5

Research and Psychometric Data

This chapter provides the technical information that went into the development of the Morphological Awareness Test for Reading and Spelling™ (MATRS™) (i.e., the technical documentation). We first describe the item development (Years 1 and 2), including the participants who participated in Years 1 and 2, the procedures for data collection, how the data were analyzed, and the results of those analyses. We then describe the task validation studies (Years 1 and 3). This section also provides information on how the data were analyzed and their outcomes.

ITEM DEVELOPMENT PRIOR TO STUDIES

Item development occurred prior to Year 1 of MATRS as well as in Years 1 and 2. Prior to Year 1, the authors developed a potential set of items that aligned well with the specifically defined aspect of the MATRS definition/conceptual framework. For each subtest, the authors created items that represented inflectional and derivational morphemes to capture developmental variations in inflectional versus derivational morphological awareness (e.g., Carlisle & Nomanbhoy, 1993). Depending on the subtest, they also developed specific derivational items that were transparent with their base forms and represented a phonological shift, an orthographic shift, or a combined phonological and orthographic shift based on research that suggests differences in ease of response to items may be due to transparency between base and derived forms (e.g., Apel & Thomas-Tate, 2009; Carlisle & Stone, 2005). In consultation with a corpus expert, the authors used websites, such as English Lexicon Project (http://elexicon.wustl.edu), to determine transparency/shifts. Task items for each subtest were developed specifically for each grade level (SPELL-Links Word List Maker, 2010; Zeno et al., 1995). The authors ensured similar word frequency levels (Zeno et al., 1995; corpus of 60,527 words sampled from children's texts) among grades, with ranges of word frequencies within tasks to ensure range of difficulty. The authors purposely did not use low-frequency words that represented rare words (Standard Frequency Index below 30; Carlisle & Katz, 2006). Affixes were chosen from lists of the most common prefixes and suffixes (e.g., Baumann et al., 2002; Honig et al., 2000; White et al., 1999).

After the list of items was developed, items in each subtest were reviewed by a panel of experts to determine whether every item matched the defined aspect of the four components of MATRS. The panel included two consulting experts with proficiency in language and literacy, particularly in morphological awareness. Each expert rated each item for its suitability for the subtest (i.e., coherence with the definition) based on a 1–5 scale, with 1 representing lack of coherence with the definition and 5 representing solid coherence with the definition. Only items with an average score of 3 or better were maintained. Additionally, the experts judged whether the

items represent a range of difficulty suitable for students in grades 1–6 and fairness to intended population subgroups (e.g., by gender, race/ethnicity, socioeconomic status, or disability) based on their educational and clinical expertise. If the experts believed any items were inappropriate or unfair, those items were replaced. The authors also consulted the Smarter Balanced Assessment Consortium's *Bias and Sensitivity Guidelines* (2012) to ensure the items were culturally and/or geographically sensitive. The end product was a pilot set of items per subtest that characterized the intended construct and represented a range of inflectional and derivational forms.

The pilot set of items then was used in Year 1 of MATRS item development. The main purpose was to test out a representative number of items per subtest to determine the exact factor structure at each grade level that then would guide writing test items for Year 2. Specifically, Year 1 item development was to determine whether all MATRS subtests should be included in a field test version of the MATRS (Year 2) and, if so, which items were most discriminative.

In Year 2 of item development (field test version of MATRS), the authors determined the psychometric adequacy of MATRS. Specifically, the authors administered MATRS to a large number of students to evaluate each item within each subtest for the informational value it provided to the assessment of morphological awareness. Some students were readministered MATRS a few months after the initial administration to determine reliability. The end goal was to reduce the total number of items to the most parsimonious, reliable item set.

Year 3, the task validation study, included administration of two versions of MATRS based on field testing from Year 2. In addition, students were assessed with multiple measures of language and literacy. Using these two versions, the authors assessed the convergent and predictive validity of MATRS, which involved how well MATRS related to and predicted students' performances on norm-referenced measures of word-level reading and reading comprehension, spelling, and vocabulary. By the end of Year 3, the authors had a solid understanding of the convergent and predictive ability of MATRS.

PARTICIPANTS

Across 3 years of testing, a total of 5,402 students from first through sixth grade were assessed. During the item calibration studies in Years 1 and 2, a total of 4,059 students served as participants. The students in these 2 years served as the normative sample for the psychometric analyses. Year 3, the task validation study, included another 1,343 students. These students served as the normative sample for the purpose of constructing initial grade-based norms. Across all 3 years, the students were recruited from both public and private schools in the southeastern region of the United States. Parents of all students provided consent for their children to participate, using a consent form approved by the university institutional review board. Specific demographic information for all students is contained in Table 5.1.

Item Calibration Full Sample

Across the 2 years of the item calibration study, we collected data in the southeastern United States, specifically Georgia and South Carolina. A total of 4,059 students participated in the 2-year item calibration studies in Grade 1 (N = 807), Grade 2 (N = 726), Grade 3 (N = 756), Grade 4 (N = 644), Grade 5 (N = 626), and Grade 6 (N = 501). Full-sample demographics were 53.97% female; race/ethnicity was 43.81% white, 37.70% black, 9% multiracial, 5.24% Latinx, 1.50% Asian, <1% Native American, and 1.75% no response. More than 90% of the sample reported mother and father education as having completed high school or having earned a General Educational Development (GED) credential. Ninety-four percent of the sample spoke English as the primary language, 3% spoke Spanish as the primary language, and 3% of the sample spoke another primary language in the home. Relative to the U.S. Census data, the item calibration sample across the 2 years included a larger percentage of black students and a smaller percentage of Latinx and Asian students. To understand the potential differences in representation that might manifest

Table 5.1. Demographic Characteristics of the Participants by Year

		Year 1		Year 2		Year 3	
Demographic	Category	N	%	N	%	N	%
Sex	Male	512	45.63	1,357	46.2	630	46.91
	Female	610	54.37	1,580	53.8	713	53.09
Race/ethnicity	White	461	41.09	1,318	44.85	672	50.04
	Black	430	38.32	1,101	37.46	405	30.16
	Latinx	72	6.42	141	4.8	87	6.48
	Asian	18	1.6	44	1.5	23	1.71
	Native American	4	<1%	17	<1%	3	<1%
	Multiracial	109	9.72	192	6.5	107	7.97
	No response	28	2.5	126	4.3	46	2.64
Mother's education	Less than high school	23	2.05	35	1.19	14	1.04
	Some high school	63	5.62	158	5.38	57	4.24
	GED	37	3.3	92	3.13	40	2.98
	High school grad	289	25.76	644	21.94	259	19.29
	Some college	258	22.99	570	19.42	289	21.52
	Associate or technical degree	138	12.3	327	11.14	144	10.72
	Bachelor's degree	151	13.46	537	18.3	233	17.35
	Graduate degree	102	9.09	382	13.02	228	16.98
	No response	61	5.43	94	6.48	79	5.88
Father's education	Less than high school	25	2.23	46	1.57	21	1.56
	Some high school	83	7.4	207	7.06	65	4.84
	GED	50	4.46	123	4.19	60	4.47
	High school grad	373	33.24	791	26.96	369	27.48
	Some college	193	17.2	411	14.01	243	18.09
	Associate or technical degree	91	8.11	233	7.94	97	7.22
	Bachelor's degree	102	9.09	426	14.52	185	13.78
	Graduate degree	47	4.19	255	8.69	138	10.28
	No response	158	14.08	447	15.06	165	12.29
Language	English	1053	93.85	2,789	93.54	1265	94.19
	Spanish	54	4.81	87	2.96	56	4.17
	Other	15	1.34	63	3.5	22	1.64
Special services	No services	744	66.31	2,137	72.71	942	70.14
	Speech only	82	7.31	215	7.32	131	9.75
	Language only	7	1	16	1	5	<1%
	Reading/writing only	97	8.65	251	8.54	110	8.19
	Multiple services	129	11.5	247	8.4	135	10.05
	Other	63	5.23	73	2.03	20	1.87

through bias in item-level performance, differential item functioning analyses were conducted (see results below). The majority of students participating in the item calibration study did not receive special education services (70%), with approximately 7% of the sample receiving speech services only, 9% receiving reading/writing services only, 9% receiving multiple services, and 5% noting other services that were provided.

A description of the participants in Year 1 and Year 2 of the item validation sample follows.

Year 1 Item Calibration Sample A total of 1,122 students participated in Year 1 of the item validation study in Grade 1 (N = 179), Grade 2 (N = 224), Grade 3 (N = 167), Grade 4 (N = 210), Grade 5 (N = 180), and Grade 6 (N = 162). The sample, as shown in Table 5.1, was 54.37% female; race/ethnicity was 41.09% white, 38.32% black, 9.72% multiracial, 6.42% Latinx, 1.6% Asian, <1% Native American, and 2.5% no response. More than 90% of the sample reported mother and father education as having completed high school or a GED. Approximately 94% of the sample spoke English as the primary language, 4.81% spoke Spanish as the primary language, and 1% spoke another primary language in the home. The majority of students participating in the item calibration study did not receive special education services (66%), with approximately 7% of the sample receiving speech services only, 8.65% receiving reading/writing services only, 11.5% receiving multiple services, and 5% noting other services that were provided.

Year 2 Item Calibration Sample A total of 2,939 students participated in Year 2 of the item validation study in Grade 1 (N = 628), Grade 2 (N = 502), Grade 3 (N = 589), Grade 4 (N = 434), Grade 5 (N = 446), and Grade 6 (N = 339). The sample, as shown in Table 5.1, was 53.8% female; race/ethnicity was 44.85% white, 37.46% black, 6.5% multiracial, 4.8% Latinx, 1.5% Asian, <1% Native American, and 4.3% no response. More than 90% of the sample reported mother and father education as having completed high school or a GED. Approximately 94% of the sample spoke English as the primary language, 2.96% spoke Spanish as the primary language, and 3.5% spoke another primary language in the home. The majority of students participating in the item calibration study did not receive special education services (73%), with approximately 7% of the sample receiving speech services only, 8.54% receiving reading/writing services only, 8.4% receiving multiple services, and 2% noting other services that were provided.

MATRS Task Validation Full Sample

In Year 3 of the study, a total of 1,343 students participated in a validation study in Grade 1 (N = 293), Grade 2 (N = 262), Grade 3 (N = 233), Grade 4 (N = 197), Grade 5 (N = 173), and Grade 6 (N = 185). Table 5.1 shows that the sample was 53.09% female; race/ethnicity was 50.04% white, 30.16% black, 7.97% multiracial, 6.48% Latinx, 1.71% Asian, <1% Native American, and 2.64% no response. More than 90% of the sample reported mother and father education as having completed high school or a GED. Approximately 94% of the sample spoke English as the primary language, 4.17% spoke Spanish as the primary language, and 1.64% spoke another primary language in the home. The majority of students participating in the item calibration study did not receive special education services (70%), with approximately 10% of the sample receiving speech services only, 8.19% receiving reading/writing services only, 10% receiving multiple services, and 1.87% noting other services that were provided.

OTHER MEASURES ADMINISTERED

In addition to completing the MATRS tasks, the participants were administered norm-referenced tests measuring their word-level reading, reading comprehension, spelling, phonological awareness, and receptive vocabulary skills; these tests are considered to be "gold standard" measures of literacy and literacy-related skills. Morphological awareness has been shown to be related to and predictive of language (e.g., Sparks & Deacon, 2015), literacy (e.g., Larsen & Nippold, 2007; McCutchen et al., 2008; Singson et al., 2000; Wolter et al., 2009), and other linguistic awareness skills (e.g., Apel, Wilson-Fowler, et al., 2012). These standardized, norm-referenced measures

were administered to assess convergent and concurrent validity of MATRS. Convergent validity was tested through correlations among the MATRS tasks with each other. Concurrent validity was tested through the correlations of the MATRS tasks with the gold standard measures. For all of the other norm-referenced measures, the instructions in the test manuals were followed. The order of administration of the tests varied within and across grades and sessions.

Reading Measures

A number of measures were administered to assess the students' word-level reading and reading comprehension skills. For word-level reading, the students were administered subtests from either the Test of Word Reading Efficiency, Second Edition (TOWRE-2; Torgesen et al., 2012) or the Woodcock Reading Mastery Tests, Third Edition (WRMT-3; Woodcock, 2011). On the TOWRE-2, the Sight Word Efficiency (SWE) subtest measured student skills in real-word reading (i.e., stored knowledge of written words). The Phonemic Decoding Efficiency (PDE) subtest assessed the students' ability to decode pseudowords. The students were given 45 seconds to read as many real and pseudowords, respectively. The TOWRE-2 authors' manual reports an alternate form reliability of .93.

Word-level reading also was measured using the Word Identification and Word Attack subtests of the WRMT-3. The Word Identification subtest measured the students' abilities to read single real words in isolation. Split-half reliability for this subtest is .91. The Word Attack subtest measures the students' skills to decode single pseudowords until a ceiling is reached. Split-half reliability for this subtest is .89.

We administered the Passage Comprehension subtest of the WRMT-3 to measure the students' reading comprehension skills. For this subtest, students read short passages and then stated the word(s) missing from those passages. Split-half reliability coefficients are between .84 and .96 for first- through sixth-grade students.

Spelling Measure

The Test of Written Spelling–Fifth Edition (TWS-5; Larsen et al., 2013) was used to assess students' spelling abilities. On this test, the examiner said a word, used it in a sentence, and then said the word again. The students then were required to spell the word. This task was conducted at the group level to decrease testing time; thus, all students were required to spell all 50 words on the test. However, each student's responses were scored according to the measure's basal and ceiling guidelines. According to the authors' manual, test–retest reliability is .92 for grades 1–5 and .90 for grades 6–8.

Phonological Awareness Measures

The Elision and Blending Words subtests from the Comprehensive Test of Phonological Processing–Second Edition (CTOPP-2; Wagner et al., 2013) were administered to measure the students' phonological awareness skills. The Elision task measured the students' abilities to delete a whole word from a compound word, onsets from rime units, codas from words, and phonemes within rime units. The Blending Words task required that the students listen to sounds of words presented in segments and then blend the sounds together to create real words. According to the authors' manual, test–retest reliability for the Elision task ranges from .77 to .93 for the ages involved. For the Blending Words subtest, test–retest reliability ranges from .74 to .79 for the ages of our participants.

Receptive Vocabulary Measure

We administered the Peabody Picture Vocabulary Test, Fourth Edition (PPVT-4; Dunn & Dunn, 2007) to measure the students' receptive vocabulary skills. On this test, students point to one of four pictures that matches a verbally presented word. The authors' manual states that split-half reliability is .94.

PROCEDURES ACROSS ALL YEARS OF ITEM CALIBRATION AND TASK VALIDATION

All students who participated across the 3 years were administered all tasks within their school building. In Years 1–3, all MATRS tasks were administered (note: the Suffix Choice and Written Relatives tasks were not administered to first- and second-grade students). In addition, in Years 1 and 3, the norm-referenced measures were administered. Typically, the Segmenting and Spoken Relatives tasks of MATRS were administered one on one; the remaining MATRS tasks were administered in small groups of same-grade students. In Years 1 and 3, all norm-referenced tasks were administered one on one except for TWS-5, which was administered in small groups of same-grade students. All tasks were administered by individuals who were trained by the main investigators on all procedures used in the study. All tasks were scored by two different raters to ensure accuracy of scoring. Interscorer agreement was calculated for each task for 15% of the participants, and average agreement ranged from .92 to .99 across all MATRS tasks and grades.

DATA ANALYSIS AND RESULTS: ITEM CALIBRATION STUDIES

Tables for all psychometric results are contained in the MATRS Technical Report, which is available on the Download Hub (see the About the Downloads page in the front of this manual for access information). For both Years 1 and 2 of item calibration, data analyses were conducted after data were obtained to study the dimensionality of scores as well as the difficulty and discrimination of parameters. The data analyses were essentially the same for both Years 1 and 2 and included data screening, multiple-group item response modeling, reliability analysis, and differential item functioning. These analyses are part of item response theory (IRT), which allows researchers to select items based on item-level characteristics.

Data Screening

Data screening was chosen as the first step in data analysis to identify items that demonstrated strong floor or ceiling effects in response rates ≥95%. Such items are not useful in creating an item bank because there is little variability in whether students are successful on the item. In addition to evaluating the descriptive response rate, we estimated item-total correlations. Items with negative values are indicative of poor functioning because they suggest that individuals who correctly answer the question tend to have lower total scores. Items with low item-total correlations indicate the lack of a relation between item and total test performance. Items with correlations <.15 were flagged for removal.

Item-level *p* values were calculated for each item by year and grade. Dichotomous items demonstrating floor effects (e.g., item means ≥95% incorrect) or ceiling effects (e.g., item means ≥95% correct) were flagged for potential removal prior to the IRT modeling.

Results of Data Screening

Task 1: Segmenting The mean percent correct for Task 1 items was 0.46 (standard deviation [SD] = 0.28) with a minimum of 0.00 and a maximum of 0.97. Two items presented with floor effects (i.e., T1_12_6 and T1_13_23), and three items demonstrated ceiling effects (i.e., T1_14_4, T1_1_6, T1_6_3). Vertical linking items for Task 1 showed that 13 of the 20 items demonstrated reasonable increases in percent correct by grade level. See Tables D2 and D10 for results.[1]

Task 2: Affix Identification Task 2 contained a mix of two-category and three-category items. The mean percent correct for Task 2 dichotomous items was 0.34 (SD = 0.22) with a minimum of 0.02 and a maximum of 0.95. Seven items presented with floor effects (i.e., T2_85_1, T2_76_2,

[1] Throughout this chapter, "D" refers to tables found in the MATRS Technical Report, which is available as a download; refer to this book's About the Downloads page for access instructions.

T2_55_2, T2_72_1, T2_72_2, T2_75_12, T2_82_4), and one item demonstrated ceiling effects (i.e., T2_64_456). The mean response for Task 2 polytomous items was 1.23 (SD = 0.14) with a minimum of 1.01 and a maximum of 1.62. Vertical linking items for Task 2 showed that 21 of the 25 items demonstrated reasonable increases in percent correct by grade level. See Tables D3 and D11 for results.

Task 3: Affix Meaning The mean percent correct for Task 3 items was 0.47 (SD = 0.18) with a minimum of 0.03 and a maximum of 0.88. One item presented with floor effects (i.e., T3_40_1), and no items demonstrated ceiling effects. Vertical linking items for Task 3 showed that 40 of the 51 items demonstrated reasonable increases in percent correct by grade level. See Tables D4 and D12 for results.

Task 4: Suffix Choice The mean percent correct for Task 4 items was 0.65 (SD = 0.17) with a minimum of 0.11 and a maximum of 0.94. No items presented with floor or ceiling effects. Vertical linking items for Task 4 showed that 17 of the 23 items demonstrated reasonable increases in percent correct by grade level. See Tables D5 and D13 for results.

Task 5: Spelling Multi-Morphemic Words The mean percent correct for Task 5 items was 0.22 (SD = 0.18) with a minimum of 0.00 and a maximum of 0.76. A total of 49 items presented with floor effects, and no items demonstrated ceiling effects. Vertical linking items for Task 5 showed that 27 of the 34 items demonstrated reasonable increases in percent correct by grade level. See Tables D6 and D14 for results.

Task 6: Suffix Spelling The mean percent correct for Task 6 items was 0.60 (SD = 0.21) with a minimum of 0.13 and a maximum of 0.99. No items presented with floor effects, and two items demonstrated ceiling effects (i.e., T6_31_6, T6_10_4). Vertical linking items for Task 6 showed that 30 of the 37 items demonstrated reasonable increases in percent correct by grade level. See Tables D7 and D15 for results.

Task 7: Spoken Relatives The mean percent correct for Task 7 items was 0.51 (SD = 0.31) with a minimum of 0.00 and a maximum of 0.97. Seventeen items presented with floor effects, and six items demonstrated ceiling effects (i.e., T7_17_23, T7_21_4, T7_17_4, T7_16_45, T7_22_2, T7_27_4). Vertical linking items for Task 7 showed that 19 of the 26 items demonstrated reasonable increases in percent correct grade level. See Tables D8 and D16 for results.

Task 8: Written Relatives The mean percent correct for Task 8 items was 0.34 (SD = 0.24) with a minimum of 0.00 and a maximum of 0.85. Twenty-five items presented with floor effects, and no items demonstrated ceiling effects. Vertical linking items for Task 8 showed that 19 of the 23 items demonstrated reasonable increases in percent correct by grade level. See Tables D9 and D17 for results.

Multiple-Group Item Response Modeling (MG-IRM)

Data from MATRS were analyzed using multiple-group IRT. This approach to data analysis differs from classical test theory (CTT). In CTT, testing and analysis of items involve estimating the difficulty of the item (based on the percentage of respondents correctly answering the item) as well as discrimination (how well individual items relate to overall test performance). Although CTT practices are common in assessment development, IRT holds several advantages over CTT. When using CTT, the difficulty of an item depends on the group of individuals on which the data were collected. This means that if a sample has more students who perform at an above-average level, the items will appear to be easier, but if the sample has more below-average performers, the items will appear to be more difficult. Similarly, the more students differ in their abilities, the more likely the discrimination of the items will be high; the more similar the students are in their ability, the lower the discrimination will be. Thus, scores from a CTT approach

can be heavily dependent on the makeup of the sample on which the items are tested. The benefits of IRT are such that 1) item difficulty, item discrimination, and pseudo-guessing parameters are not dependent on the group(s) from which they were initially estimated; 2) scores describing students' abilities are not related to the difficulty of the test; 3) shorter tests can be created that are more reliable than a longer test; and 4) item statistics and the ability of students are reported on the same scale.

A series of multiple-group unidimensional IRT models were fit to the item-level data corresponding to each of the individual tasks. Following the data screening processes, these analyses allowed for an evaluation of item properties, specifically the item difficulty. The multiple-group approach to the item-response modeling facilitated a vertical equating of the item parameters across groups (i.e., grade levels) as well as the vertical scaling of ability scores across groups. One aspect of IRT modeling is that the probability of item accuracy may be modeled according to constraints associated with particular item properties including the item difficulty, item discrimination, and pseudo-guessing parameter. For example, a one-parameter logistic response model (1PL) estimates each item's difficulty value but constrains the item discrimination value to equality across all items and fixes the pseudo-guessing parameter to 0. The two-parameter logistic response model (2PL) estimates individual item difficulty and discrimination coefficients and fixes the pseudo-guessing parameter to 0; and the three-parameter logistic response model (3PL) estimates individual item difficulty, discrimination, and pseudo-guessing coefficients. A variation of the 1PL models is the Rasch model, whereby all item discrimination values are fixed to 1.0. A benefit of the Rasch model is that raw total scores from items can be more easily converted to a Rasch scale score compared to using other forms of IRT. This feature is particularly advantageous for paper-and-pencil–delivered or hand-scored assessments.

Rasch models were fit using flexMIRT software (Cai, 2013) and were evaluated using local fit (i.e., performance of the individual items) and goodness of fit based on the M_2 statistic (Maydeu-Olivares, 2013), the root mean square error of approximation based on M_2 ($RMSEA_2$), and the Tucker-Lewis Index (TLI). M_2 is often sensitive to sample size in terms of rejecting the fitted model; thus, the $RMSEA_2$ is useful for determining adequate fit (<.089), close fit (<.05), or excellent fit [.05/(k − 1), where k = number of categories]. Marginal reliability was reported by grade for each of the administered tasks using Thissen and Orlando (2001), whereby the variance of theta minus the averaged error variances is divided by the variance of theta.

Results of Multiple-Group Item-Response Modeling

Test characteristic curves and test information functions are presented in Figures D1 and D2. Item characteristic curves are available upon request.

Task 1: Segmenting Model fit for the Task 1 unidimensional MG-IRM resulted in a rejection of the null hypothesis of a correctly specified model (M_2 = 2,003.46, p <.001) with poor fit also suggested by the TLI of .82. The RMSEA was estimated as 0.02. The mean b value was 0.25 (SD = 1.50) with a minimum of −3.16 and a maximum of 3.08. Marginal reliability ranged from .67 in Grade 6 to .78 in Grade 1. The vertical scale for Task 1 ability scores did not demonstrate consistent increases across the grade levels—decreasing from grades 2 to 3 and remaining relatively constant from grades 5 to 6. See Tables D18, D19, D27, and D28 for results.

Task 2: Affix Identification Based on the threshold sensitivity of the polytomous items, scores were recoded to be dichotomous. Model fit for the Task 2 unidimensional MG-IRM resulted in a rejection of the null hypothesis of a correctly specified model (M_2 = 16,559.29, p <.001) with poor fit also suggested by the TLI of .85. The RMSEA was estimated as 0.03. The mean b value was 0.03 (SD = 1.20) with a minimum of −1.50 and a maximum of 4.48. Marginal reliability ranged from .80 in Grade 1 to .92 in grades 5–6. The vertical scale for Task 2 ability scores demonstrated consistent increases from −1.96 in Grade 1 to 1.39 in Grade 6. See Tables D18, D20, D27, and D28 for results.

Task 3: Affix Meaning Model fit for the Task 3 unidimensional MG-IRM resulted in a fail-to-reject the null hypothesis of a correctly specified model (M_2 = 6,365.34, p >.500) with acceptable fit also suggested by the TLI of 1.00 and the RMSEA of <.001. The mean b value was 0.03 (SD = 0.74) with a minimum of −1.82 and a maximum of 2.29. Marginal reliability ranged from .82 in Grade 1 to .94 in Grade 4. The vertical scale for Task 3 ability scores demonstrated consistent increases from −1.05 in Grade 1 to 0.76 in Grade 6 with similar ability in Grades 4 and 5. See Tables D18, D21, D27, and D28 for results.

Task 4: Suffix Choice Model fit for the Task 4 unidimensional MG-IRM resulted in a fail-to-reject the null hypothesis of a correctly specified model (M_2 = 1,473.90, p >.500) with acceptable fit also suggested by the TLI of 1.00 and the RMSEA of <.001. The mean b value was −0.94 (SD = 1.09) with a minimum of −3.16 and a maximum of 3.05. Marginal reliability ranged from .89 in Grade 6 to .91 in grades 3–4. The vertical scale for Task 4 ability scores demonstrated relatively consistent increases from −0.76 in Grade 3 to 0.76 in Grade 6 with similar ability in grades 4 and 5. See Tables D18, D22, D27, and D28 for results.

Task 5: Spelling Multi-Morphemic Words Model fit for the Task 5 unidimensional MG-IRM resulted in a fail-to-reject the null hypothesis of a correctly specified model (M_2 = 1,761.84, p >.500) with acceptable fit also suggested by the TLI of 1.00 and the RMSEA of <.001. The mean b value was 0.97 (SD = 1.43) with a minimum of −3.22 and a maximum of 4.48. Marginal reliability ranged from .89 in Grade 1 to .95 in Grade 4. The vertical scale for Task 5 ability scores demonstrated relatively consistent increases from −2.80 in Grade 3 to 0.00 in Grade 6. See Tables D18, D23, D27, and D28 for results.

Task 6: Suffix Spelling Model fit for the Task 6 unidimensional MG-IRM resulted in a rejection of the null hypothesis of a correctly specified model (M_2 = 5,508.05, p <.001) with acceptable fit suggested by the TLI of .97 and the RMSEA of .01. The mean b value was −0.21 (SD = 0.99) with a minimum of −3.71 and a maximum of 2.84. Marginal reliability ranged from .84 in Grade 1 to .90 in Grades 3–4. The vertical scale for Task 6 ability scores demonstrated relatively consistent increases from −0.71 in Grade 1 to 1.26 in Grade 6 with similar ability in Grades 4 and 5. See Tables D18, D24, D27, and D28 for results.

Task 7: Spoken Relatives Model fit for the Task 7 unidimensional MG-IRM resulted in a fail-to-reject the null hypothesis of a correctly specified model (M_2 = 1,720.92, p >.500) with acceptable fit also suggested by the TLI of 1.00 and the RMSEA of <.001. The mean b value was 0.00 (SD = 2.05) with a minimum of −4.09 and a maximum of 5.13. Marginal reliability ranged from .82 in grades 2 and 4 to .88 in Grade 6. The vertical scale for Task 7 ability scores did not demonstrate consistent increases across the grade levels—decreasing from grades 4 to 5. See Tables D18, D25, D27, and D28 for results.

Task 8: Written Relatives Model fit for the Task 8 unidimensional MG-IRM resulted in a fail-to-reject the null hypothesis of a correctly specified model (M_2 = 1,983.86, p >.500) with acceptable fit also suggested by the TLI of 1.00 and the RMSEA of <.001. The mean b value was 0.69 (SD = 1.76) with a minimum of −3.16 and a maximum of 6.03. Marginal reliability ranged from .93 in grades 5–6 to .95 in Grade 3. The vertical scale for Task 8 ability scores demonstrated relatively consistent increases from −1.23 in Grade 3 to 0.52 in Grade 6. See Tables D18, D26, D27, and D28 for results.

Differential Item Functioning

Differential item functioning (DIF) was chosen as the next step in the analyses. DIF refers to instances where individuals from different groups with the same level of underlying ability significantly differ in their probability to correctly endorse an item. Unchecked, items included in a test that demonstrate DIF will produce biased test results. DIF was estimated using the

difR package (Magis et al., 2020) using the Mantel-Haenszel method (Mantel & Haenszel, 1959) for detecting uniform DIF. For each of the six MATRS tasks, DIF was tested for four primary contrasts: 1) male versus female, 2) white versus sample, 3) black versus sample, and 4) Latinx versus sample. The Mantel-Haenszel chi-square statistic was reported for test by item, and the chi-square was used to derive an effect size estimate (i.e., ETS delta scale; Holland & Thayer, 1989). Effect size values ≤1.0 are considered small, values of 1.0–1.5 are moderate, and values ≥1.5 are considered large.

Results of DIF

Results for the DIF analyses are reported in Tables D29–D60. Across all tasks, five items demonstrated significant DIF (i.e., T3_1_1, T3_2_1, T4_20_3, T4_9_5, T6_14_456). Remaining items all presented with nonsignificant chi-square statistics and ETS delta values near 0.00, indicating no practically important DIF.

DATA ANALYSIS AND RESULTS: TASK VALIDATION STUDIES

In Years 1 and 3 of the study, scores from the MATRS analyses were correlated with concurrently administered standardized assessments of reading and language. (See Table 5.2 for names of abbreviated tests.) Year 1 validity analyses included grade-based MATRS task correlations with the PPVT4, TOWRE-2 Sight Word Efficiency (SWE), TOWRE-2 Phonemic Decoding Efficiency (PDE), and TWS-5; multiple regression analyses tested the additive and interactive strength of MATRS scores explaining individual differences in the PPVT-4, TOWRE -2 SWE, TOWRE-2 PDE, and TWS-5. Year 3 validity analyses included grade-based, MATRS task concurrent correlations with the PPVT-4, TWS-5, Elision, Blending, and WRMT-3 Word Identification (WID), Word Attack (WA), and Passage Comprehension (PC) subtests. Predictive relations were estimated between MATRS task scores and TWS-5, along with the WRMT-3 WID, WA, and PC tasks. For each set of concurrent and predictive correlations in Year 3, results are reported separately for Versions A and B of MATRS tasks that were administered in that portion of the study. Multiple regression analyses tested the additive strength of MATRS scores for each of Versions A and B explaining individual differences in the WRMT-3 WID, WA, and PC tasks as well as the TWS-5.

RESULTS

Correlational and multiple regression analyses were used to determine the relation among performance on the MATRS tasks and the relations between performance on MATRS and the norm-referenced language and literacy measures, hereafter referred to as "outcomes." In some

Table 5.2. Names of Abbreviated Tests and Subtests

Abbreviation	Name
PPVT-4	Peabody Picture Vocabulary Test, Fourth Edition
TOWRE-2	Test of Word Reading Efficiency, Second Edition
SWE	Sight Word Efficiency
PDE	Phonemic Decoding Efficiency
TWS-5	Test of Written Spelling-5th Edition
WRMT-3	Woodcock Reading Mastery Test, Third Edition
LWID	Word Identification
WA	Word Attack
PC	Passage Comprehension
CTOPP-2	Comprehensive Test of Phonological Processing, Second Edition

instances, there was no relation between students' performance on different MATRS tasks within a grade, which was not unexpected given that various tasks were designed to measure different domains of morphological awareness. Likewise, concurrent validity between performance on the MATRS tasks and the language and literacy measures was variable. Again, the differing magnitude of relations among the gold standard measures and each of the MATRS tasks for each grade were not altogether unexpected given that certain tasks represent specific domains of morphological awareness. For example, tasks within Domain 1 measure the awareness of spoken or written forms of morphemes, whereas tasks within Domain 4 measure awareness of the meaning connections between base words and their inflected or derived forms. Similarly, the gold standard measures administered measure various aspects of literacy and literacy-related skills. Thus, differences in the strengths of the relations were expected. See summary information below or refer to Tables D61–D80 for correlational values.

Results of Task Validity Studies

Year 1

Grade 1 Correlations among MATRS ability scores ranged from .07 between Task 1 and Task 3 to .62 between Task 5 and Task 6. Correlations between MATRS ability scores with outcomes ranged from .00 between Task 1 and the PPVT-4 to .48 between Task 5 and TOWRE-2 PDE. Multiple regression R^2 values for the additive MATRS model ranged from .08 for the PPVT-4 to .27 for the TWS-5. Multiple regression R^2 values for the MATRS two-way interactions model ranged from .17 for the PPVT to .40 for the TWS. See Tables D61 and D67 for results.

Grade 2 Correlations among MATRS ability scores ranged from .21 between Task 1 and Task 5 to .72 between Task 5 and Task 6. Correlations between MATRS ability scores with outcomes ranged from .03 between Task 1 and the TWS-5 to .53 between Task 5 and TWS-5. Multiple regression R^2 values for the additive MATRS model ranged from .10 for the PDE to .30 for the TWS-5. Multiple regression R^2 values for the MATRS two-way interactions model ranged from .17 for the PDE to .42 for the TWS. See Tables D62 and D67 for results.

Grade 3 Correlations among MATRS ability scores ranged from .08 between Task 1 and Task 4 to .74 between Task 5 and Task 8. Correlations between MATRS ability scores with outcomes ranged from .00 between Task 1 and the TWS to .48 between Task 5 and TOWRE-2 PDE. Multiple regression R^2 values for the additive MATRS model ranged from .11 for the SWE to .27 for the TWS-5. Multiple regression R^2 values for the MATRS two-way interactions model ranged from .81 for the SWE, PDE, and TWS-5 to .82 for the PPVT-4. See Tables D63 and D67 for results.

Grade 4 Correlations among MATRS ability scores ranged from .06 between Task 1 and Task 6 to .77 between Task 4 and Task 8. Correlations between MATRS ability scores with outcomes ranged from .00 between Task 1 and the TOWRE-2 SWE to .49 between Task 5 and TWS-5. Multiple regression R^2 values for the additive MATRS model ranged from .15 for the SWE to .26 for the TWS-5. Multiple regression R^2 values for the MATRS two-way interactions model ranged from .60 for the SWE to .67 for the TWS-5. See Tables D64 and D67 for results.

Grade 5 Correlations among MATRS ability scores ranged from .11 between Task 1 and Task 7 to .77 between Task 5 and Task 8. Correlations between MATRS ability scores with outcomes ranged from .00 between Task 1 and the TWS-5 to .51 between Task 5 and TWS-5. Multiple regression R^2 values for the additive MATRS model ranged from .08 for the TOWRE-2 SWE to .27 for the TWS-5. Multiple regression R^2 values for the MATRS two-way interactions model ranged from .71 for the SWE to .76 for the TWS. See Tables D65 and D67 for results.

Grade 6 Correlations among MATRS ability scores ranged from −.02 between Task 1 and Task 6 to .75 between Task 5 and Task 8. Correlations between MATRS ability scores with outcomes ranged from .00 between Task 1 and the PPVT-4 to .48 between Task 6 and TWS-5. Multiple regression R^2 values for the additive MATRS model ranged from .12 for the TOWRE-2 PDE

to .35 for the TWS. Multiple regression R^2 values for the MATRS two-way interactions model ranged from .79 for the TOWRE-2 SWE to .86 for the PPVT-4. See Tables D66 and D67 for results.

Year 3

Grade 1 Version A correlations among MATRS ability scores ranged from .04 between Task 1 and Task 6 to .36 between Task 3 and Task 7. Concurrent correlations between MATRS ability scores with outcomes ranged from .00 between Task 6 and WRMT-3 WID to .44 between Task 7 and PPVT-4. Predictive correlations between MATRS ability scores with outcomes ranged from .01 between Task 2 and WRMT-3 WA to .56 between Task 5 and TWS-5. Multiple regression R^2 values for the additive MATRS model ranged from .30 for WID to .33 for the TWS-5. See Tables D68 and D80 for results.

Version B correlations among MATRS ability scores ranged from .08 between Task 1 with Task 2 and Task 5 to .63 between Task 5 and Task 6. Concurrent correlations between MATRS ability scores with outcomes ranged from .05 between Task 1 and PC to .81 between Task 5 and TWS-5. Predictive correlations between MATRS ability scores with outcomes ranged from −.02 between Task 2 and TWS-5 to .78 between Task 5 and TWS-5. Multiple regression R^2 values for the additive MATRS model ranged from .44 for WRMT-3 PC to .66 for the TWS-5. See Tables D74 and D80 for results.

Grade 2 Version A correlations among MATRS ability scores ranged from .03 between Task 1 and Task 2 to .65 between Task 5 and Task 6. Concurrent correlations between MATRS ability scores with outcomes ranged from .05 between Task 1 and Elision to .75 between Task 5 and WRMT-3 WA. Predictive correlations between MATRS ability scores with outcomes ranged from .24 between Task 2 and WRMT-3 WID and WA to .78 between Task 5 and TWS-5. Multiple regression R^2 values for the additive MATRS model ranged from .47 for PC to .65 for the TWS-5. See Tables D69 and D80 for results.

Version B correlations among MATRS ability scores ranged from .11 between Task 1 with Task 6 and Task 7 to .58 between Task 5 and Task 6. Concurrent correlations between MATRS ability scores with outcomes ranged from .09 between Task 1 and PPVT-4 to .68 between Task 5 with PPVT-4 and TWS-5. Predictive correlations between MATRS ability scores with outcomes ranged from .00 between Task 7 and PC to .79 between Task 5 and TWS. Multiple regression R^2 values for the additive MATRS model ranged from .37 for PC to .65 for the TWS-5. See Tables D75 and D80 for results.

Grade 3 Version A correlations among MATRS ability scores ranged from .00 between Task 1 with Task 5 and Task 8 to .65 between Task 5 and Task 8. Concurrent correlations between MATRS ability scores with outcomes ranged from .03 between Task 6 and Blending to .76 between Task 5 and WRMT-3 WA. Predictive correlations between MATRS ability scores with outcomes ranged from .01 between Task 1 and WA to .76 between Task 5 and TWS-5. Multiple regression R^2 values for the additive MATRS model ranged from .61 for WA to .71 for the WRMT-3 WID. See Tables D70 and D80 for results.

Version B correlations among MATRS ability scores ranged from .01 between Task 1 and Task 2 to .73 between Task 5 and Task 7. Concurrent correlations between MATRS ability scores with outcomes ranged from .12 between Task 1 and Blending to .75 between Task 5 and WRMT-3 WID. Predictive correlations between MATRS ability scores with outcomes ranged from −.01 between Task 1 and WRMT-3 WA to .78 between Task 5 and TWS-5. Multiple regression R^2 values for the additive MATRS model ranged from .35 for PC to .69 for the TWS-5. See Tables D76 and D80 for results.

Grade 4 Version A correlations among MATRS ability scores ranged from .02 between Task 1 with Task 7 to .73 between Task 5 and Task 8. Concurrent correlations between MATRS ability scores with outcomes ranged from .01 between Task 1 and TWS to .75 between Task 5 and WRMT-3 WA. Predictive correlations between MATRS ability scores with outcomes ranged

from .09 between Task 1 and TWS to .77 between Task 5 and WRMT-3 WID. Multiple regression R^2 values for the additive MATRS model ranged from .46 for PC to .73 for the WID. See Tables D71 and D80 for results.

Version B correlations among MATRS ability scores ranged from .01 between Task 1 and Task 2 to .73 between Task 5 and Task 7. Concurrent correlations between MATRS ability scores with outcomes ranged from −.01 between Task 1 and PPVT-4 to .77 between Task 8 and TWS-5. Predictive correlations between MATRS ability scores with outcomes ranged from .16 between Task 1 and WRMT-3 LWID to .78 between Task 5 and TWS-5. Multiple regression R^2 values for the additive MATRS model ranged from .42 for PC to .67 for the WID. See Tables D77 and D80 for results.

Grade 5 Version A correlations among MATRS ability scores ranged from .00 between Task 1 with Task 7 to .70 between Task 4 and Task 8. Concurrent correlations between MATRS ability scores with outcomes ranged from .14 between Task 1 and PC to .75 between Task 8 and WRMT-3 WID and PC. Predictive correlations between MATRS ability scores with outcomes ranged from .04 between Task 1 and WRMT-3 PC to .92 between Task 8 and TWS-5. Multiple regression R^2 values for the additive MATRS model ranged from .76 for WID to .92 for the TWS-5. See Tables D72 and D80 for results.

Version B correlations among MATRS ability scores ranged from .09 between Task 1 and Task 3 to .65 between Task 4 and Task 8. Concurrent correlations between MATRS ability scores with outcomes ranged from .13 between Task 1 and Elision to .74 between Task 5 and WRMT-3 WID as well as between Task 8 and TWS-5. Predictive correlations between MATRS ability scores with outcomes ranged from .21 between Task 2 and WA as well as Task 7 and TWS-5 to .75 between Task 4 and TWS-5. Multiple regression R^2 values for the additive MATRS model ranged from .45 for WRMT-3 PC to .70 for the TWS-5. See Tables D78 and D80 for results.

Grade 6 Version A correlations among MATRS ability scores ranged from .05 between Task 1 with Task 4 to .70 between Task 4 and Task 8. Concurrent correlations between MATRS ability scores with outcomes ranged from −.01 between Task 1 and Elision to .84 between Task 8 and TWS-5. Predictive correlations between MATRS ability scores with outcomes ranged from −.03 between Task 1 and WRMT-3 PC to .82 between Task 5 and TWS-5. Multiple regression R^2 values for the additive MATRS model ranged from .59 for WRMT-3 WA to .86 for the TWS-5. See Tables D73 and D80 for results.

Version B correlations among MATRS ability scores ranged from −.02 between Task 1 with Task 3 and Task 7 to .70 between Task 5 and Task 8. Concurrent correlations between MATRS ability scores with outcomes ranged from .00 between Task 1 with TWS-5, CTOPP-2 Elision, and Blending to .72 between Task 8 and TWS-5. Predictive correlations between MATRS ability scores with outcomes ranged from .22 between Task 4 and WRMT-3 PC as well as Task 7 and TWS-5 to .88 between Task 5 and WRMT-3 WID. Multiple regression R^2 values for the additive MATRS model ranged from .59 for PC to .91 for the WID. See Tables D79 and D80 for results.

Multiform Construction and Estimated Reliability

The Rasch model item difficulties were used to construct approximately equivalent forms of items for each task by grade. The vertical scale results (Table D28) combined with the domain representativeness informed the task selection and item selection by grade. The intersection and representation of Domain and Task by grade level were selected as follows: Domain 1 = Task 1 (grades 1–2), Task 2 (grades 3–6); Domain 2 = Task 3 (grades 1–4), Task 4 (grades 5–6); Domain 3 = Task 5 (grades 4–6), Task 6 (grades 1–3); Domain 4 = Task 7 (grades 1–2), Task 8 (grades 3–6). For each of the tasks by grade, the theta location point for each task by grade (Table D28) was used as the initial point of item selection. A driving mechanism for constructing the forms was to achieve a marginal reliability estimate of at least .80 at the theta location for each task by grade.

The total number of items needed to achieve the reliability threshold of .80 ranged from 21 to 25. We selected approximately 10 items per form by task and grade that were centered on the theta location for each task by grade. Approximately eight items were selected that were distributed by 1.0 SD below the theta location, and approximately seven items were selected that were distributed 1.0 above the theta location. The resulting reliability of scores for all tasks by grade then approximated or was greater than .80.