ANOVA from Classical Test Theory (CTT) and to 8) investigate whether the instrument as a whole was biased against any student groups.

3. Methods

In the winter of 2016, 33 schools from four districts in a Northwestern state used the PMA. Within these four districts, 1530 kindergarten students, 1553 first grade students, and 1558 second grade students took the PMA-Screening (PMA-S) test. Data from all 4641 completed PMA-S tests and demographic data on the participating students were used in this study. Prior to the analysis, the data were cleaned and reviewed for missingness and for potential invalid values due to data entry errors. Sex and school names were recoded to facilitate statistical analyses and numbers of correct answers were calculated to provide a test score for each student.

3.1 Part 1

We implemented a two-part approach to study PMA-S reliability and validity. Part 1 examined the conceptual framework from which the PMA-S was developed and how well the items performed within the measurement model. The first step in Part 1 was a test of the validity of the six-dimension PMA-S framework by conducting a factor analysis. The current implementation of the PMA-S uses 18 test items on six dimensions (i.e., latent constructs) identified as number sequencing (ns), operations/number facts (nf), contextual problems (cntxt), relational thinking (rt), measurement (meas), and spatial reasoning (spa). See Figure 1 for the conceptual framework.

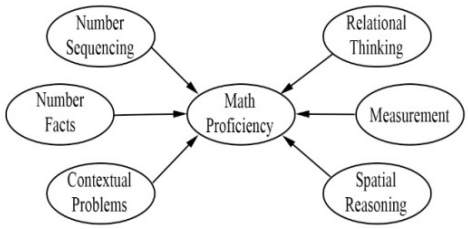


Figure 1. PMA-S conceptual framework

Mplus 7.4 (Muthen & Muthen, 2015) was used to conduct an Exploratory Factor Analysis (EFA) on 18 dichotomous variables (see Appendix A for Mplus script). The 18 variables represent the 18 PMA questions dichotomously scored 1 (correct) or 0 (incorrect). We selected WLSMV for our estimator setting. Mplus defines WLSMV as "weighted least square parameter estimates using a diagonal weight matrix with standard error and mean- and variance- adjusted chi-square test statistic that uses a full weight matrix" (p. 533) and is the default estimator for an EFA using Mplus. For the rotation, we implemented Geomin oblique to review PMA-S dimensionality for each of three grades (kindergarten, first grade, and second grade) individually and then together. Geomin was selected because of the theoretical support that the six dimensions are correlated and because "Geomin rotation is recommended when factor indicators have substantial loadings on more than one factor resulting in a variable complexity greater than one" (Muthen & Muthen, 2015, p. 537). Following the EFA, we conducted a Confirmatory Factor Analysis (CFA) for each of the three grades separately, with the three grades combined, and with covariates. Testing of multiple CFA models was necessary to explore model performance fully among the grades and when covariates were included.

The next step in Part 1 included a comprehensive item examination using Item Response Theory (IRT) using Mplus 7.4. The National Council on Measurement in Education website glossary describes IRT as, "A theory of testing based on the relationship between individuals' performances on a test item and the test takers' levels of performance on an overall measure of the ability that item was designed to measure" (NCME, 2017). Again, we selected WLSMV as our estimator for consistency and because of the nested model (i.e., restricting one estimate of the model compared with the same model with no restriction) testing capability. The testing of nested models helped identify which of the IRT models is appropriate for PMA data modeling—1PL-Rasch for examining difficulty; 2PL for examining difficulty and discrimination; and 3PL for examining difficulty, discrimination, and guessing. Examination of the parameter estimates for IRT indicated how well each item fits within the appropriate IRT model, and showed the performance of each test item for assessing student mathematical proficiency. Both the testing of nested models and examining parameter estimates contributed to a final selection as to which IRT model is most appropriate for the PMA-S and which of the 18 items were not as effective in assessing mathematical proficiency.

3.2 Part 2

In Part 2 we examined the reliability and validity of the PMA-S using SPSS 22 to explore descriptive data for the various student groups. Whereas Part 1 focused on the conceptual framework, Part 2 focused on the respondents and how well items measured mathematical proficiency without bias. The first step in Part 2 of our analysis investigated potential unexplained performance differences among groups of students based simply on the number of correct answers. For example, if a low performing district scored significantly higher on the PMA-S when compared to a higher performing district, we would use this discovery to examine the two groups. Other variables used to group students included sex, ethnicity, teacher, and grade. We used One-way ANOVA to investigate potential group differences, along with comparing descriptive information and graphs representing group scores. Using these simple review procedures allows for a more comprehensive examination of PMA-S's reliability and validity.

The second step of Part 2 closely examined each PMA-S item for the possibility of having Differential Item Functioning (DIF) characteristics. PMA-S items with DIF are biased in their ability to fairly assess mathematical proficiency between groups. For example, if wording on a PMA-S test item places a Hispanic student at a disadvantage, then the item is considered inappropriate and unable to measure mathematical proficiency effectively. Using an item that is biased has the potential to deprive a student of an equal educational opportunity due to a test score that does not represent their true mathematical proficiency (e.g., teacher intervenes with an inappropriate lesson mismatched with the students' capabilities). In addition, we used DIF detection to verify the earlier discoveries of group differences to determine whether these differences were problematic for PMA-S or

simply true variations in student proficiencies.

4. Findings

Our examination of the data discovered no missingness with regards to the 18 PMA-S items or sex. However, substantial missingness was found for the variable identifying English language learners (ELL) that removed our opportunity to include ELL in the study. Overall, we found no values that were identified as invalid, but we did see some departure from normality for the number of correct answers, which is discussed later. Next, we calculated descriptive statistics for our sample of participating students. See Table 1 for details.

Table 1. Study demographics for participating students

	Kindergarten	First Grade	Second Grade
Participants	1530 (707 female, 818 male, 5 other)	1553 (772 female, 777 male, 4 other)	1558 (765 female, 783 male, 10 other)
Schools represented	32	29	30
Ethnicity:			
not marked-0	132	83	88
American Native-1	11	12	14
Asian-2	47	48	77
Black/African American -3	39	41	53
Native Hawaiian/Pacific Islander -4	8	10	9
White not Hispanic -5	1074	1130	1078
Latino -6	148	157	180
Two or more races-7	71	72	59
ELL-LEP	5	10	5
IEP	64	98	116

4.1 EFA Findings

EFA was conducted on kindergarten (grade 0), grade 1, and grade 2 separately, and on a dataset with the grades combined. Our primary focus for the EFA was to explore the best number of dimensions to model the PMA-S items by only following the item loadings on the dimensions, and not to use EFA to explore which items should or should not be in the model. Therefore, fit indices and chi-square change significance were recorded for each possible number of dimensions from 1 to 8 (i.e., number of factors in the model). All recorded fit indices showed well-fitting models with Root Mean Square Error of Approximation (RMSEA) ranging in values from .047 to 0.00, using the good fit threshold of < .05. The Standardized Root Mean Square Residual (SRMR) also uses the < .05 threshold, and our data revealed ranges from .058 to .008, indicating well fitted models. The Comparative Fit Index (CFI) and Tucker Lewis Index (TLI) indices fit well when > .9, and we discovered CFI and TLI ranges from .941 to 1, so all models met this criterion. For grades 0 and 1, chi-square values were significant until models with at least six factors were reached, suggesting that the models with fewer than six dimensions did not match the data well. Chi-square non-significance for grade 2 data was reached with only four factors, but when all three grades were combined, non-significance was reached at seven factors. Therefore, with strong fit indices for all three grades and a marginal difference in grade 2 for reaching non-significance, the conceptual framework with six dimensions is supported initially by our EFA analyses.

Comparing factor correlations for six factor models among the three grades also show differences. All grade 0 (i.e., kindergarten) factor correlations were found to be significant at $p < .05$ and ranged in value from 0.128 to 0.742. Grade 1 factor correlations found 10 of the 15 factor correlations to be significant, ranging from -0.143 to .727. Grade 2 factor correlations were recognizably different from grades 0 and 1. Only three of the correlations were significant at $p < .05$, ranging from 0.036 to 0.782. The patterns of the factor correlations shown in Table 2 and Table 3 below are echoed in the factor structures of the three grades.

Table 2. Factor structure for Grade 0 and 1

	Factor Structure—Grade 0						Factor Structure—Grade 1					
	1	2	3	4	5	6	1	2	3	4	5	6
ns08	**0.839**	*0.623*	0.338	0.334	0.385	0.325	**0.739**	0.377	0.491	*0.547*	0.322	0.376
ns12	**0.631**	*0.492*	0.346	0.286	0.311	0.324	**0.718**	0.431	*0.511*	0.506	0.243	0.170
ns18	**0.685**	*0.579*	0.244	0.419	0.379	0.410	**0.708**	0.094	0.543	*0.594*	0.245	0.312
ns19	*0.432*	**0.460**	1.128	0.240	0.170	0.179	**0.788**	0.110	0.609	*0.661*	0.409	0.259
nf08_33	**0.716**	*0.673*	0.268	0.423	0.473	0.381	0.685	0.232	*0.718*	**0.753**	0.383	0.220
nf13_38	*0.566*	**0.747**	0.265	0.283	0.364	0.324	*0.587*	0.057	**0.858**	0.501	0.322	0.186
nf20	*0.667*	**0.674**	0.310	0.306	0.379	0.271	*0.618*	-0.070	**0.680**	0.610	0.330	0.294
rt12	0.429	*0.505*	0.237	**0.866**	0.346	0.361	**0.525**	-0.230	0.340	0.340	*0.490*	0.233
rt18	*0.513*	**0.777**	0.313	0.412	0.391	0.354	**0.583**	-0.210	0.450	*0.460*	0.360	0.199
rt23	*0.458*	**0.495**	0.125	0.409	0.310	0.373	**0.762**	-0.020	0.540	*0.570*	0.530	0.146
cntxt03	*0.575*	**0.697**	0.225	0.433	0.379	0.400	*0.503*	0.019	0.395	**0.700**	0.309	0.140
cntxt07	*0.567*	**0.718**	0.327	0.399	0.348	0.411	*0.641*	-0.020	0.510	**0.690**	0.350	0.252
means01	0.254	0.294	0.047	*0.319*	0.238	**0.677**	*0.423*	-0.030	0.320	0.410	**0.590**	0.167
means02	0.382	*0.418*	0.203	0.204	0.298	**0.651**	*0.406*	-0.020	0.360	0.340	**0.510**	0.216
means11	*0.320*	**0.432**	0.211	0.294	0.305	0.306	*0.210*	0.193	0.170	0.158	**0.323**	0.163
spa10	0.392	*0.417*	0.136	0.286	**0.772**	0.249	*0.452*	-0.080	0.410	0.430	**0.630**	0.402
spa13	0.408	*0.422*	0.094	0.294	**0.850**	0.355	0.140	-0.230	0.080	*0.160*	**0.250**	0.069
spa30	**0.287**	0.256	0.016	0.210	*0.274*	0.213	*0.254*	-0.030	0.190	0.180	0.230	**0.655**

Table 3. Factor structure for Grade 3

	1	2	3	4	5	6
ns08	0.101	**0.812**	*0.667*	0.480	0.094	0.182
ns12	0.121	**0.895**	*0.700*	0.467	0.178	-0.061
ns18	0.104	**0.725**	*0.505*	0.325	-0.170	0.082
ns19	0.109	**0.794**	*0.701*	0.473	0.263	0.159
nf08_33	0.082	*0.511*	**0.662**	0.254	0.196	0.171
nf13_38	0.117	*0.668*	**0.834**	0.418	0.242	0.149
nf20	0.103	*0.605*	**0.759**	0.434	0.282	-0.007
rt12	4.162	**0.598**	*0.541*	0.357	0.279	0.149
rt18	0.131	**0.731**	*0.612*	0.326	0.338	0.226
rt23	0.129	**0.711**	0.627	0.450	*0.637*	0.189
cntxt03	0.093	*0.503*	**0.582**	0.497	0.234	0.181
cntxt07	0.092	*0.507*	0.503	**0.900**	0.231	0.175
means01	0.077	**0.381**	*0.307*	0.221	0.084	0.272
means02	0.073	*0.373*	**0.382**	0.277	0.239	0.325
means11	0.030	*0.254*	0.211	0.168	0.010	**0.270**
spa10	0.026	0.182	0.219	0.212	*0.255*	**0.422**
spa13	0.032	0.139	*0.239*	0.196	0.034	**0.265**
spa30	0.036	*0.260*	0.203	0.136	0.144	**0.445**

Within Tables 2 and 3, bold and underlined ones identify largest loading values, while italics and underline identify second largest loading value within each grade. We focused on the factor structures for the three grades because we used oblique rotations. Tables 2 and 3 show that the identification of each factor (i.e., 1 through 6) is not the same for the three analyses (i.e., grade 0, 1, and 2), but the corresponding factors for each of the three grade analyses can be identified easily by the item loadings. Double loading was evident in many of the items. Double loading is when the largest factor loading for an item is not more than double in value from its other factor loadings. For example, item ns19 from grade 0 has a loading value of .460 with a second largest loading value of .432, and shows double loading for Grade 2 as well. This example and the other double loadings in Tables 2 and 3 indicate the degree to which the dimensions are correlated and how these items are measuring only slightly different concepts within the overall construct of mathematical proficiency. However, the presence of double loadings is an issue when developing a highly effective tool for measuring mathematical proficiency. In addition to individual items double loading on factors, some of the conceptualized groups of items do not always load on the same factor. This is especially evident for the three items for measurement (means01, means02, and means11) and for grade 2 in which no factor loadings indicate a single dimension. For all three grades combined, the loadings do not vary as much (see Table 4) but occurrences of double loadings persist.

Table 4. EFA factor loadings for grades combined

	1	2	3	4	5	6
ns08	**0.933**	*0.623*	0.509	0.526	0.469	0.318
ns12	*0.611*	0.599	**0.749**	0.543	0.258	0.237
ns18	**0.632**	0.602	*0.626*	0.568	0.335	0.346
ns19	0.591	*0.644*	**0.648**	0.534	0.446	0.206
nf08_33	*0.615*	**0.737**	0.519	0.611	0.391	0.379
nf13_38	0.528	**0.750**	0.529	*0.588*	0.409	0.292
nf20	0.564	**0.826**	0.574	*0.580*	0.321	0.333
rt12	0.438	*0.465*	0.315	0.381	**0.537**	0.298
rt18	0.486	*0.662*	**0.743**	0.478	0.277	0.461
rt23	0.321	0.335	0.247	*0.482*	**0.499**	0.033
cntxt03	0.453	*0.545*	0.396	**0.694**	0.364	0.257
cntxt07	0.463	*0.540*	0.491	**0.725**	0.409	0.219
means01	0.258	0.237	0.151	*0.307*	**0.591**	0.180
means02	0.311	0.294	0.182	*0.363*	**0.540**	0.197
means11	*0.298*	*0.298*	**0.383**	0.246	0.213	0.335
spa10	0.396	*0.440*	0.246	0.338	0.422	**0.543**
spa13	0.257	*0.345*	0.294	0.318	0.162	**0.579**
spa30	*0.260*	0.194	0.228	0.249	0.217	**0.307**

Within Table 4, bold and underlined ones identify largest loading values, while italics and underline identify second largest loading value.

The lack of clear loadings for the 18 items on six dimensions (i.e., factors) is problematic. However, fit indices and the frequency in which conceptualized items for each dimension load on the same factor when comparing among grades and when grades are combined offered some support for the initial PMA framework. Some of the items within a conceptualized dimension show that they do measure the construct effectively, while other items are not as focused on a single dimension.

4.2 CFA Findings

Given the variations in item performance in the EFA analyses, we conducted a CFA to gain additional insight into the overall structural validity of the PMA-S. Figure 2 shows the estimated relationships among the 18 items

and six latent constructs when all three grades are combined.

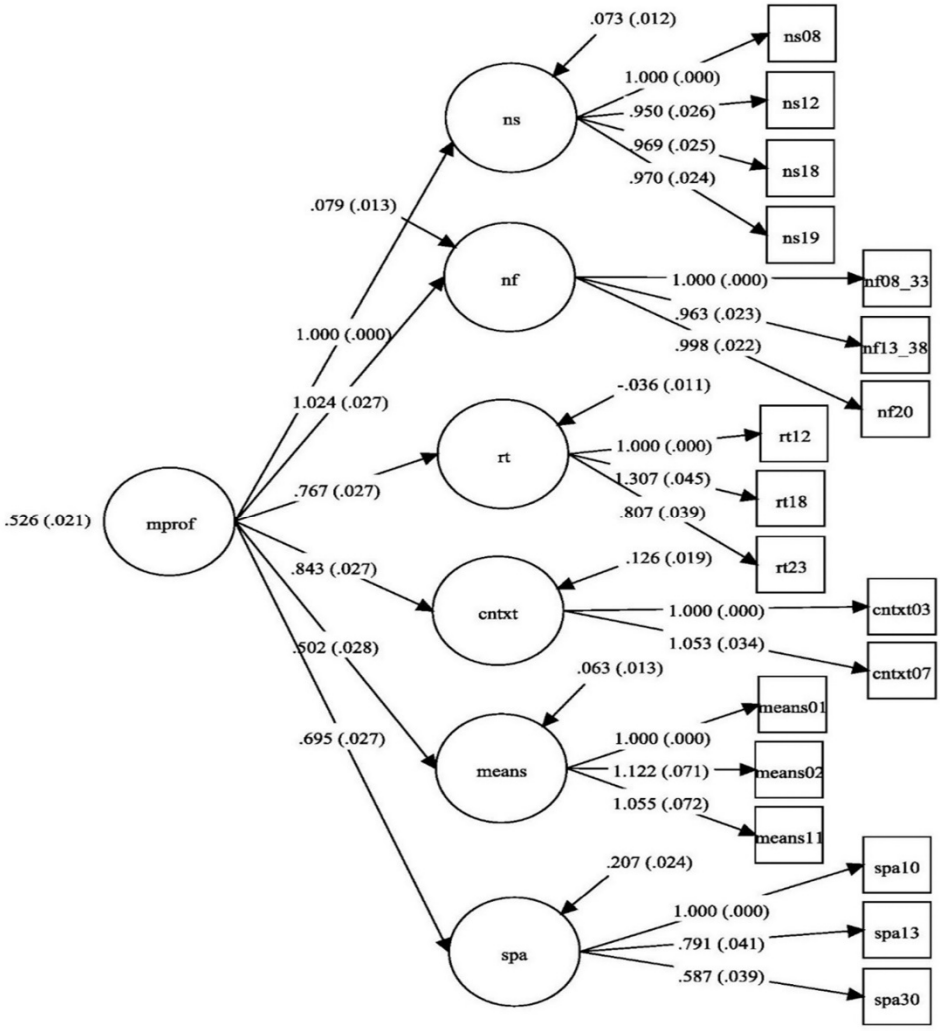


Figure 2. CFA with all three grades combined

For Figure 2, R^2 for ns, nf, cntxt, means, and spa are .879, .875, .748, .677 and .551 respectively. R^2 for rt was undefined in this model. Fit indices for this CFA model show a well-fitting model to the data prior to adding any error correlations or covariates (RMSEA = 0.035, CFI = 0.976, and TLI = 0.971). Standard errors in the model were found to be relatively small and five of the six R^2 values, representing reliability for the six latent constructs, show much of the variance within the constructs as being captured by their associated variables (R^2 represents the amount of measurement variation captured by the model). However, R^2 for spatial reasoning (spa) and measurement (means) are smaller than guidelines for well measured constructs (0.551 and 0.677, respectively).

Next covariates were added to the CFA model to investigate potential influences on specific items. After including students' grade year as a covariate for the overall latent construct of mathematical proficiency, we used Mplus's modification indices (MI) to select which of the 18 items should have a direct effect from grade, stopping when no other significant MIs exist. Adding one direct effect at a time, followed by rerunning the model, fit indices were incrementally improving even though each model, from the start, shows indices for a well-fitting model. When the suggestions ended, only five of the 18 remained unlinked (ns12, ns19, nf08_33, nf13_38, and spa30).

We used the same procedure to investigate the covariate sex. Sex was defined as 1 = female, 2 = male, and 3 = other (only 19 out of 4641 identified as other). We removed grade and added sex to the model without any direct

effects to any of the 18 items. The model with sex as a covariate resulted in a better fitting model than when grade was added, and only six of the 18 show possible significant model improvement for the MI. Unlike influence from grade, at the time MI stopped showing any suggested direct effects, 13 of the 18 remained unlinked (ns12, ns18, nf08_33, nf13_38, nf20, rt12, rt23, cntxt03, cntxt07, means11, spa10, spa13, and spa30). If adding a covariate improves model fit, then the items in the model are measuring more than their conceptualized construct. Linking the covariate to specific items is one way to identify which of the items are at fault within the model. The fact that the model improved for more links with grade than sex reveals that grade is more problematic to item effectiveness than sex.

4.3 Individual Item Performance

Following the review of the structural foundation of the PMA-S, we turned our attention to the measurement qualities of each individual item. IRT was used on each grade individually to find which of the 1-PL, 2-PL, or 3-PL models fit the data best. Given the nested nature of the models, we started with the 1-PL (i.e., Rasch) and used DIFFTEST option within Mplus (Mplus FAQ, 2017) to check for significant improvement within nested models. In addition, parameter estimates and graphs were used to examine how each model performed. Results from the DIFFTEST for each grade individually and grades combined found significant chi-square change differences going from the 1-PL to the less restricted 2-PL model, indicating that the 2-PL is a much better model for the data than 1-PL.

The slope in an Item Characteristic Curve (ICC) is the measure of discrimination for a single item. The 1-PL model (i.e., Rasch model) holds all slopes equal where the less restrictive model (i.e., 2-PL model) allows slopes to vary among items. A 2-PL model will be a better fit for the data when estimated item slopes vary. Figure 3 shows ICCs for 1-PL and 2-PL models that clearly show differences in slopes.

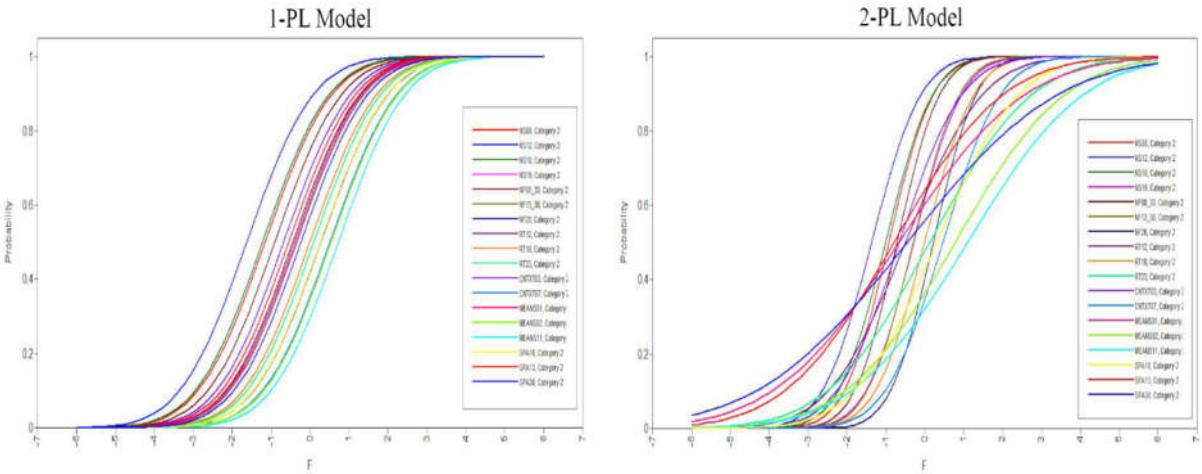


Figure 3. Comparison of ICCs for 1-PL and 2-PL models

With significant chi-square change differences and the visual confirmation from Figure 3, a 2-PL model is a better match than the 1-PL Rasch model for the PMA-S.

A different set of criteria had to be used when choosing between 2-PL and 3-PL models. DIFFTEST in Mplus is available only for three different estimators, Mean and Variance Adjusted (MLMV), Robust Weighted Least Squares (WLSMV), and Robust Unweighted Least Squares (ULSMV). However, 3-PL model parameters with dichotomous outcome variables cannot be estimated using any of these three. In addition, Maximum Likelihood Robust (MLR) is the default estimator for 3-PL modeling in Mplus, and parameter estimates from WLSMV and MLR are similar but vary somewhat, making estimate comparisons more subjective. Given this, we looked closely at the "guessing" parameter values from the 3-PL analysis using MLR estimation to determine potential differences between 2-PL and 3-PL model fit. We discovered that only three of the 72 guessing estimates (i.e., 18 items in three grade datasets and one combined dataset) had significant guessing parameters. The three occurrences were spa13 for grade 0, means01 for grade 1, and ns19 for the combined dataset. Adding the three-significant guessing 3-PL estimates to the discriminate and difficulty 2-PL estimates suggested that the 3-PL did not rise to the level of being a better model selection than the 2-PL model. If more guessing parameters

had been significantly larger than zero, making the selection between 2-PL and 3-PL more challenging, we would have conducted other model comparison tests (e.g., Wald test).

After our examination of model differences clearly selected the 2-PL model, we turned our attention to examining assessment differences among potential PMA-S student groups using classical test theory (CTT). CTT focuses more on the number of correct answers by test takers that translate typically into test scores. The distributions of test scores for the three grades of students show reasonable means and standard deviations, but they are recognizably different (see Figure 4).

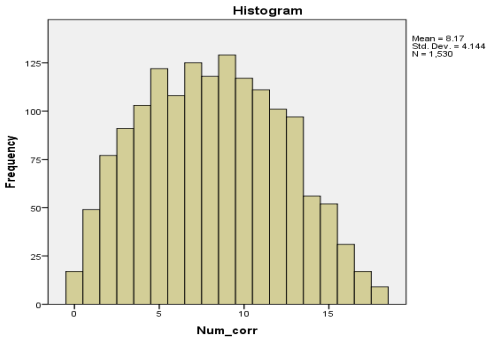
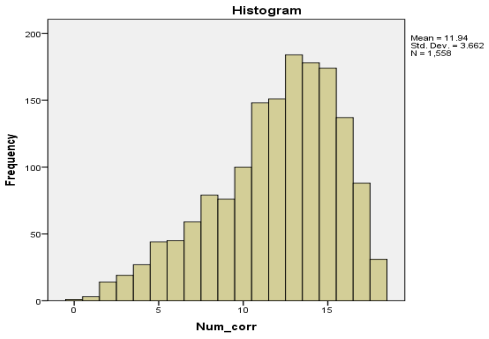
Grade	Histogram	Descriptives
Grade K		Mean = 8.17 Std. Dev = 4.144 N = 1530
Second Grade		Mean = 11.94 Std. Dev. = 3.662 N = 1558

Figure 4. Distribution of number correct scores by grade

The Kolmogorov-Smimov test for normality rejects all three distributions as being similar to a normal distribution. However, a significant departure from normality is not on its own problematic, and the changes in distribution shape from kindergarten to second grade make theoretical sense due to expected increased mathematical proficiency over time. A One-way ANOVA rejects at $\alpha = .05$ equal means for all three grades, which added to the evidence that students at different grade levels perform differently.

4.4 Item Bias

Turning our attention to potential assessment differences by sex, mean test scores for females across the three grades were 7.74, 10.03, and 11.73, for males mean test scores were 8.55, 11.00, 12.14, and mean scores for other sex were 6.60, 7.25, and 11.20 for grades 0 through 2 respectively. A One-way ANOVA found female and male scores to be significantly different at $\alpha = .05$ for grades 0 and 1, but not significant for grade 2. However, statistical significance alone is insufficient in determining differences. The large sample size can contribute to statistically significant differences between males and females, yet in our sample the other sex category did not have enough power to reach statistical significance. Practically, the mean differences between males and females were less than one correct answer apart, which does not support a finding of practically important difference. This illustrates why CTT is not the preferred approach to understanding assessment bias. IRT and DIF analyses provide more insight into which items on a test are placing a particular group at a disadvantage.

An approach commonly used to detect DIF items is the Mantel-Haenszel (Mantel & Haenszel, 1959). Mantel-Haenszel tests each item, identifying which items on an assessment reject the possibility that no DIF characteristics are present. However, because our dataset is large (i.e., high statistical power), we were concerned that too many false positives would appear. Using a ranking approach, placing items in the order of potential DIF,

offered a more hands-on approach in identifying biased items allowed us to compare items while considering their theoretical contribution to assessing mathematical proficiency. We used one such ranking approach, the Area Method proposed by Raju (Raju, 1988), to explore for DIF characteristics in the PMA-S items. The Area Method uses the amount of area found between ICCs from two separate groups to identify the potential level of DIF present. Because the calculated area represents the potential amount of DIF, items can be sorted by their calculated area values and then reviewed. We executed the Area Method following a 2-PL model on the PMA-S items to explore possible sex and grade bias by looking at each sex/grade combination against all others. Grouping sex and grade provided a better understanding of who exactly was experiencing the most bias, and which items in the PMA-S are most problematic.

Table 5. Potential DIF items using area method

Group	number sequencing				number facts			relational thinking		
	ns08	ns12	ns18	ns19	nf08_33	nf13_38	nf20	rt12	rt18	rt23
G0 female	0.039	0.455	0.483	***1.267***	0.023	0.288	0.435	0.740	***1.261***	***1.806***
G1 female	0.747	0.268	0.188	0.274	0.181	0.292	0.360	***1.100***	0.799	***1.602***
G2 female	0.422	***1.024***	0.319	0.327	0.854	0.085	0.343	0.681	0.462	***1.108***
G0 male	0.588	0.172	0.134	0.721	0.175	0.055	0.568	0.560	***1.897***	***1.911***
G1 male	0.480	0.293	0.280	0.293	0.266	0.302	0.067	***1.280***	0.697	***2.094***
G2 male	0.330	0.795	0.495	0.421	0.212	0.256	0.242	***1.078***	0.345	***1.259***

Group	contextual problems		measurement			spatial reasoning		
	cntxt03	cntxt07	means01	means02	means11	spa10	spa13	spa30
G0 female	0.494	0.310	***1.728***	***1.021***	***1.128***	0.391	***2.136***	0.741
G1 female	0.047	0.418	***1.073***	0.348	***1.428***	0.129	***2.497***	0.273
G2 female	0.396	0.402	0.030	0.230	***1.116***	***1.597***	***2.043***	0.587
G0 male	0.289	0.373	***1.316***	***1.172***	***1.365***	0.517	***2.706***	0.064
G1 male	0.322	0.429	***1.616***	0.563	***1.175***	0.365	***2.073***	0.919
G2 male	0.589	0.463	0.697	0.299	0.812	***1.972***	***1.058***	0.659

Table 5 shows calculated Area values for different groups of students according to their sex and grade. For Table 5, bold and italics identify numbers larger than one, and underlined was added to numbers larger than two. Each row in the table represents one DIF analysis using the Area Method that allows for a comparison of potential DIF by item (18 columns representing the 18 items in the PMA-S). In addition, to help the comparison of items, we bolded the Area values larger than one, and for values larger than two we added an underline. A comparison reveals that item one of the items measuring spatial reasoning (spa13) is showing signs of bias (i.e., signs of large DIF characteristics). Three other items in two different dimensions are showing signs of bias, one in relational thinking (rt23) and two in measurement (meas01 and meas11). Six items show minimal signs of DIF characteristics, but with the earlier presented findings these signs may be related more to grade bias rather than male/female differences.

5. Conclusion

EFA and CFA both provide support for the conceptual framework. The model was well fit to six dimensions, which we believe would be even stronger if the three items found to have DIF characteristics were replaced with less biased items. In addition, IRT's 2-PL model was most appropriate for modeling mathematical proficiency, primarily due to its varying discrimination capabilities for items that are not recognized with a more conservative Rasch model.

DIF analyses do show that PMA-S has some bias based on grade level and sex. Assessing measurement aptitude in students will be much improved if further review of measurement items were conducted. On the other hand, three of the six dimensions (number sequencing, number facts, and contextual problems) show very few signs of DIF.

A limitation to our DIF analysis was the lack of access to other demographic information. As new test questions are added to the item pool, it will be very important to conduct a study dedicated to DIF with a wide range of student and teacher demographics. In addition, the variations in test item measurement effectiveness increased the likelihood that item bias using DIF analysis was not identified.

Nevertheless, the PMA-S was found to assess mathematical proficiency in kindergarten to second grade students remarkably well. As we continue development of this instrument, more work needs to be conducted in understanding grade performance differences and other differences that may be problematic.

6. Implications

The findings from this study support the theoretical construct that an assessment of mathematical proficiency must include the measure of six dimensions. Other assessment tools only address a few, and in fact are not effective in early grades. The utilization of an assessment tool that measures effectively mathematical proficiency on multiple dimensions, in early childhood education, is essential to the growth of math capabilities in older students. Missing the opportunity to assist children who are experiencing a proficiency gap in one of the dimensions can haunt a student's growth for many years, if not indefinitely. Though this study indicates some important issues that need to be addressed as the development of the PMA-S continues, the findings do show a significant opportunity to improve mathematical proficiency in K-2 students.

References

Ashkenazi, S., Black, J. M., Abrams, D. A., Hoeft, F., & Menon, V. (2013). Neurobiological underpinnings of math and reading learning disabilities. *Journal of Learning Disabilities, 46*(6), 549-569. http://doi.org/10.1177/0022219413483174

Brendefur, J. L., & Strother, S. (2010). *Idaho's Primary Mathematics Assessment: 2010 Research Report*. Boise, ID: Center for Developing Mathematical Thinking.

Brendefur, J. L., Strother, S., & Thiede, K. (2012). *Primary Mathematics Assessment* (Technical Report). Boise, ID: Developing Mathematical Thinking Institute.

Brendefur, J., Johnson, E. S., Thiede, K. W., Smith, E. V., Strother, S., Severson, H. H., & Beaulieu, J. (2015). Developing a comprehensive mathematical assessment tool to improve mathematics intervention for at-risk students. *International Journal for Research in Learning Disabilities, 2*(2), 65-90.

Clements, D. H., & Sarama, J. (2007). Effects of a preschool mathematics curriculum: Summative research on the Building Blocks project. *Journal for Research in Mathematics Education, 38,* 136-163. Retrieved from http://www.jstor.org/stable/30034954

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., & Brooks-Gunn, J. (2007). School readiness and later achievement. *Developmental Psychology, 43*, 1428-1446. https://doi.org/10.1037/0012-1649.43.6.1428

Geary, D. C. (2011). Cognitive predictors of individual differences in achievement growth in mathematics: A five year longitudinal study. *Developmental Psychology, 47*, 1539-1552. https://doi.org/10.1037/a0025510

Gersten, R., Beckmann, S., Clarke, B., Foegen, A., Marsh, L., Star, J. R., & Witzel, B. (2009). *Assisting students struggling with mathematics: Response to Intervention (RtI) for elementary and middle schools (NCEE 2009-4060)*. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://ies.ed.gov/ncee/wwc/publications/ practiceguides/

Gersten, R., Clarke, B., Dimino, J., & Rolfhus, E. (2011). *Universal screening measures of number sense and number proficiency for K-1: Preliminary findings* (Report No. 2011-1). Los Alamitos, CA: Instructional Research Group.

Jordan, N. C., & Levine, S. C. (2009). Socioeconomic variation, number competence, and mathematics learning difficulties in young children. *Development Disabilities Research Review, 15*(1), 60-68.

Jordan, N. C., Kaplan, D., Ramineni, C., & Locuniak, M. N. (2009). Early math matters: Kindergarten number competence and later mathematics outcomes. *Developmental Psychology, 45*, 850-867. https://doi.org/10.1037/a0014939

LeFevre, J. A., Fast, L., Skwarchuk, S. L., Smith-Chant, B. L., Bisanz, J., Kamawar, D., & Penner-Wilger, M. (2010). Pathways to mathematics: Longitudinal predictors of performance. *Child Development, 81,* 1753-1767. https://doi.org/10.1111/j.1467-8624.2010.01508.x

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719-748.

Morgan, P. L., Farkas, G., & Wu, Q. (2009). Five-year growth trajectories of kindergarten children with learning difficulties in mathematics. *Journal of Learning Disabilities, 42*, 306-321. https://doi.org/10.1177/0022219408331037

Mplus FAQ. *How can I compute a chi-squared test for nested models with the MLMV or WLSMV estimators (difftest)?* UCLA: Statistical Consulting Group. Retrieved January 2017, from http://www.ats.ucla.edu/stat/mplus/faq/difftest.htm

Muthén, L. K., & Muthén, B. O. (1998-2015). *Mplus User's Guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

NCME. *Glossary of Important Assessment and Measurement Terms*. National Council on Measurement in Education. Retrieved January 2017, from http://www.ncme.org/ncme/NCME/Resource_Center/Glossary/NCME/Resource_Center/Glossary1.aspx?hkey=4bb87415-44dc-4088-9ed9-e8515326a061#anchorI

NRC. (2009). *Mathematics learning in early childhood: Paths toward excellence and equity*. Washington, DC: The National Academies Press.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53*(4), 495-502. https://doi.org/10.1007/BF02294403

Reese, C. M., Miller, K. E., Mazzeo, J., & Dossey, J. A. (1997). *NAEP 1996 mathematics report card for the nation and the states: Findings from the National Assessment of Educational Progress* (Publication No. NCES 97-488). Washington, DC.

Rousselle, L., & Noel, M. P. (2007). Basic numerical skills in children with mathematics learning disabilities: a comparison of symbolic vs non-symbolic number magnitude processing. *Cognition, 102*(3), 361-95. https://doi.org/10.1016/j.cognition.2006.01.005

Swanson, H. L., Howard, C. B., & Saez, L. (2006). Do different components of working memory underlie different subgroups of reading disabilities? *Journal of Learning Disabilities, 39*(3), 252-69. https://doi.org/10.1177/00222194060390030501

Appendix A

Appendix A1. Mplus Code Example for EFA

```
TITLE:
    EFA for grade combined - oblique Geomin rotation
    Kindegarden through G2
DATA:
    FILE IS GC_Data.dat;
VARIABLE:
    NAMES ARE GC_ID grade ns08 ns12 ns18 ns19 nf08_33
      nf13_38 nf20 rt12 rt18 rt23 cntxt03 cntxt07
      means01 means02 means11 spa10 spa13 spa30;
 USEVARIABLES ARE ns08 ns12 ns18 ns19 nf08_33
      nf13_38 nf20 rt12 rt18 rt23 cntxt03 cntxt07
      means01 means02 means11 spa10 spa13 spa30;
 CATEGORICAL ARE ns08 ns12 ns18 ns19 nf08_33
      nf13_38 nf20 rt12 rt18 rt23 cntxt03 cntxt07
      means01 means02 means11 spa10 spa13 spa30;
```

```
ANALYSIS:
!       TYPE = efa 6 6;    ! used for one dim model at a time
       TYPE = efa 1 8;
!       ROTATION = quartimin;    ! same as Geomin results
       ITERATIONS = 100000;
PLOT:
        TYPE = PLOT2;
```

Appendix A2. Mplus Code Example for IRT Analyses

```
TITLE:
        IRT model analysis of winter 2016
        Kindegarden data - G0
DATA:
        FILE IS GC_CFA.dat;
VARIABLE:
        NAMES ARE Dis_no Sch_no sex ethnic ell_lep iep
                grade ns08 ns12 ns18 ns19 nf08_33
                nf13_38 nf20 rt12 rt18 rt23 cntxt03 cntxt07
                means01 means02 means11 spa10 spa13 spa30;
        USEVARIABLES ARE ns08 ns12 ns18 ns19 nf08_33
                nf13_38 nf20 rt12 rt18 rt23 cntxt03 cntxt07
                means01 means02 means11 spa10 spa13 spa30;
        USEOBSERVATIONS ARE (grade EQ 0);
        CATEGORICAL ARE ns08 ns12 ns18 ns19 nf08_33
                nf13_38 nf20 rt12 rt18 rt23 cntxt03 cntxt07
                means01 means02 means11 spa10 spa13 spa30;
                        !when 1 and 2-PL modeling
!           CATEGORICAL = ns08 (3pl) ns12-spa30 (3pl);
!                       !when 3-PL modeling
ANALYSIS:
!         ESTIMATOR = MLR;
!                 ! when for 3-PL modeling
        ESTIMATOR = WLSMV;
                 ! when 1 2-PL modeling
!         DIFFTEST = mydiff.dat
!                 ! second step testing nested models
MODEL:
        f BY ns08-spa30* (1);
                    !when 1-PL modeling
!         f BY ns08-spa30*;
!                   !when 2 3-PL modeling
        f@1;
        [f@0];
```

```
OUTPUT:
        STDYX TECH1 TECH5;
PLOT:
      TYPE = PLOT1 PLOT2 PLOT3;
SAVEDATA:
         difftest is mydiff.dat;
              !first step testing nested models
```